User's Manual to

ITRS

Interrater/Test Reliability System

by

Jamal Abedi
# TABLE OF CONTENTS

1. Introduction 1-1

2. How to Install the System 2-1

3. Subsystems 3-1
   3.1 Interrater Reliability Section 3-1
   3.2 Test Reliability Section 3-4

4. Data Input 4-1
   4.1 Interactive Mode 4-1
   4.2 ASCII File Mode 4-2

5. System's Output 5-1

6. Input Example 6-1

7. Output Example 7-1

8. Comments 8-1

9. Format 9-1

10. Missing Data 10-1
    10.1 Deleting Subjects With Missing Values 10-3
    10.2 Replacing Missing Values 10-3

11. Error Messages 11-1
    11.1 Error Messages Related to File Name 11-1
    11.2 Error Messages Related to the FORTRAN-Type Format Statement 11-1
    11.3 Error Messages Related to the Input Data 11-2

12. Example Section 12-1
    12.1 Example 1 12-2
    12.1.1 Example 1 Input 12-3
    12.1.2 Example 1 Output 12-4
12.10.1 Example 10 Input 12-41
12.10.2 Example 10 Output 12-42
12.10.3 Example 10 Comments 12-45
12.11 Example 11 12-46
12.11.1 Example 11 Input 12-47
12.11.2 Example 11 Output 12-48
12.11.3 Example 11 Comments 12-49

13. Appendix A, Literature Review 13-1

14. References 14-1
1. Introduction

Many different techniques have been used for estimating interrater reliability. The variations and limitations of these statistical techniques have been well documented in the literature (see Appendix A for a review of related literature). The ITRS system was designed as a comprehensive tool to address a multitude of questions and concerns regarding interrater reliability. This system provides a unique opportunity for those who are dealing with rating of any kind whether it involves assignment of ratings to a performance-based assessment test, categorizing patients into different categories of illness, or placing social issues into different groups. For example, the most complex problem facing assessment in a portfolio assessment case is testing the reliability of the scoring rubric and the consistency between raters. The ITRS system can aid in this type of assessment by obtaining the most appropriate measures of interrater reliability and identifying any potential problems.

The system computes several different indices of interrater reliability for all raters or any combination of raters that may be employed. This capability enables the user to address various concerns such as the effects of raters' background on their ratings.

In addition to providing different indices of interrater reliability, ITRS computes the generalizability coefficient (G-coefficient) of raters and topics/tasks based on the available data. The generalizability coefficient can be expanded over any number of raters/tasks and can provide useful information needed to determine how many raters/tasks are necessary to reach the desired level of generalizability. The system can also be used for estimating reliability of multiple-choice as well as other types of tests.
2. How to install the system

I. Macintosh Version

(a) Copy the ITRS folder onto your MAC hard disk;
(b) to run the system, double-click on: ITRS;
(c) follow the instructions.

You can open ITRS text document (such as output files) using MS Word or any other word processors.

II. IBM version

To install the system do the following:

Insert the "SYSTEM/ILLUSTRATION DISK" into your floppy disk drive (Drive A or B) and then type: INSTALL

The Install program then asks you for your source drive, i.e. the drive that you have the "SYSTEM/ILLUSTRATION DISK". This is usually your "A" drive. Then, the program asks you to name the directory in which you want to store the ITRS software. It is usually C:\ITRS. For example:

-> INSTALL

What is the source drive? (ex.A: ) A:

What are your target drive and the directory name? (ex. C:\ITRS ) C:\ITRS

ITRS has now been successfully installed.

When you have installed the system, you can then enter the following DOS command to run the system (assuming that you have stored ITRS software on your drive C:, directory ITRS):

C:\ITRS <RETURN>
ITRS <RETURN>
3. Subsystems

The ITRS system, which consists of 32 subprograms, was developed mainly for estimating interrater reliability indices. However, to satisfy the needs of a broader group of researchers/practitioners, who may need to work with standardized achievement tests, the system includes sections on scoring, analyzing, and estimating the reliability of multiple-choice and other types of standardized instruments. ITRS has two different sections (subsystems): (1) the interrater reliability section, and (2) the test reliability section.

3.1 Interrater Reliability Section

For the first time, the ITRS system makes it possible for users to obtain a variety of different reliability estimates for a given instrument. Users can compare the indices and select the most appropriate statistics for a given circumstance. This section computes the following statistics as estimates of interrater reliability:

(a) Descriptive statistics including mean, standard deviation, minimum, and maximum scores (ratings);
(b) Product Moment correlation / Kedall's tau between pairs of raters, average and range of the correlations;
(c) percent of exact agreement;
(d) percent of agreement within explicit point range;
(e) Cronbach's alpha;
(f) kappa coefficient and the associated z;
(g) intraclass correlation;
(h) generalizability coefficient for the existing (actual) number of raters and tasks;
(i) a table of G-coefficient for different number of raters and tasks;
(j) Williams' index of agreement.

The system can compute any set of statistics or all of the above statistics for all raters as a group or any combination (groups) of raters. For example, suppose that a set of 200 history essays from high school students was scored by ten raters (not all of the essays were scored by all of the raters). Raters 1,2,5,7, and 8 scored all 200 essays. Raters 1, 2, and 5 are high school history teachers and raters 7 and 8 are graduate students in history. The system can provide one set of interrater reliability estimates (including all
the above statistics) for the history teachers (raters 1, 2 and 5), and another set of estimates for the graduate students (raters 7 and 8). A comparison between the two sets of interrater reliability estimates would reveal any systematic differences between the two groups of raters. These differences would register the impact of each rater's background on their ratings.

The system asks the user to specify how to group the raters by displaying the following prompt (we will display system's messages in Bold Italics):

*Enter 'ALL' for estimates based on all raters as a group, 'SEL' for selected combinations, and 'ALC' for all possible combinations.*

(a) **Overall estimates.** To select this option, respond "ALL" to the above system's command. The system then provides one set of interrater reliability estimates for all of the raters as a group. For example, in the case of the 10 raters, the percent of agreement, P.M. correlations, Alpha, Kappa, and other indices would be computed for the ten raters as a group. It must be indicated, however, that if all of the raters did not score all of the essays, this may not be an appropriate approach for data analysis. Option (b) below may be more appropriate for such cases.

(b) **Only selected combination(s).** To select this option, respond "SEL" to the above prompt. The program will provide statistics for those combinations of raters that are of interest to the user. For example, if there are 1,013 possible combinations, the user might only be interested in 10 of those combinations. If so, the system can provide the desired statistics for only the selected group(s) of raters. This option is particularly useful when all of the raters did not score all the essays. To select this option, respond "SEL" to the above system's command and tell the system which raters to put together as a group. Suppose, of the ten raters, one wants to know how consistent the history teachers (raters 1, 2, and 5) and graduate students (raters 7 and 8) are on their ratings. To do this, type each rater's number separated by a comma (,).
Each group of raters, must be separated from another group by a slash (/) and the list must be ended with a period (.). To instruct the system to form the two groups of raters mentioned above, in response to the command:

*Please input the combinations of raters for whom you need to estimate interrater reliability*

Simply type the following: 1,2,5/7,8.

(c) **Compare ratings for consistency.** The system can also compare a particular rater's ratings with all of the other raters' ratings in a group. To select this option, first follow the instructions above for selecting combinations of raters (SEL). After inputting the combinations of raters, compute the Williams' index by entering WI when asked to indicate what statistics are to be run. The system then asks you which rater you want to compare with the group. For example, suppose that we want to compare the consistency of ratings by rater 2 with other raters in the first group. That is, we want to compare rater 2 with raters 1, 2, and 5 as a group. In response to the system's question of *which rater you want to compare with others*, we type 2. It must be noted, however, that if the rater chosen is not part of the group, the system will not compute Williams' index, and will simply print "NA" (Not Applicable) in the Williams' index space.

(d) **All possible combinations of raters.** To select this option, respond "ALC" to the above prompt. In a case of 10 raters, there would be 1,013 possible combinations of 2 or more raters. The system will provide estimates for all 1,013 possible combinations. The user can simply review the estimates and look for desired combination(s).

### 3.2 Test Section

This section, which has two parts (one for multiple choice and one for other types of tests such as those with Likert-type items), performs the following functions:
(a) Computes frequencies for different options in an item and flags the options with dis-proportionate frequencies;
(b) scores the test using the key provided by user;
(c) computes mean and standard deviation for the total test and subscales as well as for the individual items;
(d) computes item-total and item-subscale correlation;
(e) computes alpha coefficient for the total test and/or for the subscales.
4. Data Input

The data can be inputted in three different ways: (a) SPSS/SAS/BMDP mode, (b) interactive mode, and (c) ASCII file mode.

(a) **SPSS/SAS/BMDP mode.** An SPSS or SAS or BMDP output file or a file from any other statistical software could be used as an input file to the system.

(b) **Interactive mode.** Based on some parameters that user provides, the system sets up a data input screen for data input and user can input the data directly to the system. In this mode, the data input section of the system checks the accuracy of data to the extent possible and saves the data permanently on a file.

(c) **ASCII file mode.** An ASCII data file (compiled by any word processor or any other software) could be used as an input file. A FORTRAN format statement should specify the location and format of the data.

Before describing different modes of data input, we must indicate that the system accepts ratings for multiple topics and/or multiple raters. Topics are first level in data input and raters are second level. In cases of multiple topics, all data for the first topic must be entered first followed by the data for the second topic, and so forth. Suppose you have a case with 3 topics/tasks and 4 raters on each topic/task. You must input ratings provided by the 4 raters for the first topic, then you start with ratings of the 4 raters for the second topic followed by the ratings of the 4 raters for the third topic. Thus, the number of data elements would be the number of topics \( (NT) \) multiplied by the number of raters \( (NR) \). For the above example with 3 topics and 4 raters rating each essay on each topic, enter 12 data elements \( (3 \times 4 = 12) \) for each subject. If all raters did not rate all essays for all topics, you need to replace the data element with a code for missing data elements such as 0 or 9. For example, suppose, the first subject obtained the following ratings from the four raters on the first topic: 3, 3, 4, 3 and the following ratings from the four raters on the second topic: 3, 4, 3, 0 (a rating of zero for the fourth rater indicates a
missing data) and finally the following from the four raters for the third topic: 4, 4, 3, 3. 
Data record for this subject would look like the following:

00133434304433

In this record, columns 1 to 3 "001" is the subject ID, columns 4 to 7 contains the ratings
for the first topic, columns 8 to 11 contains ratings for the second topic, and columns 12
and 15 contains ratings for the third topic.

