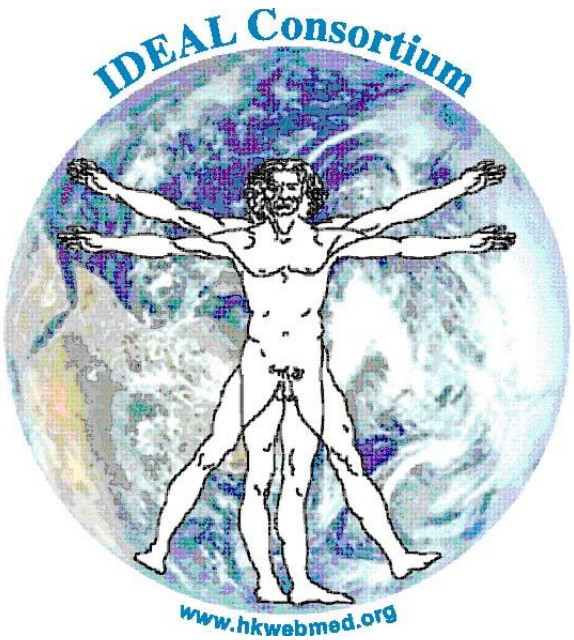
IDEAL Item Analysis Users' Guide
for
Selected and Constructed Item Formats

2nd Edition



IDEAL - HK™

International Database for Enhanced Assessments & Learning

D. Precht, C. Hazlett, S. Yip, J. Nicholls & Y. Jin

The IDEAL Consortium and its member institutions assume
no responsibility for any errors of omission or commission for any
information contained within this user guide and the IDEAL item analysis program.
Any and all responsibility stemming from its use shall be borne by the user.

October 2008

Printed in Hong Kong by Condor Production Ltd
2203 Hong Man Industrial Centre, 2 Hong Man Street, Chai Wan, Hong Kong

Acknowledgement

The item analysis program, originally developed by personnel at the University of Alberta and at the University of British Columbia, was redesigned into a windows-based application specifically for use by the IDEAL system. Mr. Daniel Precht, one of original authors at the University of Alberta, performed this work under contract with the WebMED project and the IDEAL Consortium. The WebMED project was supported by funding from the University Grants Committee, Hong Kong Government.

IDEAL Item Analysis Users' Guide

Table of Contents

Page

Chapter 3 Item Analysis for Constructed Response & Multiple Selected Response Items

Chapter 1

Introduction to IDEAL's Item Analysis Program (IA)

1.1 Accessing the IA Program

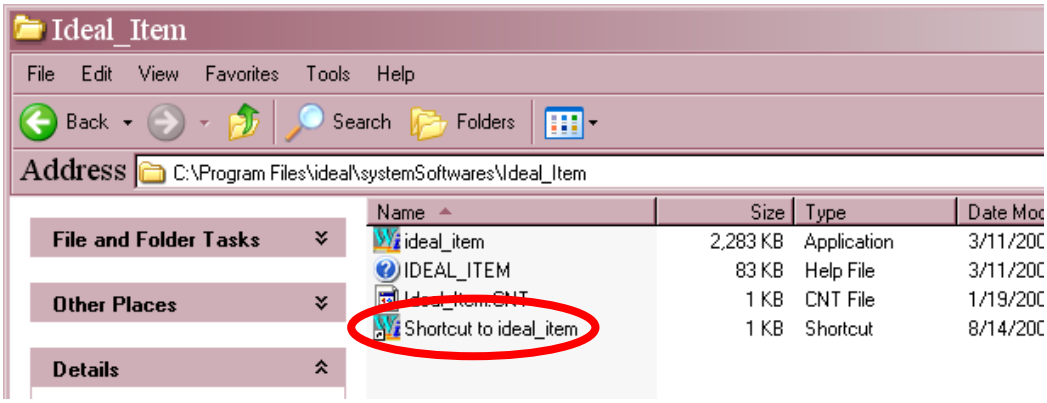
IDEAL's item analysis program (hereafter referred to IA) is included with IDEAL (4.1) PC and Intranet programs. After installation of either or both of these two IDEAL programs, the item analysis application can be found in the following respective directories:

(1) for IDEAL PC, the location is:
C:\Program Files\ideal\systemSoftwares\Ideal_Item

(2) for IDEAL Intranet, the location is:
C:\Program Files\IidealSetup\Ideal_Item

The first directory will be on one's PC, and the second will be on the server. *Right* click the application (note first file in the following illustration). Create a shortcut (circled in the illustration). Drag and drop the shortcut onto your desktop.

IA is a standalone application. If you wish to use IA independently of the main IDEAL programs, copy all files noted in the following illustration to a folder on any PC (or server) and use IA in the same manner as described in the following chapters.



1.2 Purpose and Options

IA computes and prints test and item statistical properties based on models drawn from classical test theory. It is a powerful and uniquely flexible program. IA can generate relevant psychometrics for (i) all selected response items that are usually scored in a binary fashion (e.g., answers for A-type and R-type MCQs are either right or wrong), (ii) all forms of constructed response assessments that are usually scored as continuous data (e.g., Modified Essays Questions (MEQ), Short Answer Questions (SAQ), Key Features, OSCEs and OSPEs), and (iii) any selected response items that are also scored from 0 to n options for each item (e.g., Pick n of N).

IA is particularly useful for supporting one's IDEAL item banks. IA generates a soft copy file of the psychometric properties for items used in an assessment. This file is readable by IDEAL PC or IDEAL Intranet and used for automatically updating the banked items' statistical properties after each time these items are administered in an assessment.

1.3 Overview of Manual

Chapter Two explains how to use IA to generate psychometric properties for any assessments that are scored dichotomously, such as A-type or R-type (extended matching) items. For these assessments, candidates usually enter their answers on optically scored sheets and an optical scanner is used to mark the candidates answers. The scanner can also reproduce the candidates' answers in the form of a data (*dat) or a text (*.txt) file. IA can use this latter file to compute the psychometric properties of the whole test and for each item in the assessment. However, if the user cannot optically score the items, a file similar to that produced by a scanner will need to be created by using MS Excel, Windows Notepad or MS Word.

Chapter Two is organized assuming one has an optical scanner (a common arrangement in medical education units). However, an addendum to the chapter describes the additional required steps if one does not have access to a scanner.

Chapter Three explains how to use IA for items scored in a non-dichotomous fashion. The candidates' scores on these assessments can range from 0 to any value including part scores. Normally these assessments have to be marked manually. To use IA, one needs to enter the candidates' scores into a spreadsheet using MS Excel.

In addition to describing the procedures required to run the item analysis program for each of the above types of assessments, the nature and use of the statistics generated by IA are explained in each chapter. These psychometrics assist a test committee in making more informed judgments concerning the quality of an examination.

Finally, the steps required to upload the generated psychometrics into IDEAL's item banks are also described in this manual.

Members in the IDEAL Consortium are invited to recommend any additional features that they might wish to be added to IA. The IDEAL Consortium's goal is to facilitate as much as possible the assessment needs of each of its members.

Chapter Two

Item Analysis for Dichotomously Scored (e.g., A-type, R-type) Items

2.1 Prepare an Appropriate Data File

2.1.1 Introduction

To run the item analysis application, the user must have a file containing the key (correct answers) and the responses that a group of candidates gave as their answers to selected-response questions (such as A- and/or R-Type items).

How to prepare this data file will depend on whether you administered the assessment in a paper and pencil format or via the Web using IDEAL Intranet. Section 2.1.2 describes how to prepare the data file if the assessment is administered in hard copy. Section 2.1.3 describes how to access the data that IDEAL Intranet generates automatically after a Web-based assessment is completed.

2.1.2 Paper & Pencil (Hardcopy) Tests (IDEAL PC)

The data file must include the key (correct answers) and the candidates' responses to all the items to be analyzed.

1. An optical scanner produces this file at the time the candidates' responses are optically scored. Save this file in the same directory as the IA application, or a subdirectory of one's choice. Save the file as a data file (i.e., with the extension *.dat). To read or edit this file, use IA or Windows' Notepad.

2. However, if one cannot optically score the test, enter the key and candidates' responses into a spreadsheet using Microsoft Excel and save the file with the extension *.csv (i.e., a comma delimited file and not with the default extension of *.*xls*). Alternatively, one may also use Windows Notepad or MS Word and save the file as a data (*.dat) file. Refer to the Addendum at the end of Chapter Two for further details.

3. The file, generated by an optical scanner, is constructed similar to a spreadsheet, but it must meet the following specifications:

    a. 1st line specifies the key (list of correct answers for the items to be analyzed);
    b. each subsequent line (row) corresponds to a different candidate's responses;
    c. each character space (column) in a row corresponds to each of the keyed answers and the corresponding candidates' responses to each item in the examination; and
    d. answers can be coded numerically (e.g., 1 to 9), but if the maximum number of choices for any item exceeds one digit, the answers and candidates' responses must be coded as alphabetic characters.

The output file from optical scoring normally includes the candidates' IDs. With the exception of the key, the scanner usually reproduces each ID at the beginning of each row of data. However, if the scanner reproduces the IDs in another location (e.g., after the candidates' answers), IA can accommodate this variation.

Save this file of answers and candidates' responses in either the Ideal_Item subdirectory[1] or a directory of one's own choosing [such as default directory for tests generated by IDEAL, i.e., C:\Program Files\ideal\cohortDatabase *(or another database being used)*\ExamPapers]. Use a filename that can be easily matched with the name of the file assigned when the

[1] For IDEAL PC, the default location is: C:\Programe Files\ideal\systemSoftwares\Ideal_Item.
For IDEAL Intranet, the default location is: C:\Program Files\IidealSetup\Ideal_Item.

items were selected from the IDEAL bank *and printed* (see IDEAL PC manual, Section 5.1 and entry for "Course/Module No." in the "Print Exam" window). Assigning a filename that can be always recognized in terms how it was used in assessing candidates is recommended (e.g., *answers cardio yr 2 2008.dat*).

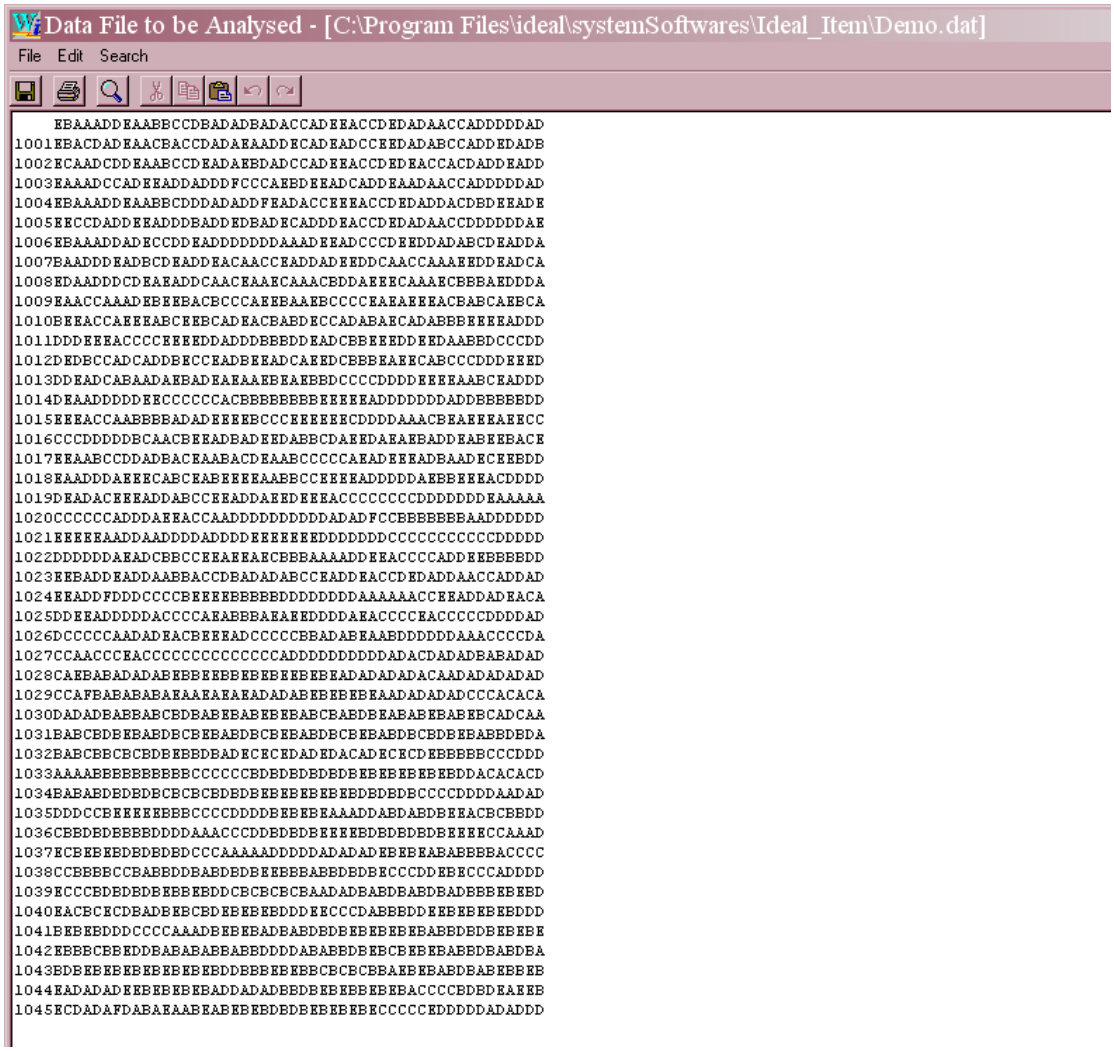
The user may not wish to use the default option of the candidate's total test score as an overall measure of performance (for purposes of calculating unadjusted and adjusted biserial and point biserial coefficients). In lieu, a reference standard for each candidate can be used and is entered immediately following the candidate's responses; a user may enter these data elsewhere in the record but will need to ensure that the Fortran Format statement correctly specifies this alternate location (see Section 2.2.2.3 for related help).