4.1 SPSS/SAS/BMDP Mode

If you have a SAS or SPSS or BMDP program for analyzing your data file, then you can
use the same program to create an input file for the ITRS. Suppose, you have a data file
consisting of the following variables: Student ID, gender, SES, ethnicity, grade points for
language, math, science, and ratings by 4 raters on each of the two topics (American
history and World history, four ratings for the essay in American history and four rating
for the essay in World history) with the following arrangement in SPSS:

```
DATA LIST FILE='MYFILE' /ID 1-8 GENDER 9 SES 10 ETHNIC 11
   LANG 12-14 MATH 15-17 SCIEN 18-20 R1T1 TO R4T1 21-24 R1T2 TO
   R4T2 25-28.

COMPUTE GPA = MEAN (LANG,MATH,SCIEN).

............ (Other SPSS codes)

............ (Other SPSS codes)

FINISH.
```

To prepare an input file to be used for your ITRS, you can simply put a WRITE or
PRINT statement within your SPSS. Following is an example of such statement within
an SPSS program:

```
WRITE OUTFILE='MY_ITRS' /R1T1 TO R4T1 (4F1.0) R1T2 TO R4T2
   (4F1.0).
```
This **WRITE** statement generates a file called 'MY_ITRS' which contains ratings for raters 1 to 4 on the first topic and raters 1 to 4 on the second topic. The format statement of (4F1.0) instructs the computer to print each rating as a single digit number with no decimals. Therefore, each record of the file contains 8 one-digit number from column 1 to 8.

### 4.1 Interactive Mode

To input your data interactively, you should type "D" for direct input in response to the following system command:

```
Data input mode: enter "D" for direct input
"F", for input from a file->
```

When you type "D" the system asks for a file name to store the data:

```
Please input a file name for storing input data->
```

You can give any name (a combination of character and numeric) to your file (a maximum of 8 characters, plus a three character extension following a period). For example, names such as DATA.DAT or "JOHN2.FIL" or A1234567.TXT could all be valid names. After entering the file name, then the system can accept data.
Please input data based on the format you provided earlier

Enter "//" to close data input screen

For each subject you need to enter ___ data elements

Please make sure to press <RETURN> after each line>

23445

66432

45356

//

For each subject, at least one record (or one line) of data must be entered. Each line contains \( K \) data elements, \( K \) being the number of topics multiplied by the number of raters (or number of items in the TEST subsystem). After each line of data input, you should press <RETURN>. When you are done with the data input, you must type "///" followed by <RETURN>. The two slashes ("///") signals the end of data input. Any non-numeric characters would also cause the data input screen to close. After the data input was complete, the file could be used as existing file for any future analyses.

4.2 ASCII File Mode

An existing ASCII data file (compiled by any word processor or any other software) could be used as an input file. (Please note that if you use a word processor to compile your data, you should save your file as a DOS or ASCII file. See your word processor's manual for instruction of how to save a file as unformatted or ASCII.) To use an existing ASCII data file, you should respond "F" to the following system's command:
Data input mode: enter "D" for direct input
"F", for input from a file->

The system then asks for the file name. You must make sure that the name and location you enter corresponds exactly with the name and location of your existing file. If the name you enter differs from the name of the existing file, you will get the following error message:

FILE NOT FOUND.
5. System's Output

The output of the system depends on which part of the systems is selected. The interrater and the test subsystems output different statistics. The following section presents examples of system input and output.

Illustrations and Examples Section

On the system disk, there is an Illustration/Example section. This section contains 11 examples (lessons) which illustrate the application of the system on different data sets. Users can select the example(s) most relevant to their cases. Each lesson includes a brief description of a real or hypothetical study, system input (system control cards and data), and system output. The lessons also present a brief interpretation of the results. The 11 examples cover a wide range of applications of the system, from performance-based and standardized assessments to interrater reliability in general. To run the Illustration/Example part of the system, type: ITRS and then type: E and follow the instructions.

The Illustration/Example section of this system has two major goals: (1) to provide several different opportunities for users to practice and to run the system, and (2) to present some real applications of the concept of interrater reliability and to show how interrater reliability indices are computed and how they are interpreted. We therefore urge our users to pay special attention to the Illustration/Example section of this system particularly when they want to use ITRS for the first time.

On the system disk, we have included 11 different data sets to be used for this purpose. These data sets are called EXAMPL.E01 to EXAMPL.E11. To practice running ITRS, you can use any or all of these examples/files. These examples have been fully described in the Illustration/Example section of this manual. In that section, actual input to the system and output from the system have been presented. First, read each
example carefully, then execute the system by following the sequence of input for each example. The input system's command for an input from the user is displayed in the manual in courier font, and the user's response is in bold. After the user provides the appropriate data, the system computes and outputs all requested statistics. The system's output is also given for each example. Finally, there is a discussion section of the results for each example. Following is an example (this example is an addition to the 11 example mentioned above):

A group of 40 eighth-grade students was rated by their math, science, language art, and history teachers in creativity on a scale of 1 to 5 (1 being least creative and 5 being most creative). First, we want to know how consistent teachers are on rating their students in general, and second, if the math and science teachers as a group is more consistent than language art and history teachers as a group.

The scores were stored in an ASCII file which was named CREAT.TXT (subject's IDs were typed into columns 1-4 and ratings were typed into columns 5-8).
6. Input Example

In a large school district, a plan for a performance-based assessment was established in history, science, and math. The main concern of the Board of Education in this district in implementing this plan was the scorer or interrater reliability. They wish to implement a scoring procedure which could yield scores with high levels of reliability. They randomly selected a group of 109 8th-grade students. The students were asked to write a short essay on world geography. The essays were scored by two math and two social science teachers. The following are questions concerning the interrater reliability of this study:

1) Are the teachers' backgrounds related to their scoring of essays in geography? If so, which group of teachers would provide a more reliable set of scores (math or social science teachers)?

2) How many raters should score each essay in order for the average ratings to reach an acceptable level of interrater reliability (.80 or above).

3) If the budget does not allow to hire more than one rater, how that could work? Would one rater provide scores reliable enough to be considered in a city-wide performance-based assessment.

The following is an example of the setup instructions used to run the system. The system's messages are in italic/bold, the user's inputs are in bold, and explanations are in plain text and in parentheses.
-> ITRS

****

INTERRATER/TEST RELIABILITY SYSTEM (ITRS)

****

For Interrater Reliability System enter: I
For Test System enter: T
For Missing Data Handling enter: M
For Illustration/Example System enter: E

****

Please enter your choice at the prompt

****

-> I

INTERRATER/TEST RELIABILITY SYSTEM (ITRS)

The Interrater Subsystem

Enter number of topics/tasks: -> 1

Since we have only one topic, i.e., world geography, we entered 1.

Enter number of raters: -> 4

We have 4 teachers, therefore, we entered 4.

Enter "ALL" for estimates based on all raters as a group, "SEL" for selected combinations, and "ALC" for all possible combinations:

SEL

Since we were interested in some specific combinations of raters, we entered "SEL".

Enter the combinations of raters for whom you want to estimate interrater
reliability:

1,2/3,4/1,2,3,4.

The above set of numbers indicate that we want raters 1 and 2 (Math teachers) to be in one group, raters 3 and 4 (social science teachers) to be in another group, and raters 1, 2, 3, and 4 (all teachers) be in another group. Note that a slash (/) separates the first group of raters from the second group of raters, a comma (,) separates the raters within each group, and the set of numbers ended by a period (.)

Enter a FORTRAN-type FORMAT for the data:

(4X, 4F1.0)

We skip 4 columns and we read 4 one-digit numbers with 0 decimals. For a detailed description of the FORMAT statement, see the section on FORMAT.

Data input mode: enter "D" for direct input

"F", for input from a file-> F

Because we have already stored our data into a file named "GEOGR.TXT", we entered "F".

Enter file name-> GEOGR.TXT

This is the name of the file which contains the data.

Please enter the full name or the first two characters of the statistics you wish to be computed. Separate names by a comma (,) and end the list with a period (.)

Enter "ALL" for all statistics.

ALL

We requested that all of the interrater statistics available in this software be computed, therefore, we entered "ALL".

Which rater do you want to compare with the others? 2

We wanted to compare the consistency of ratings by rater 2 with other raters,
therefore, we entered 2.

Enter the filename on which the output must be saved or enter "LPT1" to direct output to a printer: GEOGR.OUT

GEOGR.OUT is the name of the file that all the system's output will be stored on.

You can type: TYPE GEOGR.OUT>LPT1 to print an already stored output file.

Print descriptive statistics? (Y/N) Y

Please stand by, computing.....

Your output is saved under: GEOGR.OUT

You can now print the content of the file or read the file using any word processor.

Stop - program terminated.
7. Output Example

Descriptive statistics for Topic: 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>109</td>
<td>2.83</td>
<td>1.56</td>
<td>.00</td>
<td>5.00</td>
</tr>
<tr>
<td>2</td>
<td>109</td>
<td>2.89</td>
<td>1.18</td>
<td>.00</td>
<td>5.00</td>
</tr>
<tr>
<td>3</td>
<td>109</td>
<td>2.93</td>
<td>1.35</td>
<td>.00</td>
<td>5.00</td>
</tr>
<tr>
<td>4</td>
<td>109</td>
<td>2.93</td>
<td>1.35</td>
<td>.00</td>
<td>5.00</td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2 (math teachers)

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
<td>Coef</td>
</tr>
<tr>
<td>1</td>
<td>109</td>
<td>49.54</td>
<td>97.25</td>
<td>.88</td>
<td>.88</td>
<td>.88</td>
<td>.85</td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

3, 4 (social science teachers)

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
<td>Coef</td>
</tr>
<tr>
<td>1</td>
<td>109</td>
<td>90.83</td>
<td>100.0</td>
<td>.97</td>
<td>.97</td>
<td>.97</td>
<td>.97</td>
</tr>
</tbody>
</table>
Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2, 3, 4 (all teachers)

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>109</td>
<td>48.62</td>
<td>92.66</td>
<td>.88</td>
<td>.97</td>
<td>.91</td>
<td>.85</td>
</tr>
</tbody>
</table>
8. Comments

The results show some discrepancies between the different indices of interrater reliability. However, the results indicated that, in general, raters in group 2 (the social science teachers) provided scores with higher levels of interrater reliability. That is, they are more consistent in their ratings of a geography essay than the math teachers. These results suggest that raters' background does have an impact on their rating practices. The percent of agreement for group one (math teachers) is 49.5% as compared with 90.8% for group two (social science teachers). The intraclass correlation for group one is .85 as compared with .97 for group two. Group one's kappa is .36 with a $z$ of 7.4 as compared with a kappa of .88 and a $z$ of 17.94 for group two. The Williams' index (comparing rater 2 with math teachers, $w = .50$ and comparing rater 2 with all teachers, $w = .69$) showed that rater 2 was fairly consistent with the other three teachers. This suggests that rater 2 alone would have provided fairly reliable ratings if there was a limit set on the number of raters. Williams' index was not computed for raters 3 and 4 as a group, so an "NA" (Not Applicable) was printed because that rater was not a part of this group.