A data file developed according to these above guidelines is readable by IA.

To show clearly how this file must be prepared the following example is provided.

Assume one administered an examination composed of 50 A-type items, five choices for each item were coded as A to E, and the candidates had entered their four-digit ID along with their answers on optically scored sheets.

A file typically produced when optically scoring the answer sheets is illustrated below.



Note that the first line has no data entered into the first four columns. These blank spaces are followed by 50 columns of letters (A to E). This first line is the answer key, and since a list of answers does not need an ID, the first four spaces (columns) are blank. The keyed answer to the first item in this example test is choice E – i.e., see the fifth character position (column 5) in the first line; the keyed answer for the second item is B (column 6); and so on until the keyed answer for the last item is D (last entry being column 54).

The remaining lines in this illustration begin with four numeric digits, followed by 50 letters varying from A to E. The first candidate has an ID of 1001 and his/her answer to first question is E, then B and so forth. The next candidate has an ID of 1002 and his/her answer to the first two questions is E and C. The last candidate has an ID 1045 and provided answers of E and C for the first two questions.

The user has options to vary conditions from those illustrated in the above example:

1. test may use letters or numbers for designating different choices in each item provided the total number of choices does not exceed nine; if 10 or more choices are used, the code must be alphabetic, i.e., A, B, C, . . . Z;

2. the number of choices across items may vary from item to item and range from 2 to 26 options (i.e., A to Z);

3. atypical alphabetic character sequences can be used (e.g., R, A, I and H for 'recommended', 'appropriate', 'inconsequential' and 'harmful' procedures);

4. candidates' IDs may be of any length and expressed in alphanumeric format;

5. one may concomitantly analyse a mix of item types (e.g., A-, and R- type items);

6. one may analyse each item type or any subset of items separately by restricting the number of items analyzed in each analysis; to do this, modify the first line of keyed answers by leaving blank the columns that correspond to items which are to be ignored in a particular analysis.

How to make use of these options is explained in more detail in Section 2.2.1.

2.1.3 Web-based Assessments (IDEAL Intranet)

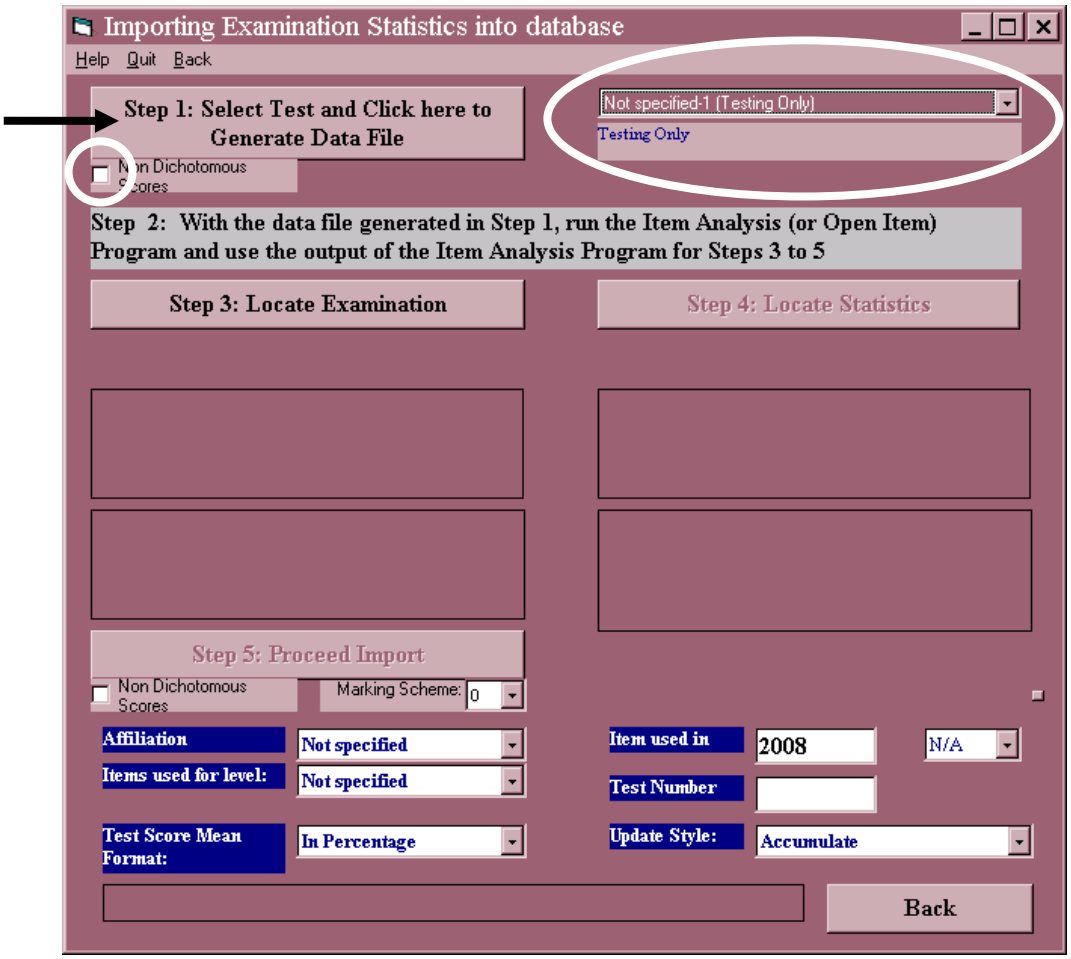
The data file containing the keyed answers and the candidates' responses for a web-based assessment is generated by the IDEAL Intranet program. How to generate and access this file is now explained.

In the Management window of IDEAL Intranet (illustrated at the beginning of Section 8.4.2 in the IDEAL Intranet Users' Guide), click the button "Import Statistics". The window illustrated below will open.

In IDEAL, any assessment administered on the Web will be A-type, R-type or Pick n of N, and will have a filename. This name must be assigned by the instructor or the item bank administrator at the time the test is created. Following the administration of the test, the filename can be located by clicking the drop down arrow located immediately right of the text box in upper right of the window (see circled section in the following illustration).

Find and highlight the appropriate filename. If the items are all Pick n of N, check (√) the small box "Non Dichotomous Scores" (small circle in the illustration), otherwise leave this check box blank (the default). Click the button that is labeled "Step 1: Select Test and Click here to Generate Data File".

A data file (with the extension *.dat) is then generated by IDEAL Intranet; the file will contain the key and the candidates' responses to each item that was used in the Web-based assessment. This file will be placed, by default, on the server, the location for which is C:\Program Files\IidealSetup\Item_Analysis*.dat (where * represents the filename assigned by the teacher for the Web-based assessment). This file is read by the item analysis program in order to compute the items' psychometric properties.

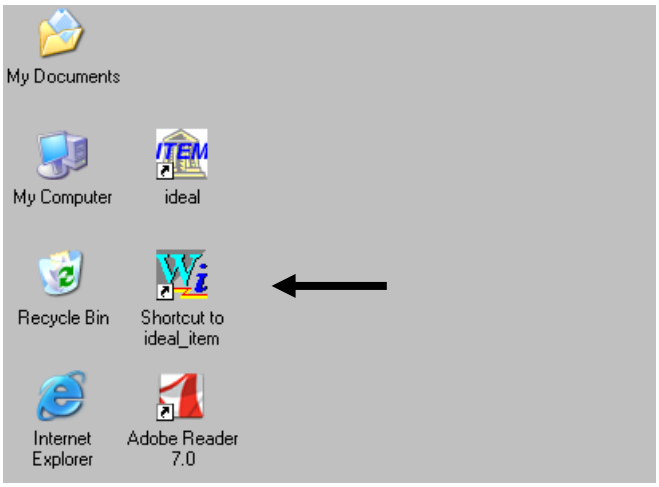


2.2 Analyse the Candidates' Answers

2.2.1 Steps to Follow

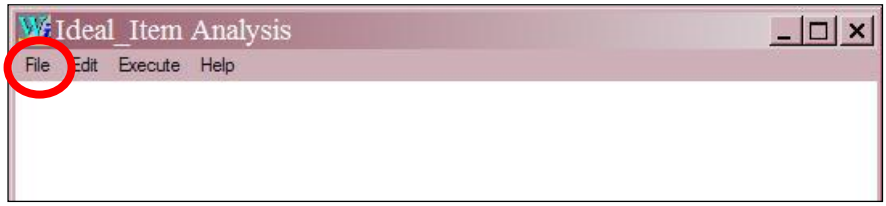
To analyse the data file generated by the optical scanner or IDEAL Intranet (see Section 1.2), follow these instructional steps in the sequence indicated.

1. Open the item analysis program: double click the application's shortcut placed on the windows desktop (illustrated below), or navigate to the appropriate directory and double click the application's icon.

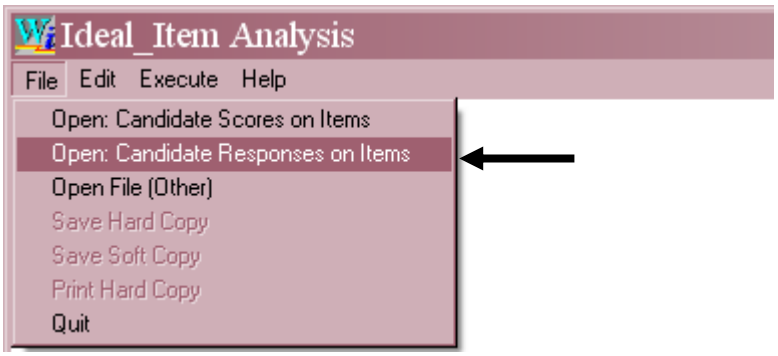


The item analysis window opens (illustrated below). An inserted message box will also appear (not illustrated), providing an overview for using IA; this message will disappear as one begins to use the program (i.e., executes step 2 as described below).

2. Click File (see upper left corner of the item analysis main window).

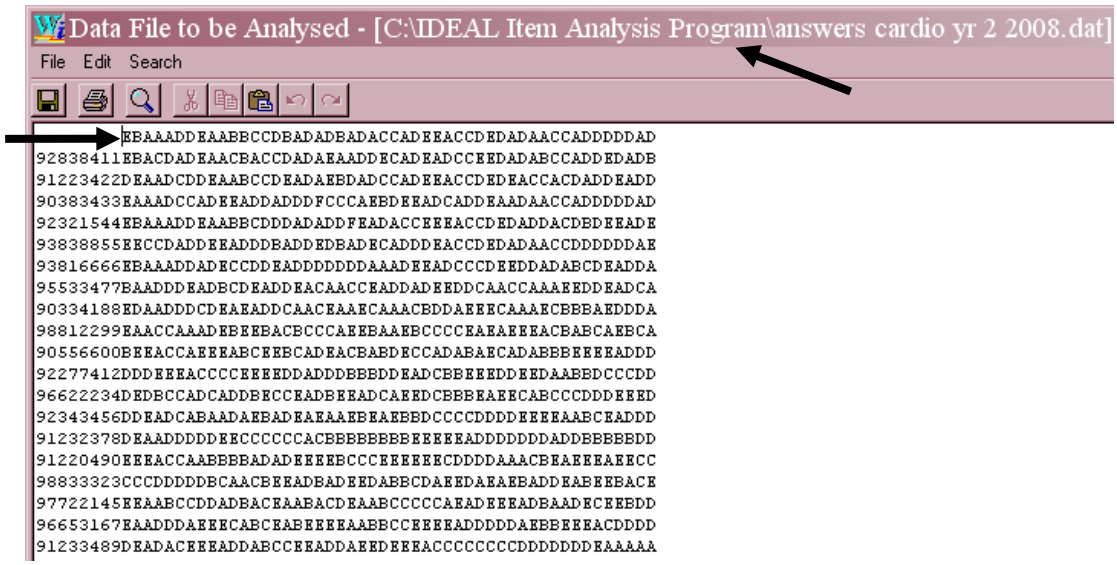


3. Click "Open: Candidate Responses on Items" (in the drop down menu).



The default directory for the data file will be displayed; if you previously saved the file generated in Section 2.1 in another directory, navigate to the appropriate directory.

4. Double click the desired filename and the file opens as illustrated below.



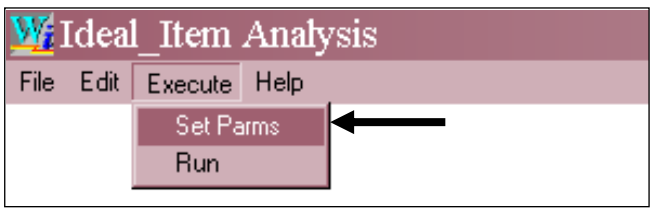
5. Confirm that the opened file contains the data that you wish to analyse; if you need to modify the data file, click "Edit" and enter any required changes.

6. In the first row, place your cursor immediately to the left (i.e., in front of) the letter (or number) that represents the answer to the first item to be analyzed:
   a. in the above illustration, the cursor is placed in front of the first item in the test which has E as the correct answer; the location is column 9 because a eight digit ID is being used;
   b. if one wanted to skip some initial items the cursor may be placed further to the right, the location depending on how many items are initially skipped.

7. Click "File" then "Exit"; however, if you edited the file since opening it, you must click "Save" before clicking "Exit".

You will be returned to the main item analysis window.