To practice using the system, follow the steps in the input section of the example using GEOGR.TXT file. The same results will be obtained if the input instructions are followed carefully.
9. Format

A FORTRAN-type format statement must specify the length and the locations of data elements (ratings or responses to test items). The ITRS system accepts data elements in real mode with "F" notation. The F-notation permits the input of numbers with decimals. F-notation consists of an "F" followed by a number which indicates the length of the data element, followed by a period '.' and then by another number which indicates the length of the decimal points. Use "X" notation to skip columns and use slashes to skip the remainder of the current record and move to the next record. Two slashes (//) skips the entire next record and three slashes (///) skips two records. You can use slashes to skip as many records as you want. As an example of a FORMAT statement, suppose that one-digit ratings for six raters have been typed in columns 11 to 16. To read this data set you must provide the following Format statement to the system: (10X, 6F1.0). In this statement, 10X instructs the system to skip the first 10 columns and start with column 11. 6F1.0 instruct the system to read six one-digit real numbers from each line starting at column 11 (6 means read 6 numbers, F is the notation for real number, 1 after F means each number is a one-digit number, and a zero following the period means that there is no decimal points). As another example, suppose you have 5 raters who assigned a two-digit rating (with one digit as decimal such as 2.5, 3.2, etc.) to each of the two topics in a study. These digits are entered without a decimal point. For example, 25 is entered as 25. The ratings for the first topic were entered in columns 5 to 14 and the ratings for the second topics were typed from columns 40 to 49. The Format statement for this data set would look like the following:

(4X, 5F2.1, 25X, 5F2.1).

In this statement, 4X instructs the computer to skip 4 columns and start from column 5, then read 5 two-digit numbers (each number with one decimal point) starting from column 5. Then skip 25 columns to reach to column 40 and finally read another 5
two-digit numbers (each number with one decimal) starting from columns 40. As a third example, suppose that we have a 10-item test. Responses to the test items are entered into columns 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20. To read this data, one must type the following format:

\[(1X, F1.0, 1X, F1.0, 1X, F1.0, 1X, F1.0, 1X, F1.0, 1X, F1.0, 1X, F1.0, 1X, F1.0, 1X, F1.0, 1X)\].

The format statement must always be enclosed in a pair of parentheses "()". If the Format statement is not correct, the following error message will occur and the system will stop:

\[\text{FORMAT ERROR!!!!}\]

If any notation other than "F" notation (such as "T" or "A" notation) is used, the following error message may occur:

\[\text{Data Format Error in File XXXX}\] (where XXXX is the data file name).
10. Missing Data

In practice, there are many different instances in which a small portion of the data is missing. Since there is not an easy way in the commonly used statistical packages for missing data replacement, these cases usually are deleted from the analyses. This problem can cause substantial loss of data and therefore can prevent relevant data analyses. **ITRS enables the user to replace missing elements in cases where a small portion of one subject’s data is missing, thus, saving records of data.**

The issue of missing values is extremely complex and requires special attention. There have been numerous discussions on this topic in the literature and different strategies have been suggested for the replacement of missing values. Among the most commonly used procedures mentioned in the literature are replacing missing values based on the group means, correlations, multiple regression, and IRT.

Frequently, the missing values on a variable are replaced with the average of that variable over the entire sample. While this is a practical and a very simple approach, it sometimes produces unsatisfactory results. This occurs because the aggregate measures are not usually a valid replacement for individual scores especially when the within subject variations are too high. Unfortunately, the details of this problem cannot be addressed in this summary. However, we examined the different techniques for missing data replacement and we found the following strategy to be the most efficient and the most reliable: **replace the missing value with the average of the average of the rows and columns.** An average of the scores of the variable with the missing element over all other subjects who have valid scores on that variable would provide an estimate for the missing value which is the aggregate of the scores over all subjects with complete data. As mentioned earlier, this might be biased if there is high within subject variability and it may not be quite consistent with the score of that particular student whose score is missing in that particular variable. For example, an estimate of a missing value of a low
achieving student who is among a group of high achieving students may be biased upward if the group average is used as replacement for the missing value. On the other hand, if the missing value is replaced by the subject's valid score on other non-missing data elements, another source of bias could result because the mean of the variable with missing values might be considerably different than the mean of other variables and the correlation between the variable with missing values and other variables may not be high. An average of these two averages could provide a better estimate of the missing value. This can be done by ITRS.

You can declare a certain number to represent the missing values in your numeric data files. ITRS provides two different options for dealing with missing values: option A, ignores cases with missing values from the analyses; and option B, replaces missing values with the average of the valid scores of other variables for the same subject OR the average of valid scores of other subjects on the variable with the missing elements OR an average of the two averages. This will be done when the proportion of missing values is not large.

10.1 Option A: Deleting Subjects With Missing Values
If you have a few records with a substantial number of missing data elements, you must instruct the system to ignore those records in analyses. To do this, run MISSING subsystem, go to ITRS or to your root directory, type the following, then follow the instructions given:

```
ITRS <RETURN>
M <RETURN>
```

10.2 Option B: Replacing Missing Values
ITRS replaces the missing value with the average of the variable with missing data over
all subjects (columns), the average of the other variables with no missing values for the same subject (rows), or the average of the averages of the rows and columns. To explain the procedure of replacing missing values with the average of rows and columns, a mock data set has been designed to help elucidate usage and explanation. Suppose you have a set of essays on two different topics. Each essay was scored by four raters on six different dimensions. Suppose that rater number one did not score subject number five on the first dimension for the first topic. To replace this missing value, the system computes an average of the scores for the first topic and the first dimension for all the subjects. The program then computes an average of the scores that other raters (raters 2, 3, and 4) have assigned to the first dimension for subject number 5. Then, the program computes an average of the two averages.

The system keeps track of the number of missing values. If the number of missing values is greater than one-third of the total number of subjects, then the system does not make any replacement.

To do this option, run MISSING subsystem by going to ITRS or to your root directory and typing:

    ITRS <RETURN>
    M <RETURN>

After you are done with missing data replacement, go to the Interrater or Test section of the system and use the data file name which you used for the missing subsystem. The system recognizes that file name and treats the missing values in that file based on the instructions given to the system through the MISSING subsystem.

10 - 3
11. Error Messages

The most common forms of error messages are messages on (a) file name, (b) format statement, and (c) data input. We will explain each of these categories of error in more detail.

11.1 Error Messages Related to File Name

Errors of this type occur when the file that was introduced as the input data file either does not exist or is in a different location (different directory or folder than the location specified by the user). To avoid this type of error, you must make sure that you gave the correct spelling and location (directory and sub directory or folders) for the file. If there is a spelling error on the file name, or the exact location of the file has not been correctly specified, then the following error message will occur: "File not found". If you get this error message, simply check the exact spelling and location and re-input both parameters. Another type of error message occurs when a non-numeric field is encountered in numeric data fields. For example, if the data are to be read based on the following format: (10X, 10F1.0), then columns 11 to 20 must all be numeric. A non-numeric data element such as a comma (,) or any other characters would cause an error message to occur. This will be explained later in this chapter.

11.2 Error Messages Related to the FORTRAN-Type Format Statement

As described in the FORMAT section, you must specify the length and location of data in "F" notation. You can use "X" notation to skip columns or slash (/) to skip records. You must separate characters or notations by commas. Other characters may not be acceptable and could cause error messages. It was also indicated that the FORMAT statement must be enclosed in a pair of parentheses. A missing right or left parentheses, characters not recognizable by a FORTRAN Format statement, or extra commas will cause error. Any kind of error in a FORMAT statement could cause the following error
message to occur:

\textit{FORMAT ERROR!!!!}

If this happened, you should check your format and make sure everything follow the guidelines explained in the Format section and then run the program again.

11.3 Error Messages Related to the Input Data

The data elements that are read by "F" notation must all be numeric. Non-numeric data elements would cause the following error message to occur:

\textit{Data Error in File XXXX (where XXXX is the name of the file)}

A similar error could occur if the data file has been saved as a formatted (non-ASCII) word processing file. To correct this problem, you must check your data files and make sure that the scores or ratings are correctly entered. If problem is due to a formatted file, resave the data file as a unformatted or ASCII file (see your word processing manuals for instructions on how to save a file as unformatted or ASCII). It is permissible to use a data set with alpha or other characters as long as these characters are not read as data and the columns with such characters are skipped by "X" notation in the FORMAT statement.

(c) Other Types of Error Messages. You may encounter other types of errors and error messages that may not be quite clear to you. If that happens simply run the program again. If error persists, call or write to us for help.
12. Example section

The Example subsystem of ITRS displays the following message when chosen:

```
INTERRATER/TEST RELIABILITY SYSTEM
(ITRS)
THE EXAMPLE SUBSYSTEM

There are 11 different examples illustrating the application of ITRS. For each of these examples, there is one data file which could be used as an input file. The following are the names of these files:

(1) For example 1 the file name is: EXAMPLE.E01
(2) For example 2 the file name is: EXAMPLE.E02
(3) For example 3 the file name is: EXAMPLE.E03
(4) For example 4 the file name is: EXAMPLE.E04
(5) For example 5 the file name is: EXAMPLE.E05
(6) For example 6 the file name is: EXAMPLE.E06
(7) For example 7 the file name is: EXAMPLE.E07
(8) For example 8 the file name is: EXAMPLE.E08
(9) For example 9 the file name is: EXAMPLE.E09
(10) For example 10 the file name is: EXAMPLE.E010
(11) For example 11 the file name is: EXAMPLE.E011
```

You can use any of these files to practice running ITRS. To run ITRS with any of the above files, go to the main ITRS menu, select I for examples 1 through 9 and select T for examples 10 and 11. Follow the input sequence of that example which is in your ITRS User's guide. You will see the same results (output) as displayed in the EXAMPLE section of the ITRS User's Guide.
12.1 Example 1

A publishing company advertised a position and 100 people applied for the position. All 100 applicants were asked to write a short essay on "how to improve the quality of printing." The company then asked a group of 8 supervisors to score the essays on an overall "general content quality" category on a 4-point scale, 1 being a poor content quality and 4 being very rich content quality. The company wanted to know how consistent the scores provided by the 8 raters were, and if the rating by the chief supervisor (rater 6) would be sufficient for future hiring.

12.1.1 Example 1 Input

->II (for the interrater subsystem)

INTErrATER/TEST RELIABILITY SYSTEM
(ITRS)

THE INTErrATER SUBSYSTEM

Enter number of Topics/Tasks? 1
Enter number of Raters? 8
Enter "ALL" for estimates based on all raters as a group, "SEL" for selected combinations, and "ALC" for all possible combs

ALL

Enter a FORTRAN-type FORMAT for the data

(8F1.0)

Data input mode: enter "D" for direct input
"F", for input from a file-> F

Enter file name-> EXAMPLE.E01

Please enter the full name or the first two characters of the statistics you wish to be computed. Separate names by a comma (,), and end the list with a period (.), enter "ALL" for all statistics.

PE, PM, IN, KA, WI, AL.
Which rater do you want to compare with the others? 6
Enter the filename on which the output must be saved or enter "LPT1" to direct output to a printer
OUTPUT1.DAT
Print descriptive statistics? (Y/N) Y

12.1.1 Example 1 Output

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>38.00</td>
<td>91.00</td>
<td>61.0</td>
<td>.99</td>
<td>.86</td>
<td>.75</td>
</tr>
</tbody>
</table>

12.1.3 Example 1 Comments

In this example, all indices of interrater reliability are relatively high and indicate high levels of interrater agreement except for the percentage of exact agreement, which is only 38%
The reason for this low percentage is the high number of raters. The higher the number of raters, the more one can expect chance agreement or disagreement to occur.

The range of P.M. correlations between the pairs of raters is high. Correlations range from .61 to .99 with an average of .86. This high range indicates differences between the levels of consistency of the pairs of raters. Some raters are more consistent in their ratings than others. The average P.M. of .86, however, indicates that in most cases pairs of raters are highly consistent in their ratings. In this study, the kappa coefficient, which is an index of agreement of raters that controls for chance agreement, is higher ($k = .69$) than the simple percent of agreement. A z value of 45.75, which is highly significant,
indicates that this statistic is significantly greater than zero. That is, the level of agreement between the raters in the group is significantly greater than pure chance agreement.

The Williams' index, which shows the degree of consistency between the chief supervisor's rating (rater number 6) with others, is very high ($w = .91$) and suggests that the supervisor is very consistent with others in his/her rating. This leads to the conclusion that the supervisor alone could provide reliable ratings for the essays.

Finally, the alpha of .98 indicates a near perfect agreement between raters and shows that there was not much rater effect in this study. In other words, no significant rater dimensionality was observed in this study.
12.2 Example 2

A graduate student researcher in educational measurement was interested in studying reliability of scores in a portfolio assessment. She asked a group of 92 eighth-grade students in three classes to collect a portfolio which consists of samples of their writings in science. The researcher then asked a group of 6 teachers (three math teachers: raters 1, 2, and 3, and three science teachers: raters 4, 5, and 6) to score the students' portfolios on a 5-point scale. By doing this study, the researcher attempted to answer the following questions:

1. In general, how consistent are the 6 raters on their ratings?

2. Would the degree of consistency be different for the two groups of teachers (math teachers versus science teachers)?

3. Would ratings by teacher number 3 (third rater) be consistent with other teachers? If that is the case, then only teacher number 3 would be asked in the future to score the essays.
12.2.1 Example 2 Input

INTEGRATOR/TEST RELIABILITY SYSTEM

(ITRS)

THE INTEGRATOR SUBSYSTEM

Enter number of Topics/Tasks? 1
Enter number of Raters? 6
Enter "ALL" for estimates based on all raters as a group,
"SEL" for selected combinations, and "ALC" for all possible combs

SEL

Enter the combinations of raters for whom you want to estimate interrater reliability

1,2,3/4,5,6/1,2,3,4,5,6.

Enter a FORTRAN-type FORMAT for the data

(6F1.0)

Data input mode: enter "D" for direct input

"F", for input from a file-> F

Enter file name-> EXAMPLE.E02

Please enter the full name or the first two characters of the statistics you wish to be computed. Separate names by a comma (,), and end the list with a period (.), enter "ALL" for all statistics.

PE,PM,IN,KA,WI,AL.

Which rater do you want to compare with the others? 3

Enter the filename on which the output must be saved or enter "LPT1" to direct output to a printer

OUTPUT2.DAT

Print descriptive statistics? (Y/N) Y
12.2.2 Example 2 Output

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2, 3

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>92</td>
<td>70.65</td>
<td>96.74</td>
<td>.87</td>
<td>.98</td>
<td>.92</td>
<td>.89</td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

4, 5, 6

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>92</td>
<td>81.52</td>
<td>96.74</td>
<td>.95</td>
<td>.97</td>
<td>.96</td>
<td>.93</td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2, 3, 4, 5, 6

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>92</td>
<td>68.48</td>
<td>96.74</td>
<td>.87</td>
<td>.98</td>
<td>.94</td>
<td>.90</td>
</tr>
</tbody>
</table>
12.2.3 Example 2 Comments

In this study, six raters scored a science portfolio. Raters 1, 2, and 3 were math teachers and raters 3, 4, and 5 were science teachers. We want to know how consistent all raters are as a group and if rater 3 alone could provide reliable ratings. We also want to know if science teachers as a group provide more consistent ratings than the math teachers as a group. For these analyses, we formed three groups of raters: group one consisted of the math teachers, group two consisted of the science teachers, and group three consisted of all six teachers. The results of analyses for the first group (math teachers) indicate a high level of agreement between the raters. Percent of agreement between the three raters is 70.65%, and percent of agreement within one point is 96.74%. Pearson product moment (P. M.) correlation between pairs of raters ranges from .87 to .98 with an average correlation of .92, indicating a very high level of agreement between the raters. (Recall that a P.M. correlation ranges from -1.0 to +1.0 and a correlation with an absolute value of 1.0 indicates a perfect correlation or agreement.) Interclass correlation, another index of agreement, is also high ($\rho = .89$). Kappa coefficient is at a relatively high level of .74 with a $z$ value of 22.86, which is highly significant and indicates significant agreement above chance level. Furthermore, an alpha of .97 shows a very high level of consistency between raters' rating. The William's index of .77 comparing rater 3 with others in the first group indicates that the rater provided ratings which were consistent with the others. However, a William's index of .77 may not a very high level of consistency.

For the second group of raters (science teachers), all indices of interrater reliability are very high. Comparing the interrater reliability indices for science teachers with those for the math teachers indicates that science teachers are slightly more consistent in their ratings than the math teachers. This slightly higher level of consistency between raters with a science background matches anticipated results because the content of the essays scored by the raters was science.
The results of analyses for the total group of raters also indicate a high level of interrater consistency. However, the percent of exact agreement for the total group (68.48) is considerably lower than this index for the first group (70.65) and for the second group (81.52). This is quite expected because as mentioned earlier, the higher the number of raters the higher the level of chance agreement which effects the size of the percent of agreement. However, the Kappa coefficient which an index of agreement corrected for chance agreement is not much different for the total group than the subgroups. Alpha coefficient for the total group is higher (.99) than for the first group (.97) or for the second group (.98). The reason for this is that the size of alpha increases as the number of raters (or item in a test) increases given that the raters (or items) are unidimensional. Williams' index comparing consistency of rater 3 with others is .93 which is higher than this coefficient obtained for the first subgroup. This coefficient of .93 indicate that rater 3 alone could provide ratings which are highly reliable.
12.3 Example 3

A scoring rubric was developed for scoring essays in history. To test the validity of the rubric, a group of 59 eleventh-grade students were asked to write three essays: one on American history, one on World history, and one on the history of the American Civil War. These essays were scored by 4 raters on a 5-point scale. All raters scored all essays. Therefore, each student received 12 scores, 4 scores for each topic. There were two main hypotheses or research questions related to the validity of the rubric. The first was regarding consistency among raters (interrater reliability). It was hypothesized that the clear instructions given to the raters on the scoring rubric would lead all of the raters to score the essays similarly. That is, the rubric will provide scores with a high level of interrater reliability. The second hypothesis was regarding generalizability over topics/tasks.
12.3.1 Example 3 Input

INTERRATER/TEST RELIABILITY SYSTEM

(ITRS)

THE INTERRATER SUBSYSTEM

Enter number of Topics/Tasks?  3
Enter number of Raters?  4
Enter "ALL" for estimates based on all raters as a group,
"SEL" for selected combinations, and "ALC" for all possible combs

ALL

Enter a FORTRAN-type FORMAT for the data

(12F1.0)

Data input mode: enter "D" for direct input
"F", for input from a file->  F
Enter file name->  EXAMPLE.E03

Please enter the full name or the first two characters of the
statistics you wish to be computed. Separate names by a comma
(,), and end the list with a period (.), enter "ALL" for all
statistics.

PE, PM, IN, KA, GE, AL.

Enter the filename on which the output must be saved
or enter "LPT1" to direct output to a printer

OUTPUT3.DAT
Print descriptive statistics? (Y/N) Y

For combination number: 1
Expand G-Coeff over larger number of topics/raters?
   Enter "Y" for YES or "N" for NO -> Y
Over how many topics/tasks?  5
Over how many raters?  8
12.3.2 Example 3 Output

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2, 3, 4, 5, 6

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
</tr>
<tr>
<td>1</td>
<td>59</td>
<td>64.41</td>
<td>100.0</td>
<td>.90</td>
<td>1.00</td>
<td>.95</td>
<td>.93</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>59.32</td>
<td>100.0</td>
<td>.89</td>
<td>96</td>
<td>.93</td>
<td>.90</td>
</tr>
<tr>
<td>3</td>
<td>59</td>
<td>62.71</td>
<td>100.0</td>
<td>.91</td>
<td>1.00</td>
<td>.95</td>
<td>.92</td>
</tr>
</tbody>
</table>

G-Coefficient for this combination is: .99

Expanding G-Coefficient over 5 topics and 8 raters

<table>
<thead>
<tr>
<th>Topics</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>.94</td>
<td>.97</td>
<td>.98</td>
<td>.98</td>
<td>.98</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.97</td>
<td>.98</td>
<td>.99</td>
</tr>
<tr>
<td>3</td>
<td>.98</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
<td>1.00</td>
<td>.98</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>4</td>
<td>.98</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.98</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>5</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
</tr>
</tbody>
</table>

12 - 12
12.3.3 Example 3 Comments

In our third example, four raters used a newly constructed scoring rubric to score essays on three different topics. We created one group consisting of all 4 raters. We hypothesized that the raters would be consistent due to the instructions they were given prior to rating the essays. We further hypothesized that the scoring rubric could be used on multiple topics. Three lines of statistics were printed, each line corresponding to one of three topics. As it can be seen in the output section, the results of our analyses for all three topics are very similar and consistent. Percent of agreement ranges from 59.32 (for the second topic) to 64.41 (for the first topic), which does not show much variability. Percent of agreement within a one-point range is a perfect 100% for all three topics. For topic two, which has slightly less consistent ratings, the P.M. correlations range from .89 to .96. For the other two topics, the P.M. correlations are even higher, ranging from .90 to 1.0 for the first topic and from .91 to 1.0 for the third topic. Intraclass correlations are extremely high for all three topics, ranging from .90 (for the second topic) to .93 (for the first topic). Similarly, alpha coefficients are very high for the three topics ranging from .98 to .99. The kappa coefficients range from .71 for topic 2 to .77 for topic 1.