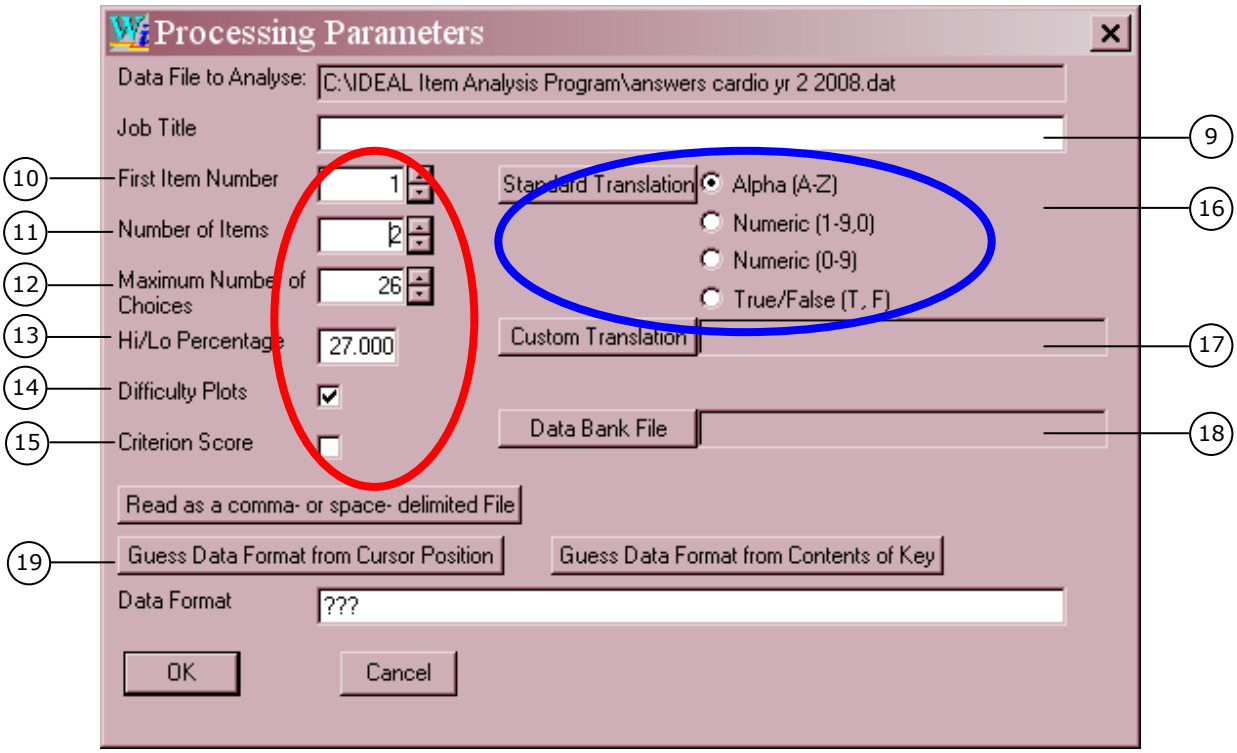
8. Click "Execute" and then "Set Parms" (set parameters) in the drop down menu.[2]



[2] If you omit clicking Set Parms (set parameters), the program will prompt you to complete the Set Parms.

The window illustrated below will be displayed. In this window the user enters his/her specific requirements for the statistical analyses and output.

This Processing Parameters window is shown if the data file has the extension *.dat. If the file was created with Excel and saved as a common delimited file (i.e., *.csv), the window will be slightly different. Refer to the Addendum at end of this chapter.



9. In the open text box to the right of "Job Title", enter any descriptive text that you want to appear at the beginning of the printed output.

10. In the box to the right of "First Item Number", select the position of the first item to be included in the analysis. This number is needed to ensure the labeling used in the output is appropriate.
    a. The default number '1' is appropriate if the entire test is being analyzed.
    b. However, assume the test has 35 items and that one intends to skip the first 20 items; specifying "21" ensures the printed output for each analysed item is correspondingly labeled 21 to 35 (not 1 to 15).

11. In the open box to the right of "Number of Items", specify the *total* number of items in the test you intend to analyse. Within this total number you have the option to omit particular items at anytime; to omit some items, leave the relevant column(s) blank in row one which correspond to the keyed answer(s) for those items that should be ignored for a particular analysis.

12. Specify the *maximum* number of choices in the items being analyzed; e.g., if the items' choices vary from 5 (A to E) to 10 (A to J), then the number is 10.

13. Specify the proportion of subjects that should be placed in the top and bottom performance groups. (The default value is 27%, matching the proportion in each group that is needed to calculate the classical discrimination index.)

14. Check the appropriate box if you wish to have item difficulty by performance groups scatter plots printed.

15. Check the box beside "Criterion Score" if a reference standard instead of the total test score is to be used for the measure of a candidate's overall ability. (Default is unchecked; if checked, the data file created in Section 2.1 must include this criterion measure, usually entered after each candidate's responses.)

16. Indicate the code used for the items' options. The default is alphabetic (A, B, C, D, E, . . . Z). If numerics are used for coding options, check the appropriate radial button. However, if a unique set of characters was used for coding options (such R and W or R, A, I and H), click the button labeled "Custom Translation" and follow step 17 in lieu of the step describe in this point 16.

17. If coding for the options is unique and the Custom Translation button is checked, a text box to the right of the latter button becomes accessible; enter the unique coding in this ungreyed text box. The characters should be entered consecutively and without blanks or commas (e.g., RW or RAIH).

18. To request a softcopy of the item statistics (that IDEAL can use to update the psychometric properties of the relevant banked items), click the button labeled "Data Bank File". Specify a file name for this softcopy statistic output in the text box located to the immediate right that will become ungreyed.

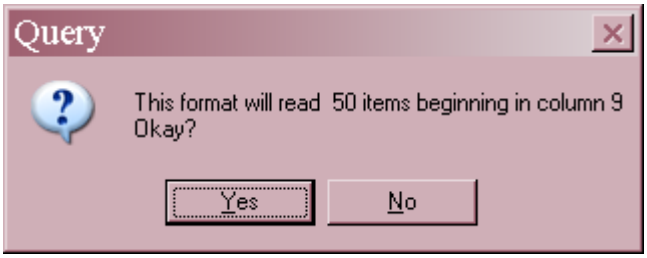
    Specify the *same* name as the filename that was assigned when the items in this test were selected from the IDEAL bank *and printed*. That is, when selected items are printed the user must specify the "Course/Module No." (see Section 5.1 in the IDEAL PC manual). IDEAL uses this course name to create a Word file (in rtf format) and a "student scoring statistics" (*.sss) file. The file containing the item analyses (generated by IA) has to be matched with the *.sss file (generated by IDEAL) at the time the item statistics are uploaded into the item bank. To help in visually making this match, use the same name. (As a different extension is used for this statistics output file (i.e., *.idb representing "*i*deal *d*ata *b*ank"), both files are distinguishable.

    One should confirm that the item analysis output is correct (and delete the softcopy output file (*.idb) if any error was made in executing the item analysis program).

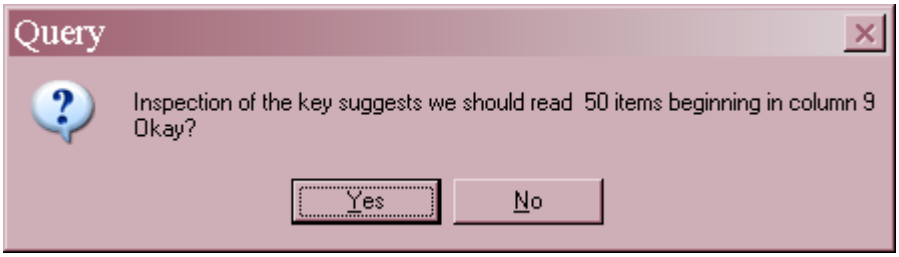
    If a filename for the Data Bank File is not specified, a softcopy will not be created. If one's purpose is to generate a file of item statistics so that IDEAL can update the relevant psychometric properties for all the banked items which were used in a given examination, then a name for Data Bank File must be specified.

19. Specify the Format of the data file by clicking one of the two "Guess Data Format" buttons. Clicking either button will instruct the program to 'guestimate' the correct Format for reading the data file.

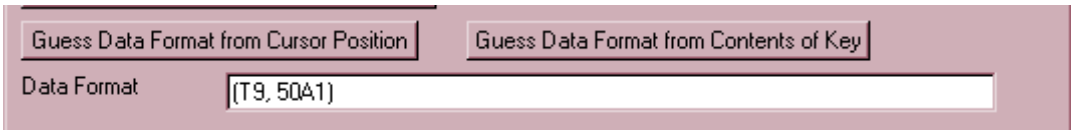
    Click the first button ('from Cursor Position") if you placed the cursor in the correct position in the data file (step 6 above) and specified the correct number of items to be analyzed (step 11 above). After clicking this button, the following illustrated query is displayed.



Alternatively, you may click the second button "Guess Data Format from Contents of Key" and IA will attempt to reproduce the correct Format based on locating the first non blank character in the key. After clicking this alternate button, the following query is displayed.



If you click "Yes" to either of these queries, a Fortran Format statement will appear in the open bottom box, replacing the "???" characters that are initially displayed.



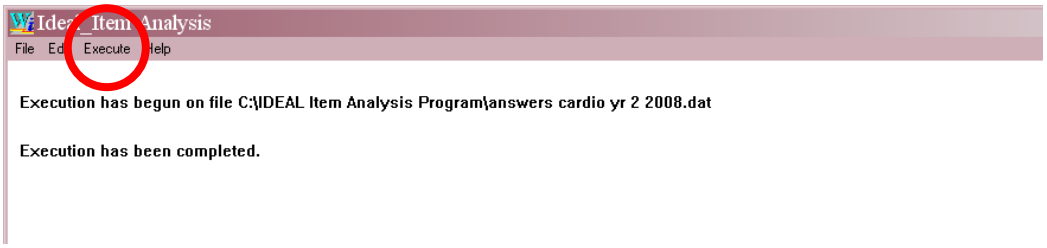
Confirm that the Format statement inserted by IA is correct, or modify the Format if required.

If a reference standard is used (indicated at step 15), ensure the Format also identifies the correct location for this criterion measure, using Tn, Fx.y (where n indicates the column number in which the measure begins, x is the maximum number of columns used for entering the criterion measure, and y is the number of decimal places used for indicating the value of the criterion measure.

If one is not familiar with Fortran Format statements, please refer to the explanatory information in Section 2.2.2 (or seek assistance from your IT resource personnel). However, for most, if not all, analyses that will be required, the Format statement generated by IA will be correct, provided the previous 18 steps were correctly followed.

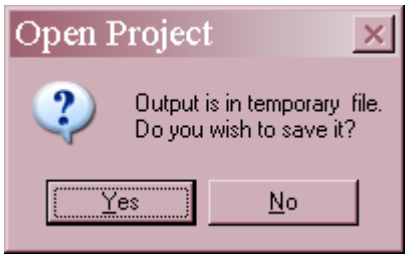
20. After confirming that all specifications are correctly entered, click the button labeled "OK". You will be returned to the main menu in the IA program.

Click "Execute" then "Run" in the drop down menu. The analyses will be executed quickly with the following messages displayed.



21. Save the results of the item analyses: click "File", then "Open File (Other)" in the drop-down menu.

You will be asked if you want to save the hardcopy output. Check "Yes".



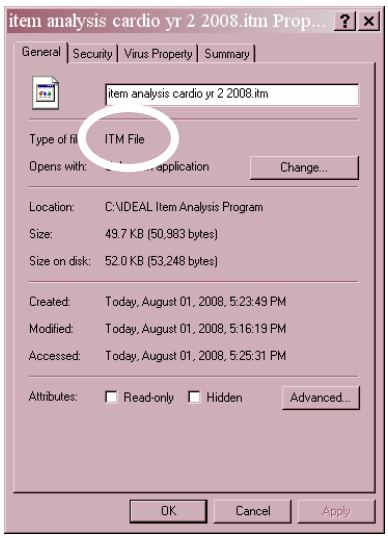
The filename you assigned for the data file (*.dat) will be displayed as a suggested filename for this hardcopy (which has an extension of *.itm). Highlight this file name and click "Save".

If you specified a filename in the "Data Bank File" open text box (in Section 2.2.1 point 18, when setting up the parameters), a window will also open asking if you want to save the soft copy output file. The suggested (default) filename again will be the filename previously assigned for the input data file (*.dat). This time the extension for the softcopy item analysis output file will be *.idb. Highlight this filename and click "Save".

If you didn't provide a filename in the Data Bank File (see Section 2.2.1 point 18), IA will not produce this softcopy file and thus, you'll be unable to save it.

To view (on your PC or server) the contents of the hardcopy (*.itm) & softcopy (*.idb) files, click "File", then "Open File (Other)" and in the subsequently displayed window select the filename you used when saving the files.

If your PC does not display the file's extensions (*.itm and *.idb), both the hard and soft copy files appear with the same names. To distinguish, hover your mouse over each file and a pop message with the extension will be presented. Alternatively, right click on the files, select "Properties" and the following illustrated window will display the extension.



2.2.2 Understanding Fortran Format Statements

2.2.2.1 Introduction

If you are familiar with Fortran formatting and can confirm that the Format statement provided by IA is correct, skip this section and go directly Section 2.2.3.

If one is not familiar with Fortran formatting, remembering the following basic rules will enable you to evaluate if the Format statements inserted by IA are correct.

a. The Format statement must begin with an open parenthesis and end with a closing parenthesis; if either or both are missing, insert them (i.e., place cursor at appropriate location and type " (" and/or ") ".

b. The letters T, A (and F if used) in the Format statement must be capitalized; use of lowercase letters will result in errors.

c. The letter T (representing Tab) is followed by a number corresponding to the column in which the first item to be analyzed is located. T9 is used for the illustrated data file in step 7 since the first item is located in column nine (i.e., following an eight digit ID used in the illustrative data set).

d. A comma must follow the combination of T and its associated number (e.g., "T9, ".)

e. The letter A (representing alphanumeric) is preceded by a number corresponding to the number of consecutive items that are to be included in the analysis. The number 1 following the letter A indicates only one column is needed to properly represent an item's coded answer; in IA this number is always 1. Thus, for example, 50A1 indicates that 50 consecutive items are located in 50 consecutive single digit columns.

f. The complete Format statement for the example illustrated at step 7 is (T9,50A1), indicating that one wants to analyze 50 items, the first of which begins in column nine and continues until column 59.