Finally, the results show an exceptionally high generalizability coefficient of .99. This near perfect G-coefficient indicates that the rubric provides instructions which leads to consistent ratings and the rubric is generalizable over topics/tasks.

We also asked the system to expand the computed G-coefficient over larger number of topics and raters to see how much improvement can be made by adding to the number of raters and/or to the number of topics. As it can be seen in the output section, we expanded the G-coefficient over 5 topics and 8 raters. Since the G-coefficient was already high, not much improvement was made by increasing number of raters or number of topics. With only one rater and one topic, the G-coefficient was at a very high level of .94. Increasing number of raters or topics could improve the G-coefficient only by a small fraction. With
only 3 topics and 7 raters, the G-coefficient reached at its maximum level of 1.0, a perfect G-coefficient.
12.4 Example 4

To examine the impact of raters' background on their ratings, a set of 60 essays was given to 9 raters. Three of the raters had extensive prior experience in scoring essays (raters 1, 2, and 3), three of them had only one time experience rating essays (4, 5, and 6), and three had no experience in scoring essays (7, 8, and 9). Of the 9 raters, 8 of them were language art teachers and one (rater number 9) was a professor in statistics. Essays were scored using a 5-point scale scheme.
12.4.1 Example 4 Input

INTERRATER/TEST RELIABILITY SYSTEM

(ITRS)

THE INTERRATER SUBSYSTEM

Enter number of Topics/Tasks?  1
Enter number of Raters?  9
Enter "ALL" for estimates based on all raters as a group,
"SEL" for selected combinations, and "ALC" for all possible
combs

SEL
Enter the combinations of raters for whom
you want to estimate interrater reliability
1,2,3/4,5,6/7,8,9/1,2,3,4,5,6,7,8,9.
Enter a FORTRAN-type FORMAT for the data
(9F1.0)
Data input mode: enter "D" for direct input
"F", for input from a file->  F
Enter file name->  EXAMPL.EO4
Please enter the full name or the first two characters of the
statistics you wish to be computed. Separate names by a comma
(,), and end the list with a period (.), enter "ALL" for all
statistics.

PE, PM, IN, KA, WI, AL.
Which rater do you want to compare with the others?  9
Enter the filename on which the output must be saved
or enter "LPT1" to direct output to a printer

OUTPUT4.DAT
Print descriptive statistics? (Y/N) Y
12.4.2 Example 4 Output

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2, 3

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext W. Pnts</td>
<td>Min  Max  Ave</td>
<td>Rho Coef Z</td>
<td>(9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>83.33 100.0</td>
<td>.96 .100 .97</td>
<td>.96 .85 19.7</td>
<td>NA</td>
<td>.99</td>
<td></td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

4, 5, 6

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext W. Pnts</td>
<td>Min  Max  Ave</td>
<td>Rho Coef Z</td>
<td>(9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>68.33 100.0</td>
<td>.92 .96 .93</td>
<td>.90 .72 16.9</td>
<td>NA</td>
<td>.98</td>
<td></td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

7, 8, 9

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext W. Pnts</td>
<td>Min  Max  Ave</td>
<td>Rho Coef Z</td>
<td>(9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>35.00 96.67</td>
<td>.83 .94 .88</td>
<td>.81 .44 10.4</td>
<td>.70</td>
<td>.95</td>
<td></td>
</tr>
</tbody>
</table>

12 - 17
Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2, 3, 4, 5, 6, 7, 8, 9

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>31.67</td>
<td>95.00</td>
<td>.82</td>
<td>1.00</td>
<td>.89</td>
<td>.79</td>
</tr>
</tbody>
</table>

12.4.3 Example 4 Comments

In this example, we want to examine the impact of raters' background on their ratings. Three of the raters in this study had extensive scoring experience (raters 1, 2, and 3), three had some experience (raters 4, 5, and 6), and 3 had no scoring experience (raters 7, 8, and 9). The raters were all language arts teachers except one (rater number 9), who was a university professor. We formed four groups of raters: group one consists of all experienced raters; group two contains raters with some experience; group three contains raters with no scoring experience; and group four consists of all 9 raters. We also want to compare rater 9 (the university professor) with all other raters for consistency.

The results of our analyses clearly show the impact of raters' background (scoring experience) on their ratings. The interrater reliability indices for group one show the highest level of consistency among the other group. Percent of exact agreement is 83.33% and percent of agreement within a one-point range is 100%. P.M. correlations range from .96 to 1.0 with an average of .97, almost a perfect correlation. The interclass correlation is also very high ($p = .96$) and indicates a high level of consistency. Kappa coefficient for this group is .85 with a $z$ of 19.75 indicating agreements significant far beyond chance level. Also, alpha has a value of .99, another index of almost perfect agreement.
For group two (semi-experienced group) the results show a good interrater agreement, but are considerably lower than indices obtained for the first group (the experienced group). The percent of agreement for the semi-experienced group is 68.33% as compared with the respective percentage of 83.33% for the experienced group. P.M. correlations range from .92 to .96 with an average of .94 indicating a high level of consistency between raters. The intraclass correlation of .90 also demonstrates a high level of consistency. Kappa coefficient for this group is .72 as compared with the kappa of .85 for the experienced group. Alpha coefficient is high ($\alpha = .98$).

For the third group (the inexperienced group), the results of our analyses suggest a moderate level of agreement between the raters. The indices of interrater agreement for this group are considerably lower than those for the first group (the experienced group) and even lower than those for the second group (the semi-experienced group). For example, percent of exact agreement for this group is 35% as compared with the percent of agreement of 83.33% for the experienced group and 68.33% for the semi-experienced group. Intraclass correlation for this group is .81 as compared with .96 for the first group and .90 for the second group.

As indicated earlier, kappa is the degree of agreement between raters when controlling for chance agreement. For the group of inexperienced raters, the kappa coefficient is .44 as compared with the kappa of .85 for the experienced group and .72 for the semi-experienced group.

The indices of interrater reliability obtained for the total group indicate high interrater reliability except for the percent of exact agreement. Percent of exact agreement for the total group is 31.67% as compared with 83.33% for the first group and 68.33% for the second group. This is probably due to there being a higher number of raters in the total group. The P.M. correlations range from .82 to 1.0 with an average of .89, intraclass correlation for this group is .79, and alpha is .99. Williams' index is .93 when comparing
rater 9 with others in the total group. This indicates high consistency between rater 9 and other raters as a group. For the total group, kappa has a value of .53 that is lower than most of the other indices and indicates that when chance agreements are controlled for, only a moderate degree of agreement remains.
12.5 Example 5

The literature indicates that training can help raters to provide reliable scores. To verify this finding, 6 social science teachers with similar backgrounds were selected as raters. For 3 of the raters (raters 1, 2, and 4), a two-day training session was held. The other 3 (raters 3, 5, and 6) did not receive any formal training. The raters were asked to score seventy-three geography essays on a 7-point scale. We want to find out if the training indeed has an impact on rater's consistency.

12.5.1 Example 5 Input

INTERRATER/TEST RELIABILITY SYSTEM
(ITRS)

THE INTERRATER SUBSYSTEM

Enter number of Topics/Tasks? 1
Enter number of Raters? 6
Enter "ALL" for estimates based on all raters as a group, "SEL" for selected combinations, and "ALC" for all possible combinations

SEL
Enter the combinations of raters for whom you want to estimate interrater reliability 1, 2, 4/3, 5, 6.
Enter a FORTRAN-type FORMAT for the data (6F1.0)
Data input mode: enter "D" for direct input "F", for input from a file-> F
Enter file name-> EXAMPLE.ES5
Please enter the full name or the first two characters of the statistics you wish to be computed. Separate names by a comma (,), and end the list with a period (.), enter "ALL" for all statistics.

PE, PM, IN, KA, AL.

Enter the filename on which the output must be saved
or enter "LPT1" to direct output to a printer

OUTPUTS.DAT
Print descriptive statistics? (Y/N) Y

12.5.2 Example 5 Output

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2, 4

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>73</td>
<td>91.78</td>
<td>100.0</td>
<td>.99</td>
<td>1.00</td>
<td>.99</td>
<td>.99</td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

3, 5, 6

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>73</td>
<td>34.25</td>
<td>95.89</td>
<td>.91</td>
<td>.93</td>
<td>.92</td>
<td>.89</td>
</tr>
</tbody>
</table>

12 - 22
12.5.3 Example 5 Comments

In this example we wanted to examine the effects of raters' training on their ratings. There were 6 raters in this study, 3 of them (raters 1, 2, and 4) had training, and 3 did not receive any training (raters 3, 5, and 6). We compared the group of raters who received training with the group that did not receive training. We formed two groups of raters: group 1 consisting of the raters who received training (1, 2, and 4); and group 2 had no training (3, 5, and 6). The results of the analyses clearly indicate that training had a significant impact on the consistency of ratings. All indices of interrater agreement are high for the trained group. Percent of exact agreement for this group is 91.78%. P.M. correlations between raters range from .99 to 1.0, almost a perfect agreement. Intraclass correlation is .99, kappa is .93, and alpha is .99. For the untrained group, however, the agreement indices are not as impressive. Percent of exact agreement for this group is 34.25% as compared with 91.78% for the trained group. Intraclass correlation is .89 as compared with .99 for the trained group. The largest difference between the indices can be seen in kappa. For the untrained group, kappa is .43 as compared with kappa of .93 for the trained group. It is interesting to note that the discrepancies between raters due to a lack of training did not have much impact on P.M. correlations and alpha coefficients. The sizes of these indices do not vary in the two groups very much. This may suggest that indices such as kappa or intraclass correlation may be more appropriate for such cases.
12.6 Example 6

Seventy-nine paintings were rated by a group of 5 artists as acceptable (1) or not acceptable (0) to be included in an art exhibition. The manager of the art exhibition can accept the judges' decision only if consistency between any groups of four raters is high (kappa higher than .80). We need to find out if the paintings can be accepted for the exhibition.