Format statements allow the user to vary which items are to be analyzed and which are to be ignored in a data file. Using the above example, if one wanted to analyze only a subset of items (e.g., items 21 to 35) one would begin in column 29 (i.e., skipping the eight digit ID and the first 20 items) and read the answers for the next 15 items only; thus, the Format would be (T29,15A1).

There are instances in which some extra care should be taken, in particular if sets of non-consecutive items are to be analysed or if a referenced standard is to be used. These cases are now explained.

2.2.2.2 Non-consecutive Item Subsets are to be Analysed

Assume the same data file as discussed above, but one wants to analyze only items 21 to 35 and items 46 to 50. The Format statement should be (T29,15A1,T55,5A1). In other words, the data begins in column 29 (as eight columns containing the IDs are skipped along with another 20 columns for items 1 to 20 that are of no interest at this time; 15 items, 21 to 35, are then read (15A1); the remaining five items (5A1) begin in column 55 so the latter is preceded by the tab T55.

One may have more than a series of two sets. The principle to follow is specify the column in which the item or items begin by inserting Tx where x is defined as the column of the first item in the series wanted. This is followed by the number (y) of consecutive items wanted at that point (yA1). One can continue with this series of T's and A's, always separating them by commas, until all desired items are identified.

2.2.2.3 Reference Standard (Criterion Score) is Used

Recall that one may use a criterion measure other than the total test score for calculating the adjusted and unadjusted biserial and point biserial coefficients. If this is intended, one must then provide the value of this criterion for each candidate and this criterion would normally follow the candidates' responses. To instruct the program where the measure is located in the data file, do the following.

After the last entry in the Format statement, but before the closing parenthesis, insert a comma, specify the column where the criterion measure begins and then enter Fx.y, where x is the total number of columns used to record the criterion measure, and y is the number of columns out of x that are used after the decimal place.

Assume, for example, that a criterion score was the grade point average of the candidates in their previous year; assume also that these grade points were calculated to the second decimal place (e.g., 4.06, 3.72, 3.51, etc). In this example x is four (as three columns are used for representing the digits and one more column is needed for the inserted decimal point); y is two (two columns are needed to specify the two decimal places). Thus, the Format for specifying this illustrated criterion measure is F4.2.

If one had included this criterion immediately after the candidates' responses, the previous example's complete Format would be (T29,15A1,T55,5A1,T60,F4.2) or more simply, (T29,15A1,T55,5A1,F4.2).

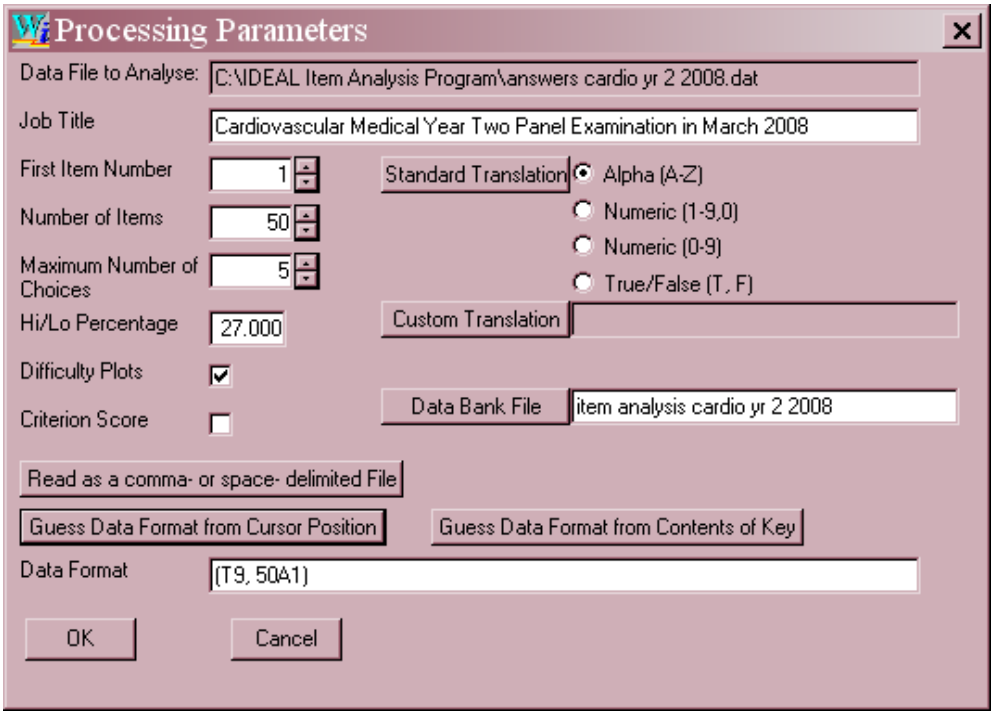
If one's criterion does not have decimals, y in Fx.y is zero. Unless the user typed a decimal point when entering such data, it is likely that the number of columns needed to represent the criterion is simply the number of digits preceding the decimal. For example, assume the criterion was a percentage score from another test and these ranged from 47 to 93. If so, the appropriate Format for this criterion would be F2.0.

Assume, however, that these percentages ranged from 1 to 100. In this case, the maximum number of columns needed to represent the criterion is three (or F3.0) as the three-digit number 100 is within the data set. One must remember that when entering the data for any values that are less than 100 (i.e., three columns in length), the values must be entered in a right justified fashion. That is, the value 1 is either "001" or "blank, blank, 1" (and similarly for all single digit values); the value for two digits, such as 10, 33, and 86 are entered as 010, 033 and 086 (or blanks can be used in lieu of the leading zero for each number). If one fails to right justify when entering the data, the Format statement will misread the correct criterion value.

2.2.3 Example of Set Parms (setting parameters)

The following window illustrates suitable specifications for running an analysis of the data file illustrated in Section 2.2.1, step 7. We assume that all fifty items are to be included in the data analysis:

- a suitable title was inserted (and will be printed in the hardcopy output);
- the default (1) was used as we need to begin with the first item (located by placing the cursor just before column 10);
- the number of items was set to 50;
- the number of choices was set to five;
- high and low performance groups were set as the top and bottom 27%
- difficulty plots were requested;
- no alternate criterion performance was used (default of total test score was used);
- the Format is one created by pressing the "Guess Data Format from Cursor Position";
- the data file had coded answers with letters so the Standard Translation was A-Z;
- as a soft copy output file for the item analysis was wanted, a file name was specified (i.e., "item analysis cardio yr 2 2008.idb").



After checking to see that these specifications are correct, one would click the OK button (lower left in the above window). The user would be returned to the main page in IA, in which s/he would select "Execute" and then select "Run" in the drop-down menu in order to have IA compute the item analyses.

2.2.4 Print the Hardcopy Output File (*.itm)

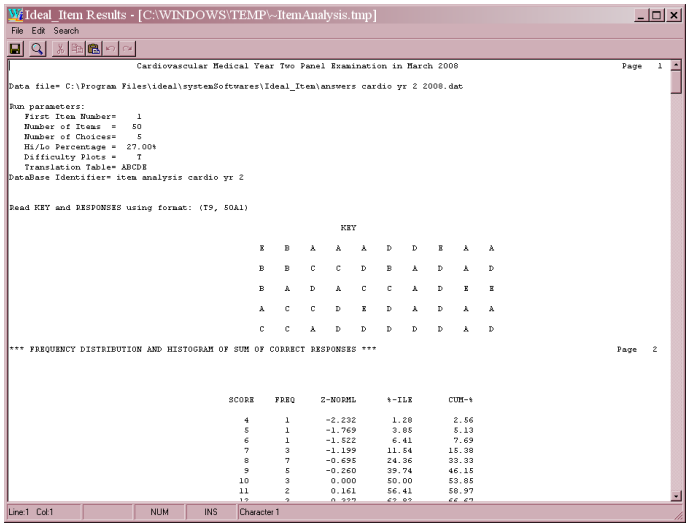
After running an item analysis, a printed copy of the hardcopy output can be obtained.

Click File in IA's main menu and then in the drop down menu, click "Print Hard Copy".

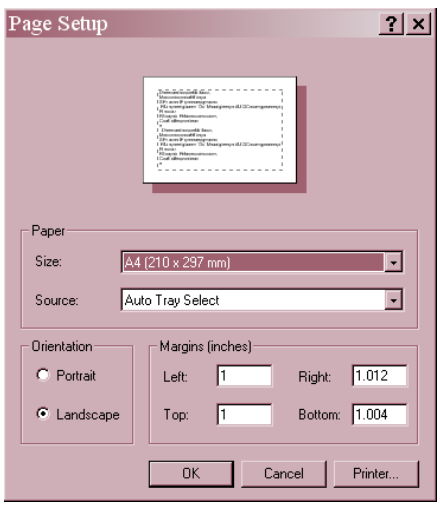
If the Print Hard Copy function is greyed (i.e., inaccessible), this probably means that while you may have previously saved the *.itm file, you also subsequently did other analyses, or you closed the application or some other procedure such that the application is unable to determine which Hard Copy to print.

Provided you have saved the file at some previous point in time, just click File then "Open" and select the appropriate *.itm file. Click Open. Subsequently, you will see that the Print Hard Copy is now no longer greyed and can be printed.

The initial printed material from the Hard Copy file will be similar to the data displayed in the following illustration.



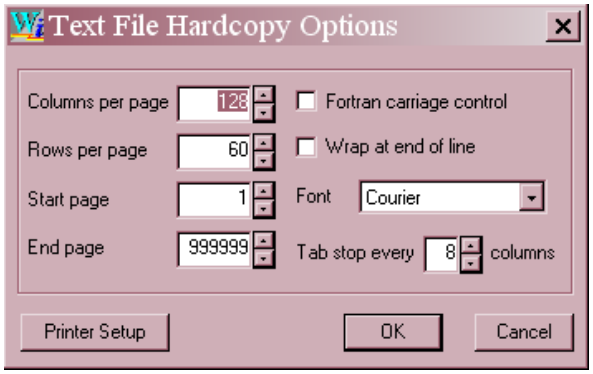
The Print Hard Copy option uses a special print routine and scales the output to fit on the page. By default, landscape orientation is chosen and normally should not be changed. The scaling procedure takes into account the size of paper selected and larger characters are used when legal-sized paper is chosen.



If the file to be printed has the extension "*.itm", it is assumed that the file has been generated by the item analysis program and this special print routine is invoked. The significant feature is that the output is scaled to fit the page.

Any other type of file may be printed using the Print Hard Copy in conjunction with the item analysis Edit function. For these general situations, the page setup is left to the user. For example, if the lines are too long to fit on the user-defined page, the lines will be wrapped onto 2nd and subsequent lines. Note the Text File Hardcopy Options window and its the first entry "Columns per page" (see illustration).

"Columns per page" represents the maximum number of characters that can appear on a single printed line. If the output to be printed comes from the current execution of the item analysis program, then the number appearing in the adjacent open box (128 is illustrated) will be the number that is most appropriate for the output. If one overrides this number with a smaller value, the output lines will be truncated.



If the output to be printed was generated on a previous occasion, one can retrieve this number by pointing the cursor to the page number, e.g., the 1 of Page 1, while in the editor function (in the item analysis main menu select "Edit") and before executing the Print Hard Copy command.

Lastly, as IDEAL supports items having up to 26 alternatives, the printed item analysis output provides the breakdown of high to low performing groups for only the alternatives that at least one candidate selected the choices. As all non selected choices would be a series of only zeros, these are not printed, making the output more readable.

For example, assume the item analysis printout lists only the responses for the high to low performing groups for choices A, D, E, G, I. One can conclude that no candidate had selected choices B, C, F, H and choices J and beyond.

2.2.5 Purpose of Soft Copy Output File (*.idb)

The Soft Copy file is readable by IDEAL and is created for purposes of easily updating the psychometric properties of banked items in its system. As noted, the Soft Copy file is created only if the "Data Bank File" option was selected and a filename was provided when setting up the item analysis parameters in the "Set Parms" window.

The first record (line) in the Soft Copy file contains the filename that was provided by the user.

The remaining records (lines) appear in sets, one set for each item in the test. Each set consists of seven lines and in each of these seven records, the following data are specified:

1. Item number in the test, Keyed answer, Difficulty level (expressed as a percentage) and Point-biserial correlation (expressed in decimal form);

2. For Total group: Number of candidates, Percentage in the total group choosing each alternative in the item;

3. For High group (P_{73} - P_{100}): Number of candidates, Percentage in this group choosing each alternative in the item;

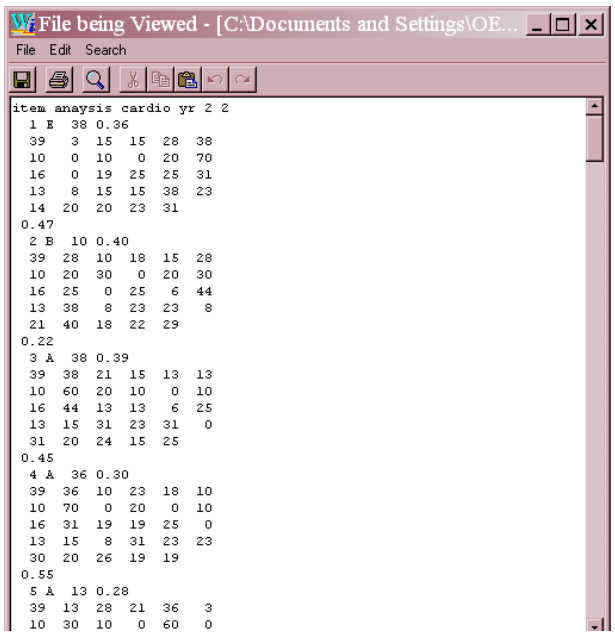
4. For Middle group (P_{27} – P_{73}): Number of candidates, Percentage in this group choosing each alternative in the item;

5. For Low group (P_1 – P_{27}): Number of candidates, Percentage in this group choosing each alternative in the item;

6. The Mean Performance on the total test (expressed as a %) for the group of examinees that choose each alternative in the item;

7. Discrimination index for the keyed response.

The Soft Copy file produced by the item analysis program, when opened and viewed on your monitor, appears as:

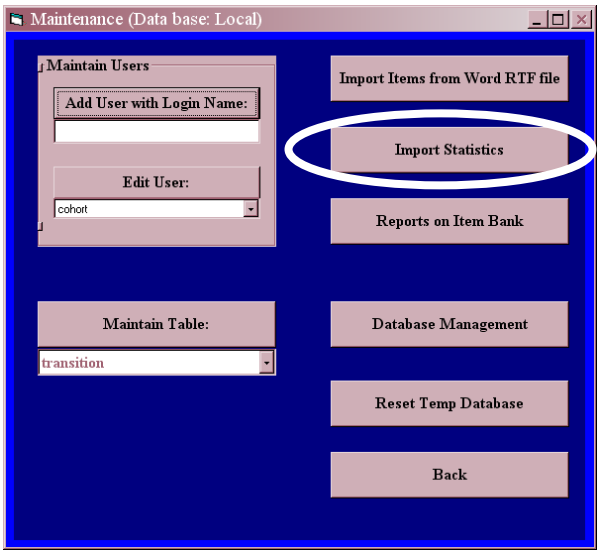


2.3 Importing Item Statistics into IDEAL

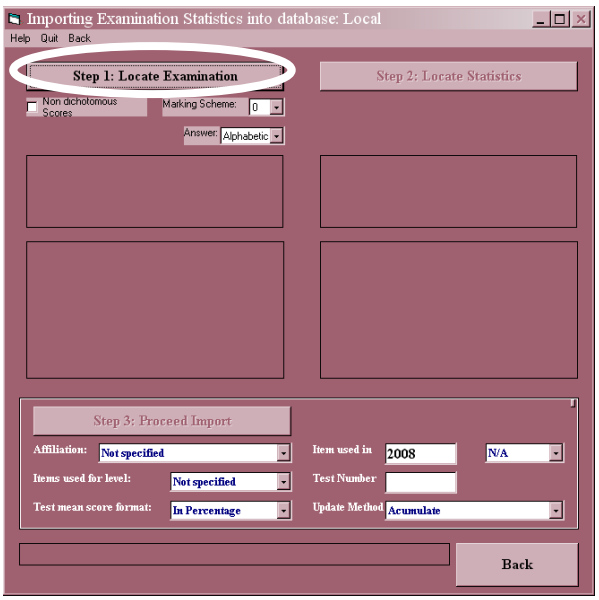
The "Import Statistics" function in IDEAL facilitates the process of importing a large number of item statistics into the item bank. The procedures for importing the item analyses vary somewhat between IDEAL PC and IDEAL Intranet. These variations are described below.

2.3.1 Importing Statistics into IDEAL PC

1. In the main menu of the IDEAL PC program, click the "Maintenance" button. The "Maintenance" window opens (illustrated below). Click the circled "Import Statistics" button.

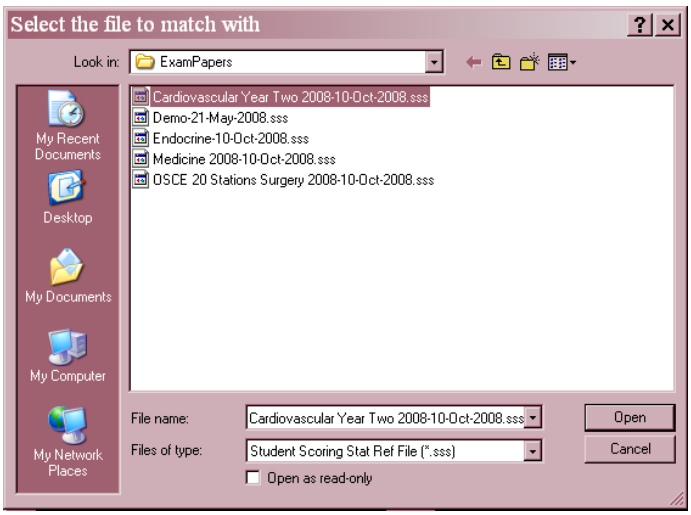


2. The "Importing Examination Statistics into database" window opens (illustrated below). One button labeled "Step 1: Locate Examination" will be accessible (circled). Beneath this button there are three small open check boxes. If you are dealing with items that have A-Z coded for options, specify "Answer" as alphabetic; if numerics were used for the coding, select numeric. If dealing with selected response items, leave the "Non dichotomous Scores" box blank. If you are referencing a test based on Pick n of N items open the box for "Non dichotomous Scores" box that is checked (√). Normally use the default value for the scoring scheme box. Subsequently, click the button labeled "Step 1: Locate Examination".

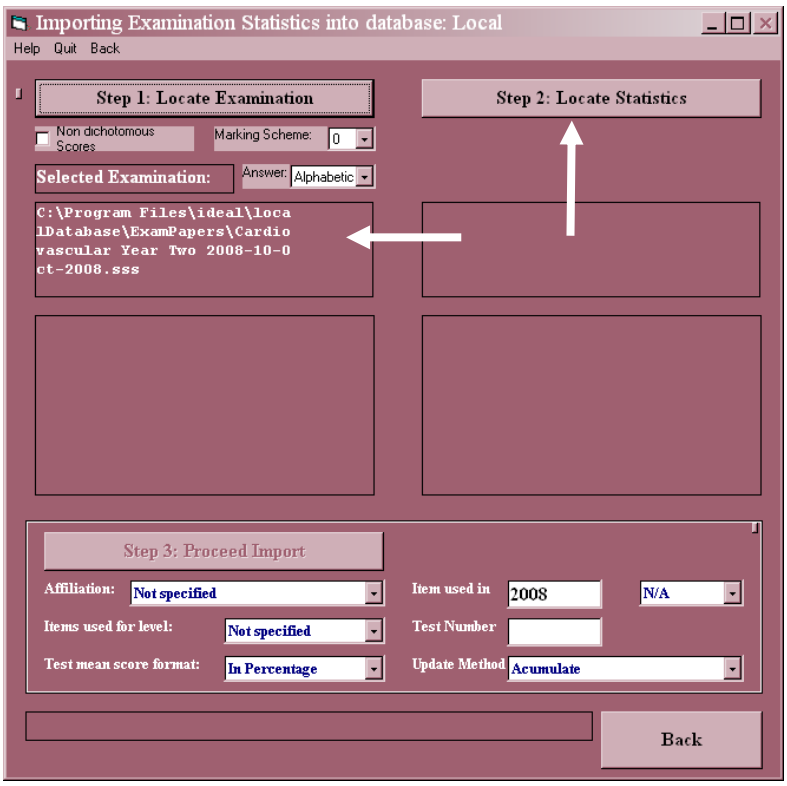


3. The following window opens, by default in the ExamPapers folder [i.e., C:\Program Files\ideal\cohortDatabase *(or another database name)*\ExamPapers].

   Locate and highlight the file name that was assigned when the items' were originally selected from the item bank *and printed*. The file name will be the course name or module number followed by a date when the test was printed. The relevant file will have the extension *.sss (standing for "student scoring statistics"). This file was automatically generated by IDEAL PC and contains information for referencing item IDs stored in the item bank with the items used in the examination and item analysed. Highlight this filename and then click the button "Open".

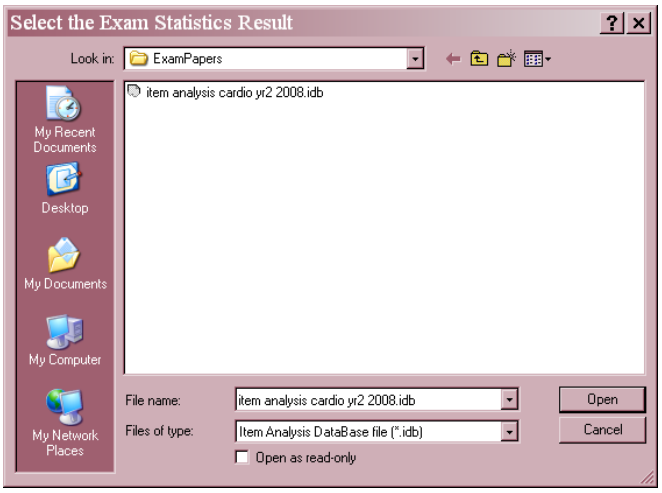


The file that you selected is then displayed (see left hand arrow in the following illustration). The button needed for the next step will become accessible (i.e., ungreyed): "Step 2: Locate Statistics" (see right hand arrow).

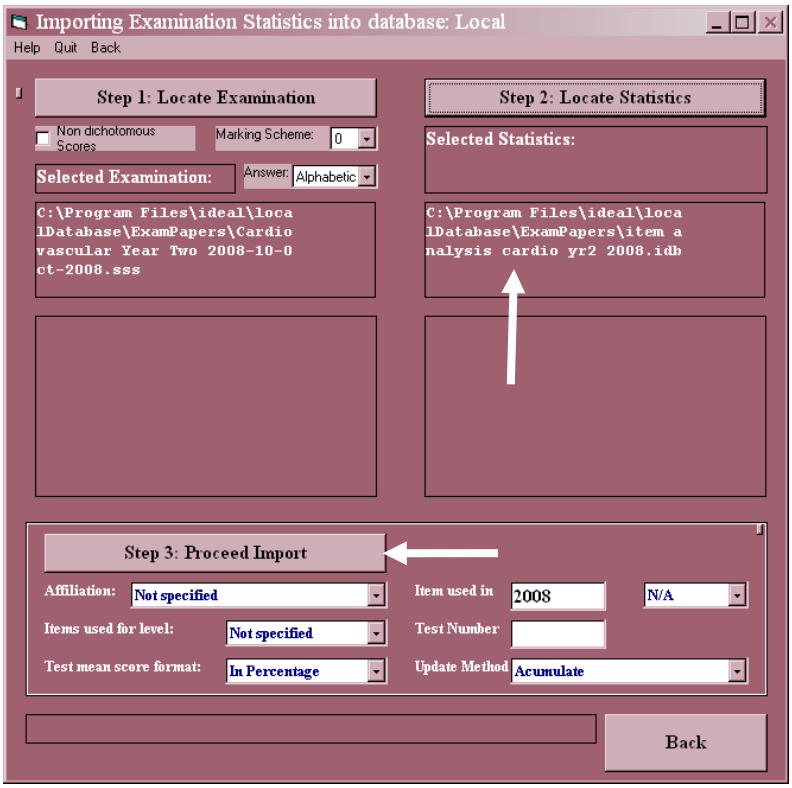


4. Click the now accessible "Step 2: Locate Statistics" button.

Locate and select the Soft Copy output file (*.idb) produced by IA for which presumably you assigned the same file name as noted in Step 1. The window for "Select the Exam Statistics Result" will open. If the relevant Soft Copy output (*.idb) file was saved in the folder "ExamPapers" [C:\Program Files\ideal\CohortDatabase *(or another database name)*\ExamPapers] you will find the file in the displayed window. If you saved this *.idb file elsewhere, navigate to the appropriate location (e.g., if the file was saved in the same folder as the IA application navigate to C:\Program Files\ideal\systemSoftwares\Ideal_Item).



The name of the selected Soft Copy output statistics (*.idb) file you just selected will then be displayed as illustrated in the following window (see right hand arrow). The third (bottom left) button becomes accessible.



5. Before clicking the Step 3 button, note the lower section in the above window. Always enter information related to the administered examination – i.e., your university, the year of the medical program in which the test was used, chronological year in which the test was administered. Use the default "Percentage" for indicating how the test mean is reported. Normally use the default "Accumulate" in specifying the updating method; the alternate choice of "Update" is used only if you previously uploaded the same statistics but made some mistake --- i.e., use of "Update" will overwrite the previous errors.

6. Click the button labeled "Step 3: Proceed Import". IDEAL will check if these two files reference the same items and in the same sequence. If the files match, all the item statistics will be imported into the IDEAL bank and the following message appears:

   "Succeeded in importing the statistics for <*total number*> records.
   Press OK to return to menu."

   If the selected files do not match in terms of their respective items, you will receive a message that the item statistics could not be uploaded into the IDEAL bank (see following illustration).

   If unsuccessful, determine the reason why. The items and sequence when analysed by IA must be exactly the same as selected from the item bank. If necessary reselect and print the items in IDEAL PC to match the items and sequence used in IA. Then repeat the importing process in this Section 2.3.1.

   The common cause for failing to successfully upload the item analysis statistics for selected response items is that the items selected and printed were different than those that were item analysed.

   To correct the problem, there are two possible solutions to consider.

   The first possible solution is to locate and use correctly matched files if you feel confident that they have been generated. You may have misidentified the relevant files for the *.sss and *.idb files. For example, if one did not use the same name assigned for the printed items, for the Data Bank File and for the corresponding Soft Copy output statistic file, it is possible one has inadvertently specified the wrong file in Step One or Step Two of the uploading procedure.

   The second possible solution is to confirm that the Softcopy item statistics file (*.idb) is based on only items that were selected and printed by IDEAL. If the administered test also included items that were not stored in the item bank, it is possible that the item analysis was for all the items in the test (rather than only the items from the bank). If this has occurred, appropriately modify the first row in the input data file that contains the key and candidates' responses. That is, in row 1, (the key) enter blanks into all the columns used for designating the items that do reside in the bank. Rerun IA and then rerun the uploading procedure described in this Section 2.3.1, using the latter generated *.idb file at Step Two.