12.6.1 Example 6 Input

INTERRATER/TEST RELIABILITY SYSTEM
(ITRS)

THE INTERRATER SUBSYSTEM

Enter number of Topics/Tasks? 1
Enter number of Raters? 5
Enter "ALL" for estimates based on all raters as a group, "SEL" for selected combinations, and "ALC" for all possible combs

ALL

Enter a FORTRAN-type FORMAT for the data (5F1.0)

Data input mode: enter "D" for direct input
"F", for input from a file-> F
Enter file name-> EXAMPLE.EO6

Please enter the full name or the first two characters of the statistics you wish to be computed. Separate names by a comma (,), and end the list with a period (.), enter "ALL" for all statistics.

PE, PM, IN, KA, AL.

Enter the filename on which the output must be saved
or enter "LPT1" to direct output to a printer

OUTPUT6.DAT

Print descriptive statistics? (Y/N) Y
### 12.6.2 Example 6 Output

**Interrater/Test Reliability System (ITRS)**

**Summary of the Results for Raters:**

3, 5, 6

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
<td>Coef</td>
</tr>
<tr>
<td>1</td>
<td>79</td>
<td>62.03</td>
<td>100.0</td>
<td>.53</td>
<td>.85</td>
<td>.63</td>
<td>.41</td>
</tr>
</tbody>
</table>

### 12.6.3 Example 6 Comments

Five judges rated 79 paintings to determine which pieces to include in an art exhibit. The ratings will not be accepted unless the manager is satisfied that the judges have agreement among themselves. To determine the overall agreement, we formed one group which consisted of all the judges. The results of our analyses indicate that most of the interrater reliability indices are low. Percent of exact agreement for this group is 62.03%. Since the ratings are binary (0 or 1), percent of agreement within a one-point range would not be acceptable because it will always equal 1. P.M. correlations range from .53 to .85 with an average of .63. Intraclass correlation has a value of .41 and kappa is .63. These values are all indicative of a moderate level of agreement at best. However, the value for alpha is .90 for this case which again indicates that alpha may not be an appropriate index of agreement for such cases.
12.7 Example 7

A group of 54 mental patients was judged by five psychiatrists to determine if they can be released from a mental hospital. Psychiatrists assigned 1 to a patient if the patient can be released and 0 otherwise. A specific judgment by an individual psychiatrist would be considered if that judgment was consistent with other judgments by the group.

12.7.1 Example 7 Input

INTERRATER/TEST RELIABILITY SYSTEM
(ITRS)

THE INTERRATER SUBSYSTEM

Enter number of Topics/Tasks? 1
Enter number of Raters? 5
Enter "ALL" for estimates based on all raters as a group,
"SEL" for selected combinations, and "ALC" for all possible combs
ALL
Enter a FORTRAN-type FORMAT for the data
(5F1.0)
Data input mode: enter "D" for direct input
"F", for input from a file-> F
Enter file name-> EXAMPLE.E07
Please enter the full name or the first two characters of the statistics you wish to be computed. Separate names by a comma (,), and end the list with a period (.), enter "ALL" for all statistics.
PE,PM,IN,KA,WA,AL.
Which rater do you want to compare with the others? 4
Enter the filename on which the output must be saved or enter "LPT1" to direct output to a printer
OUTPUT7.DAT
Print descriptive statistics? (Y/N) Y
12.7.2 Example 7 Output

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

3, 5, 6

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>54</td>
<td>64.81</td>
<td>100.0</td>
<td>.31</td>
<td>.96</td>
<td>.69</td>
<td>.50</td>
</tr>
</tbody>
</table>

12.7.3 Example 7 Comments

In this case, mental patients can be released from a hospital if there is high level of agreement between five psychiatrists who rate the patient's conditions. For the analyses, we formed one group consisting of all five psychiatrists. The results of these analyses show that only moderate level of agreement exists between the judges (psychiatrists). Percent of exact agreement for this group is 64.81. Percent of agreement within one-point range would not be acceptable because the ratings are binary (1 or 0). P.M. correlations range from .31 to .96 with an average of .69. Intraclass correlation is .50 and kappa is .69. Again, alpha has an unusually high value of .92, which is not consistent with any other index of agreement. We compared rater 4 with the group and a William's index of .64 indicates that rater 4 (or perhaps any other rater) may not be very consistent with the group.
12.8 Example 8

Six judges rated overall performance of 60 basketball players on a scale of 1 to 7.
However, because of time limitations, each judge could rate only 20 players. The
following is the rating plan:
(a) Judges 1 and 2 rated players 1 to 20.
(b) Judges 3 and 4 rated players 21 to 40.
(c) Judges 5 and 6 rated players 41 to 60.
First, we need to use the Missing Data Replacement System to treat the missing data.
Then, we can obtain estimates of inter judge consistency.

12.8.1 Example 8 Input

THE INTERRATER SUBSYSTEM
(ITRS)

THE MISSING DATA REPLACEMENT SUBSYSTEM

Please input the data-file name -> EXAMPL.EO8
What number represents missing data? 0
Replace or delete missing values? (R/D) D
The missing data replacement/deletion process is complete.
You can now type ITRS to go back to the main ITRS menu and
select Interrater or Test option
Enter number of Topics/Tasks?  1
Enter number of Raters?  6
Enter "ALL" for estimates based on all raters as a group, "SEL" for selected combinations, and "ALC" for all possible combs

SEL
Enter the combinations of raters for whom you want to estimate interrater reliability 1,2/3,4/5,6.
Enter a FORTRAN-type FORMAT for the data (6F1.0)
Data input mode: enter "D" for direct input "F", for input from a file->  F
Enter file name->  EXAMPL.E08
Please enter the full name or the first two characters of the statistics you wish to be computed. Separate names by a comma (,), and end the list with a period (.), enter "ALL" for all statistics.
PE,PM,IN,KA,AL.
Enter the filename on which the output must be saved or enter "LPT1" to direct output to a printer
OUTPUT8.DAT
Print descriptive statistics? (Y/N) Y
12.8.2 Example 8 Output

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:
1, 2

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext W. Pnts</td>
<td>Min Max Ave Rho Coef Z (0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>70.0 100.0</td>
<td>.96 .96 .96 .95 .65 6.77 NA</td>
<td></td>
<td></td>
<td></td>
<td>.98</td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:
3, 4

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext W. Pnts</td>
<td>Min Max Ave Rho Coef Z (0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>90.0 100.0</td>
<td>.99 .99 .99 .99 .88 8.65 NA</td>
<td></td>
<td></td>
<td></td>
<td>.99</td>
</tr>
</tbody>
</table>

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:
5, 6

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext W. Pnts</td>
<td>Min Max Ave Rho Coef Z (0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>60.0 95.00</td>
<td>.94 .94 .94 .93 .51 4.97 NA</td>
<td></td>
<td></td>
<td></td>
<td>.96</td>
</tr>
</tbody>
</table>
12.8.3 Example 8 Comments

Six judges rated the overall performance of the players but because each judge could not rate every player, the ratings are incomplete. Due to the incomplete data, we cannot obtain indices of agreement among all raters as a group. Since we can only group those raters who rated the same players, we must create three groups of raters. Group 1 should consist of raters who rated players 1 to 20, group 2 would be ratings for players 21 to 40, and group 3 would be ratings for players 41 to 60. We had to run the missing data program to indicate that we had missing data in this case, and that zero represents the missing values. The results of the analyses across the three groups may not be quite comparable because both the raters and the subjects in the three groups are different.

The analyses of ratings within the three groups indicate a high level of agreement between the judges. For the first group, the percent of exact agreement is 90%, P.M. is .99, intraclass correlation is .99, and alpha is .99. For the second and third groups similar results were found. The results indicated that judges in groups of two are consistent in their ratings. The percent of exact agreement for the first and third group is low compared to the other statistics for their groups.
12.9 Example 9

Eighty-two multiple-choice items in geometry for 12th grade students were inspected by four geometry experts as either acceptable (1) or not acceptable (0) to be included in the item pool for 12th grade geometry. We want to analyze the degree of consistency between the judges' decision. In particular, we need to know if any of the four judges provided ratings that were substantially different from the other three.

12.9.1 Example 9 Input

INTERRATER/TEST RELIABILITY SYSTEM
(ITRS)

THE INTERRATER SUBSYSTEM

Enter number of Topics/Tasks? 1
Enter number of Raters? 4
Enter "ALL" for estimates based on all raters as a group, "SEL" for selected combinations, and "ALC" for all possible combs ALL
Enter a FORTRAN-type FORMAT for the data (4F1.0)
Data input mode: enter "D" for direct input
"F", for input from a file-> F
Enter file name-> EXAMPLE.EO9
Please enter the full name or the first two characters of the statistics you wish to be computed. Separate names by a comma (,), and end the list with a period (.), enter "ALL" for all statistics.
PE, PM, IN, KA, WI, AL.
Which rater do you want to compare with the others? 1
Enter the filename on which the output must be saved or enter "LPT1" to direct output to a printer
OUTPUT9.DAT
Print descriptive statistics? (Y/N) Y
12.9.2 Example 9 Output

Interrater/Test Reliability System (ITRS)

Summary of the Results for Raters:

1, 2, 3, 4

<table>
<thead>
<tr>
<th>Topic</th>
<th>N</th>
<th>PC Agreement</th>
<th>P.M. Corr</th>
<th>Intraclass</th>
<th>Kappa</th>
<th>William</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ext</td>
<td>W. Pnts</td>
<td>Min</td>
<td>Max</td>
<td>Ave</td>
<td>Rho</td>
</tr>
<tr>
<td>1</td>
<td>82</td>
<td>65.85</td>
<td>100.0</td>
<td>.41</td>
<td>.90</td>
<td>.68</td>
<td>.47</td>
</tr>
</tbody>
</table>

12.9.3 Example 9 Comments

A set of 82 multiple-choice items were rated by 4 content experts as acceptable (1) and not acceptable (0). We analyzed the consistency of the raters, and also determined if only one of the raters could do the job. We formed a group consisting of all four raters. We used Williams’ index to compare one of the raters with the others. The results of the analyses showed that some of raters are highly consistent with each other but some are not. P.M. correlations ranged from .41 to .90, which indicates a high level of discrepancy between raters’ consistency. The intraclass correlation of .47 and a kappa of .63 indicates a moderate overall consistency between the raters. The Williams’ index of .79, comparing rater 1 with the group, indicates a moderate level of consistency between this particular rater and the other raters.
12.10 Example 10

A 40-item multiple choice test in math was administered to a group of 60 tenth-grade students. We need to compute the mean and standard deviation for each item as well as correlations of the items with the total test score. We also want to estimate the reliability of test taking out one item at a time. Finally, we want to obtain an overall estimate of the test reliability.

12.10.1 Example 10 Input

INTERRATER/TEST RELIABILITY SYSTEM
(ITRS)
TEST SUBSYSTEM

Enter number of items: 40
Enter number of subscales: 1
Enter a FORTRAN-type FORMAT for the data
(40F1.0)
Data input mode: enter "D" for direct input,
"F" for input from a file: F
Enter the file name: EXAMPL.E10
Enter the filename on which the output must be saved
or enter "LPT1" to direct output to the printer
OUTPUT10.DAT
Do you need item-total statistics? (Y/N) Y
Please enter "M" for Multiple-choice, "L" for Likert-type
test:
M
Please input KEY responses for scoring test items, start with
column1, you must enter 40 numbers. Enter 1 for "A", 2 for
"B", 3 for "C", etc.
2234313343412341213143123133212443312122
Print item statistics? (Y/N) Y
How many valid option each item has? 4
Stop - Program terminated.
12.10.2 Example 10 Output

INTERRATER/TEST RELIABILITY SYSTEM
(ITRS)
TEST SUBSYSTEM

ITEM STATISTICS

<table>
<thead>
<tr>
<th>Item</th>
<th>MEAN</th>
<th>S.D.</th>
<th>FREQUENCIES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OPTN 1</td>
</tr>
<tr>
<td>1</td>
<td>.62</td>
<td>.49</td>
<td>21</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.32</td>
<td>.47</td>
<td>32</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.70</td>
<td>.46</td>
<td>4</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.67</td>
<td>.47</td>
<td>8</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.22</td>
<td>.41</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>.45</td>
<td>.50</td>
<td>27</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.18</td>
<td>.39</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>.27</td>
<td>.44</td>
<td>26</td>
</tr>
<tr>
<td>9</td>
<td>.22</td>
<td>.41</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>.30</td>
<td>.46</td>
<td>4</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>.50</td>
<td>.50</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>.35</td>
<td>.48</td>
<td>21</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>.20</td>
<td>.40</td>
<td>21</td>
</tr>
<tr>
<td>14</td>
<td>.20</td>
<td>.40</td>
<td>19</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>.28</td>
<td>.45</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>.15</td>
<td>.36</td>
<td>9</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>.05</td>
<td>.22</td>
<td>28</td>
</tr>
<tr>
<td>18</td>
<td>.08</td>
<td>.28</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>.22</td>
<td>.41</td>
<td>18</td>
</tr>
<tr>
<td>20</td>
<td>.25</td>
<td>.43</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>.22</td>
<td>.41</td>
<td>11</td>
</tr>
<tr>
<td>22</td>
<td>.22</td>
<td>.41</td>
<td>8</td>
</tr>
<tr>
<td>23</td>
<td>.00</td>
<td>.00</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>.08</td>
<td>.28</td>
<td>44</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>.22</td>
<td>.41</td>
<td>7</td>
</tr>
<tr>
<td>26</td>
<td>.13</td>
<td>.34</td>
<td>8</td>
</tr>
<tr>
<td>27</td>
<td>.12</td>
<td>.32</td>
<td>14</td>
</tr>
<tr>
<td>28</td>
<td>.33</td>
<td>.47</td>
<td>17</td>
</tr>
<tr>
<td>29</td>
<td>.15</td>
<td>.36</td>
<td>10</td>
</tr>
<tr>
<td>30</td>
<td>.23</td>
<td>.42</td>
<td>14</td>
</tr>
<tr>
<td>31</td>
<td>.23</td>
<td>.42</td>
<td>30</td>
</tr>
<tr>
<td>CAUTION!!! UNPROPORTIONAL OPTION FREQUENCIES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>.27</td>
<td>.44</td>
<td>14</td>
</tr>
</tbody>
</table>

12 - 35
<table>
<thead>
<tr>
<th>Item</th>
<th>Scale Mean If Item Deleted</th>
<th>Scale Var If Item Deleted</th>
<th>Corrected Item-Tot Correlation</th>
<th>Alpha If Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.75</td>
<td>5.99</td>
<td>-.07</td>
<td>-.11</td>
</tr>
<tr>
<td>2</td>
<td>10.05</td>
<td>6.35</td>
<td>-.21</td>
<td>-.05</td>
</tr>
<tr>
<td>3</td>
<td>9.67</td>
<td>6.46</td>
<td>-.26</td>
<td>-.03</td>
</tr>
<tr>
<td>4</td>
<td>9.70</td>
<td>6.04</td>
<td>-.09</td>
<td>-.10</td>
</tr>
<tr>
<td>5</td>
<td>10.15</td>
<td>5.96</td>
<td>-.03</td>
<td>-.12</td>
</tr>
<tr>
<td>6</td>
<td>9.92</td>
<td>5.91</td>
<td>-.22</td>
<td>-.23</td>
</tr>
<tr>
<td>7</td>
<td>10.10</td>
<td>6.36</td>
<td>-.22</td>
<td>-.05</td>
</tr>
<tr>
<td>8</td>
<td>10.15</td>
<td>6.94</td>
<td>-.48</td>
<td>.04</td>
</tr>
<tr>
<td>9</td>
<td>10.07</td>
<td>5.42</td>
<td>.20</td>
<td>-.24</td>
</tr>
<tr>
<td>10</td>
<td>9.87</td>
<td>5.37</td>
<td>.19</td>
<td>-.24</td>
</tr>
<tr>
<td>11</td>
<td>10.02</td>
<td>5.17</td>
<td>.30</td>
<td>-.29</td>
</tr>
<tr>
<td>12</td>
<td>10.17</td>
<td>5.46</td>
<td>.23</td>
<td>-.23</td>
</tr>
<tr>
<td>13</td>
<td>10.17</td>
<td>5.33</td>
<td>.31</td>
<td>-.27</td>
</tr>
<tr>
<td>14</td>
<td>10.08</td>
<td>5.50</td>
<td>.17</td>
<td>-.22</td>
</tr>
<tr>
<td>15</td>
<td>10.22</td>
<td>5.60</td>
<td>.20</td>
<td>-.21</td>
</tr>
<tr>
<td>16</td>
<td>10.32</td>
<td>6.36</td>
<td>-.30</td>
<td>-.08</td>
</tr>
<tr>
<td>17</td>
<td>10.28</td>
<td>6.38</td>
<td>-.27</td>
<td>-.07</td>
</tr>
<tr>
<td>18</td>
<td>10.15</td>
<td>5.59</td>
<td>.16</td>
<td>-.20</td>
</tr>
<tr>
<td>19</td>
<td>10.12</td>
<td>5.73</td>
<td>.07</td>
<td>-.17</td>
</tr>
<tr>
<td>20</td>
<td>10.15</td>
<td>6.16</td>
<td>-.13</td>
<td>-.09</td>
</tr>
<tr>
<td>21</td>
<td>10.15</td>
<td>6.43</td>
<td>-.26</td>
<td>-.04</td>
</tr>
<tr>
<td>22</td>
<td>10.37</td>
<td>6.07</td>
<td>.00</td>
<td>-.14</td>
</tr>
<tr>
<td>23</td>
<td>10.28</td>
<td>6.38</td>
<td>-.27</td>
<td>-.07</td>
</tr>
<tr>
<td>24</td>
<td>10.15</td>
<td>5.45</td>
<td>.23</td>
<td>-.24</td>
</tr>
<tr>
<td>25</td>
<td>10.23</td>
<td>7.00</td>
<td>-.58</td>
<td>.04</td>
</tr>
<tr>
<td>26</td>
<td>10.25</td>
<td>6.12</td>
<td>-.10</td>
<td>-.11</td>
</tr>
<tr>
<td>27</td>
<td>10.03</td>
<td>6.17</td>
<td>-.14</td>
<td>-.08</td>
</tr>
<tr>
<td>28</td>
<td>10.22</td>
<td>5.53</td>
<td>.24</td>
<td>-.23</td>
</tr>
<tr>
<td>29</td>
<td>10.13</td>
<td>5.13</td>
<td>.39</td>
<td>-.31</td>
</tr>
<tr>
<td>30</td>
<td>10.13</td>
<td>5.61</td>
<td>.14</td>
<td>-.20</td>
</tr>
<tr>
<td>31</td>
<td>10.10</td>
<td>6.36</td>
<td>-.22</td>
<td>-.05</td>
</tr>
<tr>
<td>32</td>
<td>10.03</td>
<td>6.34</td>
<td>-.21</td>
<td>-.05</td>
</tr>
<tr>
<td>33</td>
<td>10.27</td>
<td>5.89</td>
<td>.06</td>
<td>-.15</td>
</tr>
<tr>
<td>34</td>
<td>10.00</td>
<td>6.03</td>
<td>-.09</td>
<td>-.10</td>
</tr>
<tr>
<td>35</td>
<td>10.15</td>
<td>6.50</td>
<td>-.29</td>
<td>-.03</td>
</tr>
<tr>
<td>36</td>
<td>10.23</td>
<td>6.45</td>
<td>-.29</td>
<td>-.05</td>
</tr>
<tr>
<td>37</td>
<td>10.08</td>
<td>4.93</td>
<td>.46</td>
<td>-.36</td>
</tr>
<tr>
<td>38</td>
<td>10.10</td>
<td>6.02</td>
<td>-.07</td>
<td>-.11</td>
</tr>
<tr>
<td>39</td>
<td>10.12</td>
<td>5.39</td>
<td>.24</td>
<td>-.25</td>
</tr>
</tbody>
</table>

TOTAL TEST: -.14
12.10.3 Example 10 Comments

Examples 10 and 11 illustrate the application of the ITRS on test items. In example 10, we have a 40-item multiple choice test and we want to compute descriptive statistics for the items as well as reliability coefficients for the total test. We asked for a print out of the frequencies of the responses to make sure that the responses are distributed fairly across the distracters.

The output shows mean (proportion of correct responses) and standard deviations for each of the 40 items. The output also displays the frequencies of each of the four options in the items and flags a warning when the responses across the distractors are not distributed fairly. For example, the mean for item 1 is .62, the standard deviation is .49, and the frequencies to options 1 to 4 are: 21, 37, 2, and 0 respectively. The correct option (key) for this item is option number 2 which has a frequency of 37.