   When the uploading of the statistics is successful, the relevant items stored in the bank will have their psychometric properties updated. The uploaded statistics will be displayed as the last used statistics; any former item statistics become the second

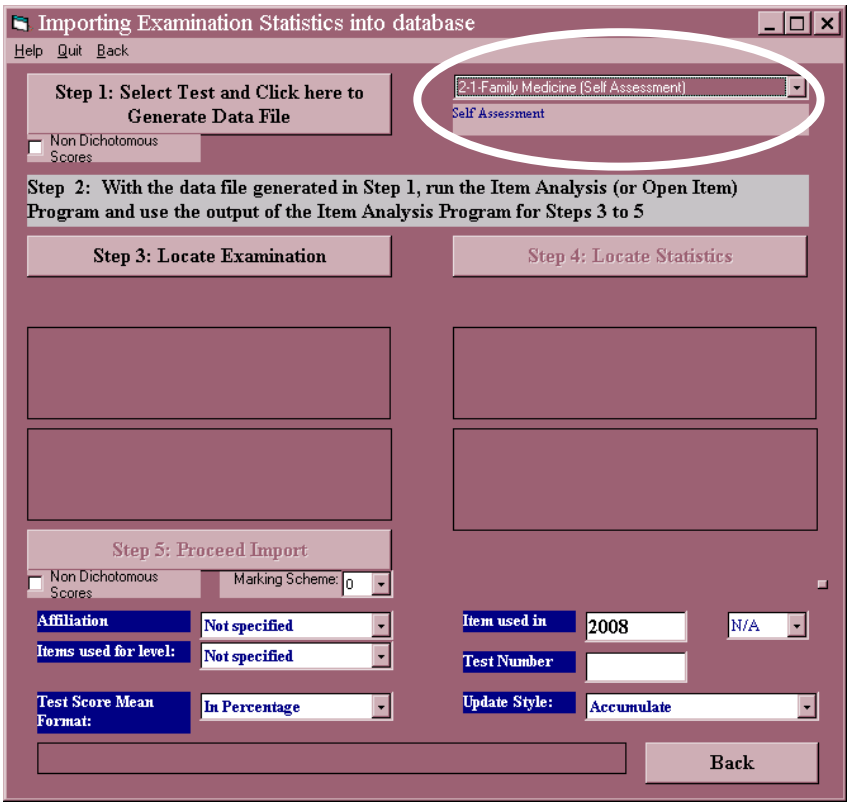
and if applicable the third last used statistics (third and later analyses are not displayed but are archived by IDEAL).

7. Finally the user may notice there is a very small (quasi-hidden) button on the right hand side of the window previously illustrated. This button when checked will provide features for the user to permanently delete an item's statistics (including the item's statistics that have been archived). This button is used if one wants to eliminate any previous errors one might have made in entering item statistics. The feature should be used with care; if the statistics are removed with this last feature, the result is not reversible.

2.3.2 Importing Statistics into IDEAL Intranet

To import the item analyses for a group of items that were used in a Web-based assessment, adhere to the following guidelines and in the sequence indicated.

Click the "Import Statistics" button in the Management window (illustrated in the beginning of Section 8.4.2 in the IDEAL Intranet Users' Guide). The following illustrated window opens.

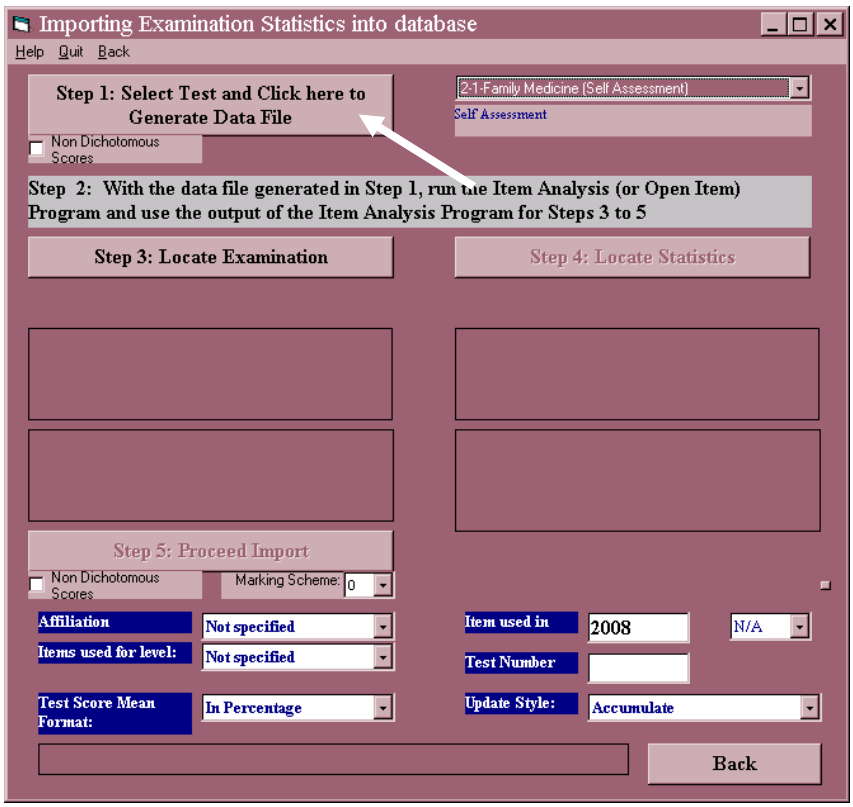


Note the circled drop-down arrow (above illustration). Click the arrow and scroll to and highlight the relevant test that was created, saved and administered on the Web using IDEAL Intranet. All tests administered with IDEAL Intranet in your institution will be listed in the upper right drop-down text box.

There are five remaining steps to take in order to import and update the psychometric properties of the banked items that were used in the Web-based assessment.

Step One: Note the open box "Non Dichotomous Scores" under the "Step 1" button. If the items are A- or R-type, this box is left blank (default). If the items are Pick n of N, check (√) those that are in the box. Please note that tests for A- or R- type items must be addressed separately from Pick n of N. This latter restriction is due to the item binary and non-binary scoring differences (i.e., must be item analysed separately).

Click the button labeled "Step 1: Select Test and Click here to Generate Data file".



Two files will be generated by IDEAL Intranet. The first file contains the key (i.e., correct answers) and the corresponding candidates' responses (i.e., their answers) for this particular Web-based assessment. This file will have the extension ".dat" and is placed into the following directory on your server:

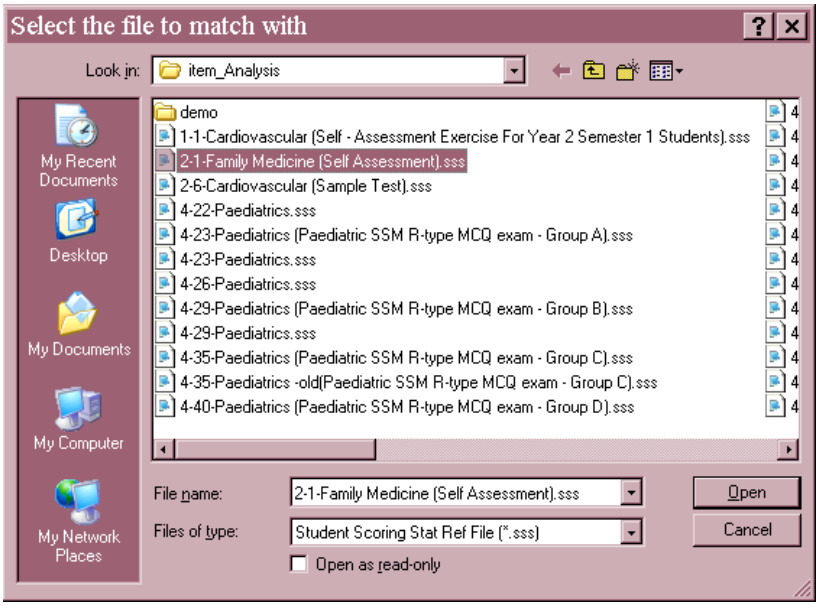
C:\Program Files\IidealSetup\Item_Analysis.

You need to access this file when executing step two (i.e., running the item analysis program). Thus make note of this default path and the filename.

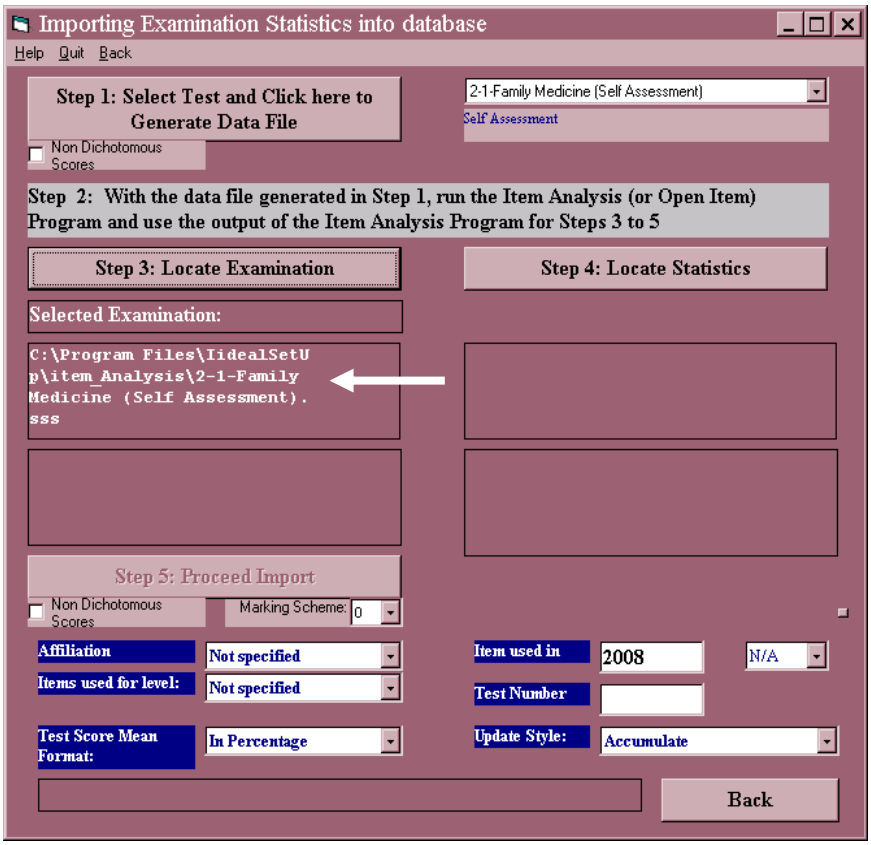
The second generated file has the extension ".sss" (student scoring statistics). This file contains information for matching items in the item bank with those used in the Web-based assessment. This file is also placed into the previously noted directory on your server. You will need to access this file when executing step three. The relevant directory will be presented by default as you execute step three.

Step Two: run IA, using the data file with the extension ".dat" that was generated in Step 1 (above). Requirements for running the item analysis program are the same as already provided in the previous sections of this Chapter Two. A Soft Copy output statistics file will be produced by the item analysis program. This file will contain all the relevant psychometrics for updating the statistical records that are stored with each relevant banked item. This Soft Copy output file will have the extension ".idb". The suggested filename is the one you used for the Data Bank File in the "Setting Parameters" window when running the IA program.

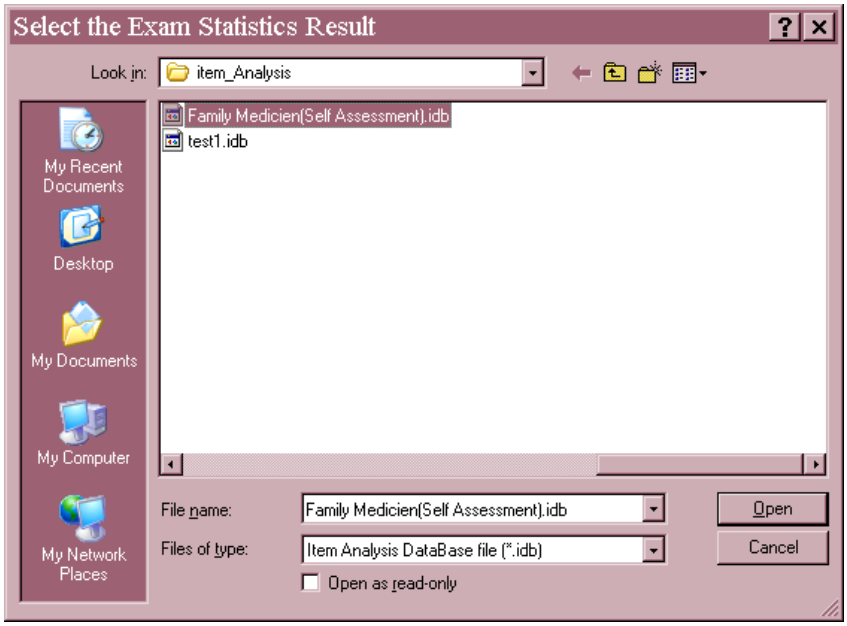
Step Three: click the button "Step 3:Locate the Examination". The following window opens in which you need to select the *.sss file created in Step One.