Since we have 60 subjects answered this test and 37 of them answered this item correctly, the remaining 23 subjects selected incorrect options. Out of this 23, 21, or 91%, of incorrect answers were on option 1, only 2 were on option 3, and none were on option 4. That is, options 3 and 4 were recognized as obviously wrong by the subjects and option 1 seemed correct to many of the subjects. This means that options 2 and 4 did not serve as good distractors. That is why a warning was presented.

The next part of the output is the reliability section of the analysis. For each item, it contains "scale mean", "scale variance", "alpha if item is deleted", and "corrected item total correlations". These statistics help test constructors to identify problematic items.
12.11 Example 11

A 30 Likert-type-items test (5-point scale) was given to a group of 50 college students. This test had three different subscales with 10 items in each subscale. We want to: (1) compute the mean and standard deviation for each of the items; (2) compute correlations between each item score and the total test score; (3) determine the reliability of each subscale; (4) determine the reliability of each subscale when an individual item is removed; and (5) determine the reliability of the test with all 30 items.

12.11.1 Example 11 Input

```
INTERRATER/TEST RELIABILITY SYSTEM
(ITRS)
TEST SUBSYSTEM

Enter number of items -> 30
Enter number of subscales -> 3

Enter a FORTRAN-type FORMAT for the data
(30F1.0)

Data input mode: enter "D" for direct input,
"F" for input from a file -> F

Enter the file name -> EXAMPL.E11

Enter the filename on which the output must be saved
or enter "LPT1" to direct output to the printer

OUTPUT11.DAT

Do you need item-total statistics? (Y/N) Y
Please enter "M" for multiple choice and "L" for Likert-type test:
L

Please enter item numbers for each subscale. Separate item
numbers by a comma, and end each line with a period (.).
Remember, you must have one line of item numbers per subscale
1,2,3,4,5,6,7,8,9,10.
11,12,13,14,15,16,17,18,19,20.
21,22,23,24,25,26,27,28,29,30.

Stop - Program terminated.
```
### 12.11.2 Example 11 Output

<table>
<thead>
<tr>
<th>Item</th>
<th>Scale Mean If Item Deleted</th>
<th>Scale Var If Item Deleted</th>
<th>Corrected Item-Tot Correlation</th>
<th>Alpha If Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary statistics for subscale 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>22.74</td>
<td>102.40</td>
<td>.94</td>
<td>.98</td>
</tr>
<tr>
<td>2</td>
<td>22.94</td>
<td>103.81</td>
<td>.94</td>
<td>.98</td>
</tr>
<tr>
<td>3</td>
<td>22.80</td>
<td>100.45</td>
<td>.95</td>
<td>.98</td>
</tr>
<tr>
<td>4</td>
<td>22.94</td>
<td>102.55</td>
<td>.94</td>
<td>.98</td>
</tr>
<tr>
<td>5</td>
<td>22.94</td>
<td>104.55</td>
<td>.87</td>
<td>.98</td>
</tr>
<tr>
<td>6</td>
<td>22.90</td>
<td>100.58</td>
<td>.59</td>
<td>.98</td>
</tr>
<tr>
<td>7</td>
<td>22.88</td>
<td>103.66</td>
<td>.90</td>
<td>.98</td>
</tr>
<tr>
<td>8</td>
<td>22.78</td>
<td>100.79</td>
<td>.98</td>
<td>.98</td>
</tr>
<tr>
<td>9</td>
<td>22.84</td>
<td>101.40</td>
<td>.92</td>
<td>.98</td>
</tr>
<tr>
<td>10</td>
<td>22.84</td>
<td>104.71</td>
<td>.90</td>
<td>.98</td>
</tr>
<tr>
<td><strong>Subscale total:</strong></td>
<td></td>
<td></td>
<td></td>
<td>.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Scale Mean If Item Deleted</th>
<th>Scale Var If Item Deleted</th>
<th>Corrected Item-Tot Correlation</th>
<th>Alpha If Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary statistics for subscale 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>24.26</td>
<td>85.54</td>
<td>.84</td>
<td>.98</td>
</tr>
<tr>
<td>12</td>
<td>24.34</td>
<td>82.15</td>
<td>.92</td>
<td>.98</td>
</tr>
<tr>
<td>13</td>
<td>24.38</td>
<td>82.09</td>
<td>.96</td>
<td>.97</td>
</tr>
<tr>
<td>14</td>
<td>24.54</td>
<td>84.54</td>
<td>.91</td>
<td>.98</td>
</tr>
<tr>
<td>15</td>
<td>24.30</td>
<td>86.42</td>
<td>.83</td>
<td>.98</td>
</tr>
<tr>
<td>16</td>
<td>24.44</td>
<td>89.84</td>
<td>.80</td>
<td>.98</td>
</tr>
<tr>
<td>17</td>
<td>24.31</td>
<td>89.36</td>
<td>.89</td>
<td>.98</td>
</tr>
<tr>
<td>18</td>
<td>24.22</td>
<td>85.15</td>
<td>.97</td>
<td>.97</td>
</tr>
<tr>
<td>19</td>
<td>24.02</td>
<td>85.12</td>
<td>.89</td>
<td>.98</td>
</tr>
<tr>
<td>20</td>
<td>24.36</td>
<td>82.52</td>
<td>.97</td>
<td>.97</td>
</tr>
<tr>
<td><strong>Subscale total:</strong></td>
<td></td>
<td></td>
<td></td>
<td>.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Scale Mean If Item Deleted</th>
<th>Scale Var If Item Deleted</th>
<th>Corrected Item-Tot Correlation</th>
<th>Alpha If Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary statistics for subscale 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>28.20</td>
<td>103.47</td>
<td>.87</td>
<td>.98</td>
</tr>
<tr>
<td>22</td>
<td>28.18</td>
<td>108.15</td>
<td>.86</td>
<td>.98</td>
</tr>
<tr>
<td>23</td>
<td>28.20</td>
<td>104.41</td>
<td>.83</td>
<td>.98</td>
</tr>
<tr>
<td>24</td>
<td>28.38</td>
<td>103.36</td>
<td>.87</td>
<td>.98</td>
</tr>
<tr>
<td>25</td>
<td>28.10</td>
<td>105.48</td>
<td>.93</td>
<td>.98</td>
</tr>
<tr>
<td>26</td>
<td>28.20</td>
<td>105.10</td>
<td>.84</td>
<td>.98</td>
</tr>
<tr>
<td>27</td>
<td>28.32</td>
<td>105.00</td>
<td>.92</td>
<td>.98</td>
</tr>
<tr>
<td>28</td>
<td>28.44</td>
<td>99.39</td>
<td>.94</td>
<td>.97</td>
</tr>
<tr>
<td>29</td>
<td>28.42</td>
<td>99.68</td>
<td>.95</td>
<td>.97</td>
</tr>
<tr>
<td>30</td>
<td>28.34</td>
<td>98.43</td>
<td>.98</td>
<td>.97</td>
</tr>
<tr>
<td><strong>Subscale total:</strong></td>
<td></td>
<td></td>
<td></td>
<td>.98</td>
</tr>
<tr>
<td><strong>TOTAL TEST:</strong></td>
<td></td>
<td></td>
<td></td>
<td>.99</td>
</tr>
</tbody>
</table>
12.11.3 Example 11 Comments

In this example, we have illustrated the application of ITRS on tests with Likert-type items. A test with 30-Likert-type items was administered to a group of 50 students. We wanted to obtain descriptive statistics for the item scores as well as for the total test score. The test has three subscales. We will obtain the relevant information for the subscales as well.

The results show scale mean, scale standard deviation, item-total correlation, and alpha for each item in each subscale. The scale mean is the mean of the subscale in which the item belongs when that item is removed from the subscale. Similarly, scale S.D. and alpha are the scale statistics when the item is removed from the scale. Item-total correlation is the correlation between individual items with the total scale score, and alpha for the total test is the reliability index for the 30-item test. As the results indicate, alpha coefficients for the subscale as well as alpha for the total test are exceptionally high (α = .99 for the first subscale, α = .98 for the second subscale, α = .98 for the third subscale, and α = .99 for the entire test.)
13. Appendix A, Literature Review

According to the literature, interrater reliabilities for performance assessment have usually been estimated by applying Cronbach's alpha (Atkinson and Murray, 1987; Lehmann, 1990; Scherer and McKee, 1992;) or by computing a Product-Moment (PM) correlation between the pairs of scores of the raters (Bers and Smith, 1990; Crews, 1991; Lundberg, Westermark, and Rasch, 1990; Magnan, 1987; Root, 1987; Rosenthal, DeMers, Stilwell, Graybeal, and Zins, 1983; Rothstein, 1990; Shechtman, 1992). Other statistical techniques for estimating interrater reliability, such as percent of exact agreement (Bunch and Littlefair, 1988; Crews, 1991; Gay, 1981; Kaplan and Johnson, 1992), Goodman-Kruskal, Kendall's Tau (Magnan, 1987), Williams' index of agreement (Posner, Sampson, Caplan, Ward, and Cheney, 1990), various forms of intraclass correlation (Bartko, 1976; Bloch, 1989; Collis, 1985; Kaplan and Johnson, 1992; O'Hara and Rehm, 1983), and different forms of kappa coefficient (Cohen, 1960, 1968; Collis, 1985; Fleiss, 1971; Fleiss, Cohen, and Everitt, 1969; Fleiss, Nee, and Landis, 1979; Kaplan and Johnson, 1992), have also been reported (see also Posner et al., 1990). The literature consistently reports that Cronbach's alpha is sensitive to the number and dimensionality of raters/items (see for example, Abedi, Baker, and Herl, 1993; Cortina, 1993). Moreover, the percent of agreement index may underestimate the agreement between scores because it does not take into account the random variations in scores provided by the raters. The literature has also reported that PM correlation usually overestimates the interrater reliability when there is a systematic error between the scores (Bartko, 1976; Crews, 1991; Gay, 1981), and kappa can be inflated or deflated by changing the number of categories (Kraemer, Bliwise, and Bliwise, 1991). Thus, various statistical techniques yield different estimates of reliability coefficients not only due to the nature of the statistics employed, but also because of differences on the level of systematic or random error in rater agreement, the type of distribution of scores, the
number of subjects, and the number of raters especially when more than two raters are involved. Access to different estimates of interrater reliability would help researchers/practitioners to select the one that would be more appropriate for their case.
14. References