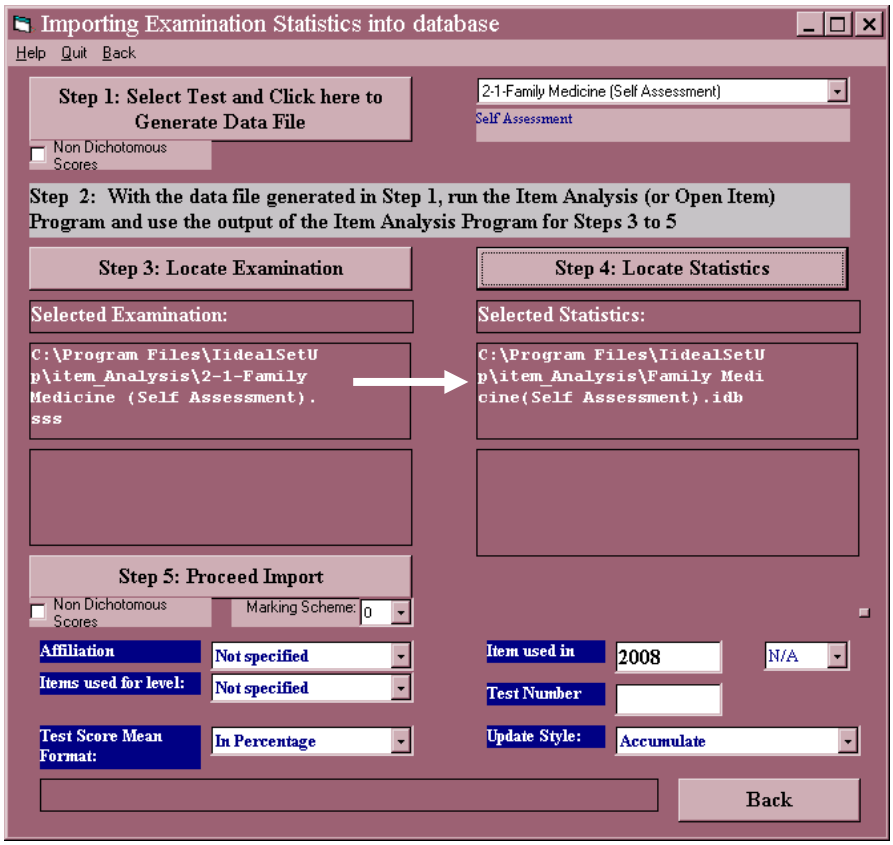
Highlight the file name and click the button "Open". The selected file is then displayed as illustrated below (see arrow).



Step Four: click the button "Step 4: Locate Statistics" and the following window opens in which you must select the *.idb file created in Step Two.



Highlight the file name and click the button "Open". The selected file is then displayed as illustrated below (see arrow).



Step Five: before clicking the button "Step 5: Proceed Import", enter information concerning the institution (i.e., 'Affiliation'), year in the medical program, calendar year in which this test was administered, etc.

Click the button, "Step 5: Proceed Import".

IDEAL will now import and store all the updated item statistical data with the appropriate banked items.

Note that these latest statistical data will be stored and displayed as the statistics for the last use of these items in a Web-based test. Any previously stored item psychometrics for these items will become the second last use. If three or more sets of statistics are available, IDEAL will archive these latter data (which are retrievable when needed).

2.4 Psychometric Properties Defined

The Item Analysis program provided with IDEAL produces test and item statistics based on classical test theory. Output includes a frequency distribution and histogram of total test scores, item statistics such as difficulty level and point-biserial correlation, discrimination index, item responses by high, middle and low-scoring groups, item-difficulty plots of each item against the quintile-groupings of total test scores and a scatterplot of difficulty level by biserial correlation.

The following item (*vs* test) statistics produced by this item analyses program are stored with the respective items in the bank. Statistics kept in the institution's bank will be those derived from the item's use in one's own institution unless the item has been only used elsewhere (i.e., in another institution within the Consortium).

The head office of IDEAL will archive all institutions' respective item statistics so that any member may use the latter database for benchmarking purposes.

1. Difficulty Level:

   Percentage of students who answered the item correctly.

2. Discrimination Index:

   This index provides information about the item's ability to differentiate between high and low scoring examinees.

   Discrimination Index is calculated as: $P_{high} - P_{low}$

   where P_{High} is the proportion of examinees in the upper 27% of the total test score distribution that answered the item correctly and P_{Low} is the same proportion but in the lower 27% group.

3. Point Biserial Coefficient:

   Correlation between score on the item and score on the total examination is defined as the point biserial. A positive value indicates that examinees who answered the item correctly scored relatively higher on the whole examination. A negative value indicates that examinees who answered the item correctly scored relatively lower on the whole examination (and vice versa).

4. Angoff Probability:

   A group of subject matter experts collectively estimates the likelihood that a minimally acceptable candidate would be able to answer the item correctly.

5. Nedelsky classification method:

   A group of subject matter experts estimates the expected score on the item by a minimally acceptable candidate (derived by eliminating the alternatives that this type of candidate would consider obviously incorrect).

2.5 Interpreting the Item Analysis Report (Hardcopy)

Overview

The item analysis program computes an array of test and item statistics (based on classical test theory) for selected response item formats. Examiners and assessment committees regard these data as useful aids for assigning grades, diagnosing the quality of an overall examination and/or its individual items, and as a means for improving subsequent assessments.

The IDEAL Consortium has determined that some of the item statistics are appropriate psychometric properties to store with X-, A-, and R-type banked items in the International Database for Enhanced Assessments and Learning (IDEAL). If banked Short Answer and OSCE/OSPE assessments include only dichotomously scored tasks, the psychometrics produced by IDEAL's item analysis program are appropriate.

However, for multiple response item formats and modified essays, this item analysis program is not suitable for estimating relevant assessment performance characteristics (psychometrics).

The following explanatory material is intended as an aid for those users who wish to print and use the analysis report. The format and sequence of presentation follows the sequence of a typical report produced by the item analysis program.

2.5.1 Scoring Key

The printed report begins by displaying the answer key that was used to score the student responses. If an ***** is printed, no keyed answer was supplied for the corresponding item (illustrated below).

```
                    KEY
*    D    A    D    D    B    B    B    A    C
A    A    D    A    B    D    B    D    B    B
A    C    B    A    B    D    A    D    D    B
C    D    B    B    D    D    B    A    C    B
```

Purpose: Whenever the reported results for the test or an item appear confusing or inconsistent, check the above to determine if the scoring key was indeed correct.

2.5.2 Frequency Table of Students' Total Scores

Frequencies (number of examinees) that obtained each test score are provided, along with the corresponding Z-score, the percentile associated with the score and the cumulative percentage of examinees obtaining a score at or below the listed value.

The two cutting points that were used to divide the class into high, middle and low groups are indicated after the frequency table. Unless otherwise specified by the user, 27% of the class is placed into each of the high and low groups with the remaining 46% in the middle. The 27% figure is an approximation and depends on the size of the class and the presence of tied scores at the cutting points.

Purpose: Among the several uses that a table of this kind provides, one is the increased convenience for determining cutoff points for assigned student grades, while another is the easy identification of the cumulative proportion that has met one's criterion reference standard.

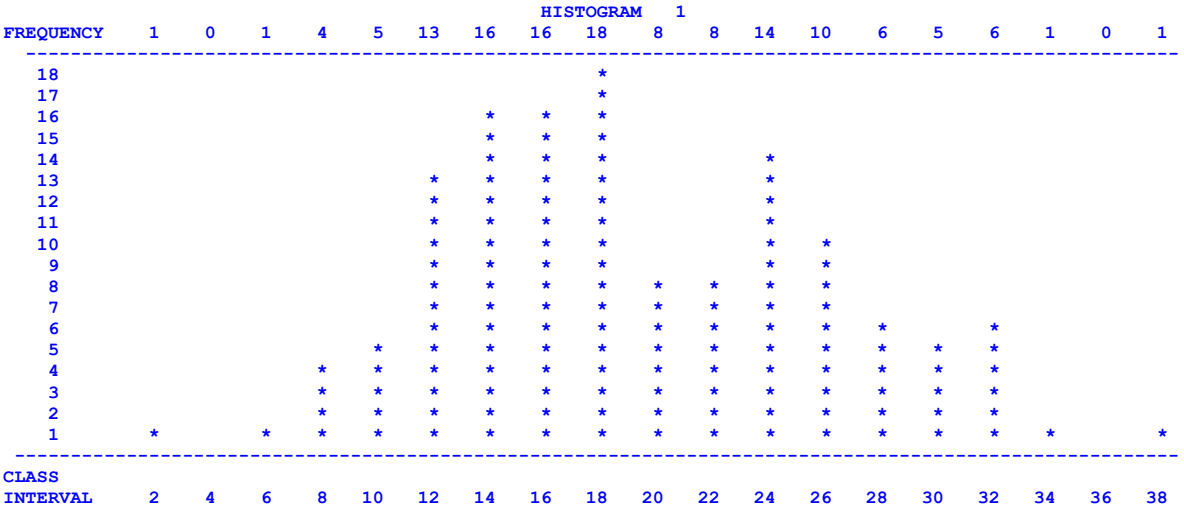
*** FREQUENCY DISTRIBUTION AND HISTOGRAM OF SUM OF CORRECT RESPONSES ***

SCORE	FREQ	Z-NORML	%-ILE	CUM-%
2	1	-2.673	.38	.75
3	0	-2.432	.75	.75
4	0	-2.432	.75	.75
5	0	-2.432	.75	.75
6	1	-2.281	1.13	1.50
7	2	-2.004	2.26	3.01
8	2	-1.780	3.76	4.51
9	1	-1.656	4.89	5.26
10	4	-1.494	6.77	8.27
11	5	-1.273	10.15	12.03
12	8	-1.035	15.04	18.05
13	6	-.831	20.30	22.56
14	10	-.633	26.32	30.08
15	10	-.417	33.83	37.59
16	6	-.257	39.85	42.11
17	7	-.132	44.74	47.37
18	11	.038	51.50	55.64
19	7	.208	58.27	60.90
20	1	.286	61.28	61.65
21	4	.336	63.16	64.66
22	4	.417	66.17	67.67
23	6	.522	69.92	72.18
24	8	.680	75.19	78.20
25	4	.831	79.70	81.20
26	6	.972	83.46	85.71
27	3	1.119	86.84	87.97
28	3	1.232	89.10	90.23
29	4	1.387	91.73	93.23
30	1	1.523	93.61	93.98
31	1	1.586	94.36	94.74
32	5	1.828	96.62	98.50
33	0	2.170	98.50	98.50
34	1	2.281	98.87	99.25
35	0	2.432	99.25	99.25
36	0	2.432	99.25	99.25
37	0	2.432	99.25	99.25
38	1	2.673	99.62	100.00

The cutting points used to divide the class into HIGH, MIDDLE and LOW groups
for the item statistics presented below are: 14.5 and 23.5

2.5.3 Histogram of Candidates' Test Scores

A histogram representing the plotted student scores on the examination is provided (illustrated below).



Purpose: If the user finds the previous tabled frequencies too detailed, a graphic representation will normally reduce the amount of detail and also aid in visually estimating if individuals or groups cases are sufficiently different to merit different grades or if atypical cases merit exceptional consideration.

2.5.4 Overall Test Statistics

Following the histogram, the report specifies the overall test's mean, variance, standard deviation, reliability and standard error of measurement.

NUMBER OF EXAMINEES	133
TOTAL SCORE	
MEAN	18.83
VARIANCE	46.29
STANDARD DEVIATION	6.80
KR-20 RELIABILITY	.8359
S.E. OF MEASUREMENT	2.7566

Purpose: These summarized statistics of the overall test have multiple purposes, not the least of which provides the user with data summary of the test's overall difficulty (mean), variability across examinees (variance and standard deviation), inter item consistency (reliability) and a helpful reference for estimating if and where examinees' performance are sufficiently different to consider them real differences (standard error of measurement).

2.5.5 Definitions

Immediately preceding the analyses of all items, a suitable reference source is cited for further explanations of each of the computed item statistical properties. This is followed by the definition for each abbreviated statistical term that is used throughout the remaining report.

```
    An overview of item analysis statistics appears in
         Osterlind, Steven J.   Constructing Test Items. 2nd ed.
         Boston: Kluwer Academic Publishers, 1998.


                     G L O S S A R Y

DIF:    DIFFICULTY - proportion of examinees answering the item correctly
RPB:    Point-biserial correlation between item and total test score
CRPB:   Corrected RPB - RPB between item & total score not including present item
CON:    Confidence limits on CRPB
RBIS:   Biserial correlation between item and total test score
CRBIS:  Corrected RBIS
IRI:    Item reliability index - RPB times square root of DIF times one minus DIF
N:      Number of examinees in group
INV:    Number of examinees not providing a valid response to this item
NF:     Number of examinees not finishing the test from this item onwards
OMIT:   Number of examinees omitting this item
*:      indicates the correct answer (keyed response)
HIGH:   Approximately  27.0% of the total group scoring highest on the total test
LOW:    Same proportion as 'HIGH' having lowest scores

TEST SCORE MEAN %: Mean % on the total test for the individuals giving the
                   Indicated response to this item

DISCRIMINATING POWER: The difference between the proportions of the HIGH and LOW
                      groups giving this response

STANDARD ERROR OF D.P.: Standard error of discriminating power
```

Purpose: IDEAL does not keep all possibly relevant item performance properties. Within the above, the user may find alternative properties that s/he considers also useful or needed.

2.5.6 Item Analyses

The item analysis program accommodates questions that have up to 26 choices. Due to space requirements for printing, only the options that are chosen by at least one examinee are tabulated.

Items, which the best performing group has frequently chosen a particular incorrect response more often than the correct response, should be examined closely. Items that the poorer performing group was able to select the correct response more frequently than the high group, also should be reviewed as to why this unexpected pattern emerged.

The following is a statistical report for 2 items (#34 and #37) in an example test.

```
ITEM  34: DIF= .459, RPB=   .480, CRPB=   .407 (95% CON=   .255,  .540)
                     RBIS=  .603, CRBIS=  .511, IRI= .239
     GROUP    N   INV   NF  OMIT     A       B*      C       D
     TOTAL  133     1    0    1     .12     .46     .24     .17
      HIGH   37     0                .00     .78     .14     .08
      MID    56     0                .14     .43     .21     .21
      LOW    40     1                .20     .20     .38     .20
     TEST SCORE MEAN %               64      74      66      67
     DISCRIMINATING POWER          -.20     .58    -.24    -.12
     STANDARD ERROR OF D.P.         .07     .11     .10     .08


ITEM  37: DIF= .391, RPB=   .041, CRPB=  -.031 (95% CON=  -.200,  .140)
                     RBIS=  .052, CRBIS= -.039, IRI= .020
     GROUP    N   INV   NF  OMIT     A       B*      C       D
     TOTAL  133     2    1    1     .21     .39     .22     .17
      HIGH   37     0                .19     .41     .19     .22
      MID    56     0                .25     .38     .21     .16
      LOW    40     2                .17     .40     .25     .12
     TEST SCORE MEAN %               65      66      60      68
     DISCRIMINATING POWER           .01     .01    -.06     .09
     STANDARD ERROR OF D.P.         .09     .11     .09     .09
```

In the first two lines for each item analysis, the psychometric values previously defined are printed (difficulty level, point-biserial, etc.). Refer to the list of definitions if the abbreviation remains unfamiliar or to the reference source if its nature or purpose is unknown.

The proportion of students selecting each choice is then listed – organized by the total and the top, middle and bottom groupings as previously defined. Please note the discrimination power associated with the correct choice: an item that discriminates in the intended direction will have a reasonably high (positive) discrimination index associated with the correct choice.

Purpose: among the numerous uses these data have, diagnosing the quality of the item is aided. For example, item 34 has a discrimination index of 0.58 (difference in proportion of top 27% who could answer this item correctly and the proportion in the bottom 27% who could answer item 34 correctly: i.e., 0.78 – 0.20 = 0.58. However, for item 37, the discrimination index is only 0.01 (i.e., the difference between 0.41 and 0.40). That is, those who did get the answer were as likely to be poorer performers as good performers. The difficulty level of 0.391 indicates that only 39% of students answered the question correctly.

Thus, regardless of one's intention (setting either a criterion or normative reference assessment), the content to which item 37 relates was, in all likelihood, either not taught well, the associated content is of suspect relevance and/or the item has inherent problems in composition.

2.5.7 Scatter Diagrams

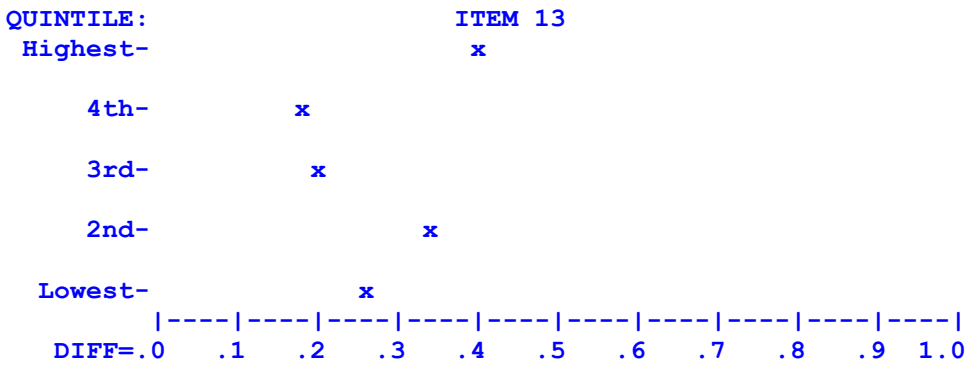
The user has the option for printing scatter diagrams, in which the difficulty levels of each item are plotted by 5 levels of examinee performance. In the examples below, the plots for items 12 and particularly item 34 illustrate that the items perform as expected, while item 13 may require closer review and possible revision.

```
PLOT OF TEST SCORE GROUPS (QUINTILES) BY ITEM DIFFICULTY
The sizes of the quintile groups from 'Highest' to 'Lowest' are:
25        27        25        32        24


  QUINTILE:                   ITEM  12
   Highest-                                        x

       4th-                          x

       3rd-                       x

       2nd-                             x

    Lowest-      x
           |----|----|----|----|----|----|----|----|----|----|
     DIFF=.0   .1   .2   .3   .4   .5   .6   .7   .8   .9  1.0
```

Purpose: the above and following graphs illustrates how one is able to see if an item is performing in an expected or desired manner. In the above graph, over 80% in the top group answered item 12 correctly. In the middle three groups, about 50% answered the item correctly. In the bottom group, only 10% of the examinees correctly answered the question. Thus, item 12 discriminated well among the better, average and poorer performing examinees (at least as defined by their performances on the total test). Item 34 (below) appears to have discriminated even more so, i.e., across all five performance levels (quintiles).

```
 QUINTILE:                    ITEM  34
   Highest-                                        x

       4th-                                  x

       3rd-                       x

       2nd-                  x

    Lowest-         x
           |----|----|----|----|----|----|----|----|----|----|
     DIFF=.0   .1   .2   .3   .4   .5   .6   .7   .8   .9  1.0
```

However, it is likely that item 13 (plotted below) has some inherent problems (and/or the adequacy of the teaching for the related content is in question). Two of the top three groups perform more poorly than the bottom two groups. Closer examination of the item, the item key or the teaching of the item's related content may reveal why this unexpected distribution occurred.



2.5.8 Relevant Resources

Inquiries about scoring and analyzing assessments often begin with the question "Which references provide helpful advice on the construction of quality assessments?"

There are numerous publications available, some of which are provided in this manual's list of references. Among these, the IDEAL Consortium highly recommends the text published by the National Board of Medical Examiners and written by Drs Susan Case & David Swanson.

2.6 Addendum: How to Use IA if an Optical Scanner is Unavailable

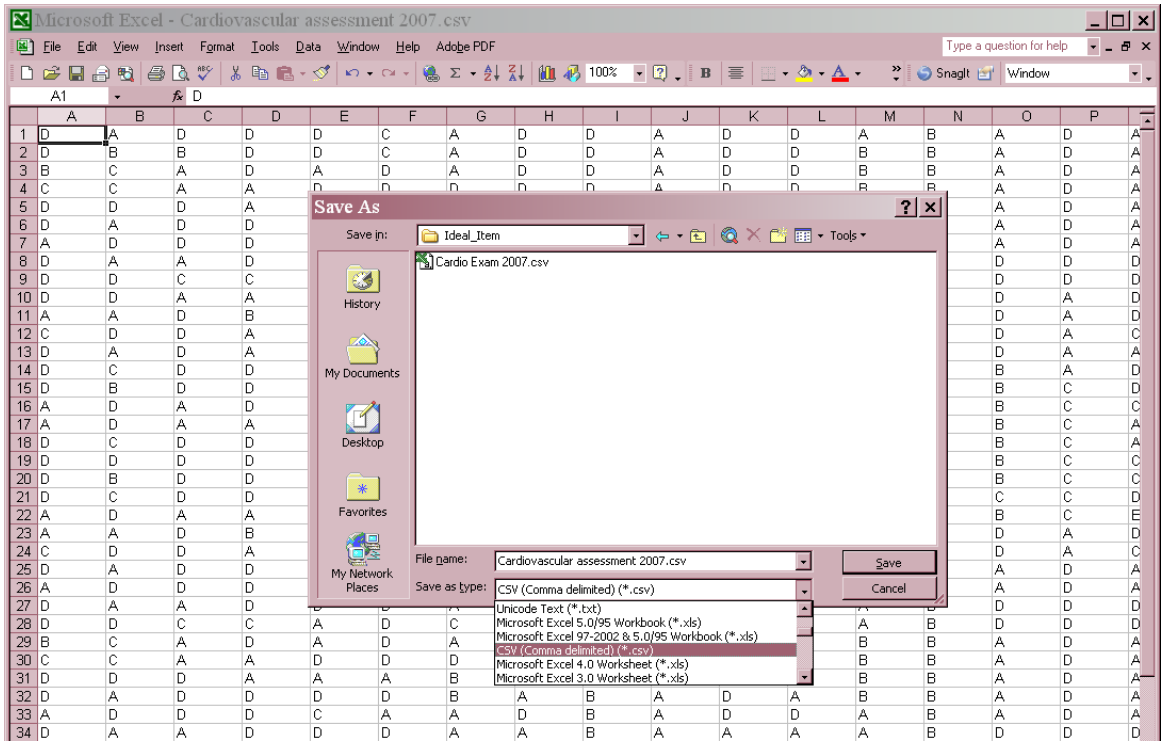
If an optical scanner cannot be used to generate an appropriate data file containing the key and the candidates' responses to the items, users will likely prefer to enter these data into an Excel spreadsheet (refer to following Sections 2.6.1 and 2.6.2). Alternatively, one can use Windows Notepad or MS Word to create a data spreadsheet (refer to the subsequent Sections 2.6.3 and 2.6.4).

2.6.1 Prepare Data as a Comma Delimited (*.csv) File Using Excel

Launch MS Excel and create a new spreadsheet. In the first row, enter the key (correct answers) for each item. Each column in the Excel spreadsheet represents a different item. For subsequent rows, enter the candidates' responses. Each of the rows from 2 to 'n' (n being the total number of candidates) represents a different candidate. Each column within the rows is used for the same item as specified in row one for the key.

In the following example, assume one administered an examination composed of 50 A-type items, five choices for each item were coded as A to E. An Excel file with the key and candidate's responses is illustrated below.

The key (a list of correct answers) cannot be located anywhere other than the first line; if the key exists elsewhere in the file, cut and paste appropriately so that the misplaced key becomes the first row in the file.



If the user does not wish to use the default option whereby the candidate's total test score is the overall measure of performance (for purposes of calculating unadjusted and adjusted biserial and point biserial coefficients), a reference standard for each candidate is entered in the column immediately after the column reserved for the last item in the test.

If you wish to enter the candidates' IDs, enter these data as the last column in each row.

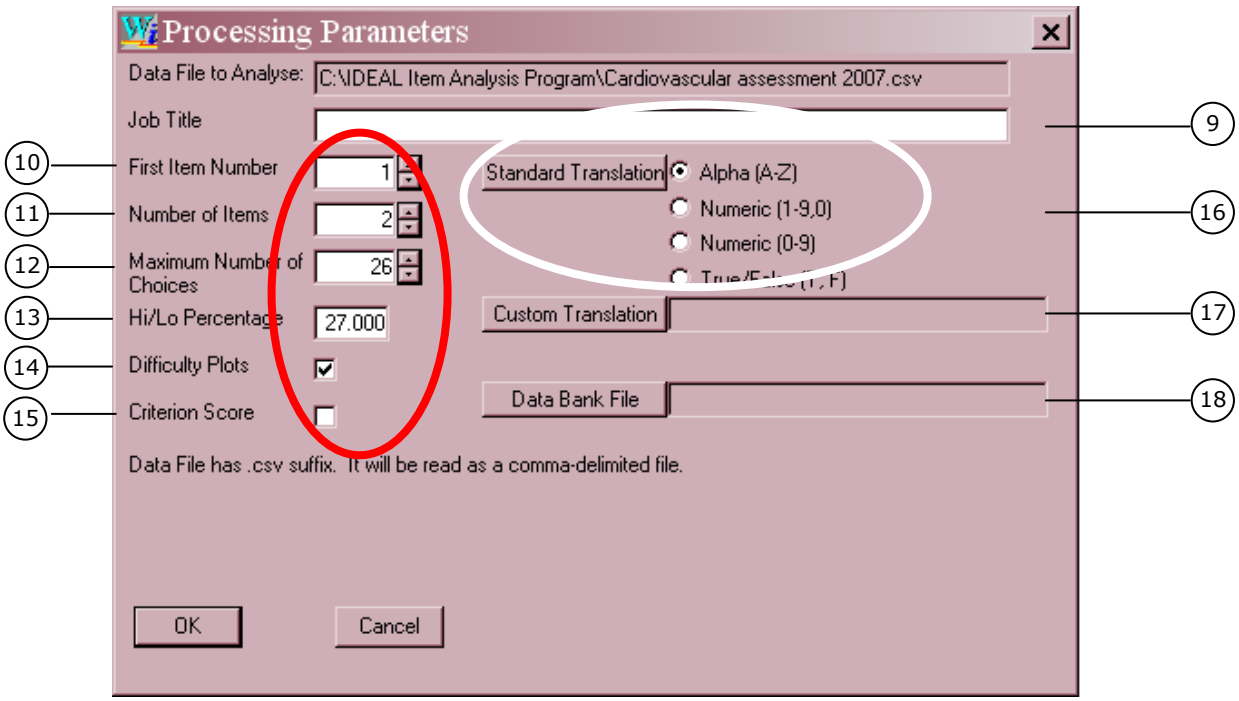
If there are rows or columns in the spreadsheet which contain information other than these above noted data, delete the non essential rows and/or columns.

After entering the data, save the spreadsheet as a comma delimited file (i.e., click "Save As" and scroll down to and select the extension *.csv). Save the file in either the IDEAL_Item subdirectory or a directory of one's own choosing. Specify the same filename that was used when the items were selected from the IDEAL bank *and printed*. That is, use the course name or module number as the filename. As an extension of *.cvs is used (e.g., *Cardiovascular assessment 2007.csv*), both files will be maintained and can be easily matched when updating the banked item's psychometrics.

A file developed according to these above guidelines is readable by IDEAL's item analysis program.

2.6.2 Analyse Common Delimited (*.csv) Data

If one is using a comma delimited file (*.csv) for data input with IA, the following illustrated window will be displayed when setting the parameters for running an item analysis. The required features designated 9 to 18 in the following illustration are addressed exactly in the same manner as described for data (*.dat) files. Refer to Section 2.2.1., steps 9 to 18).



After confirming these specifications are correctly entered, click the button labeled "OK". You will be returned to IA's main page. There you execute the analysis in the same manner as described for *.dat files (see Section 2.2.1, step 20 (p.11)).

If one wishes to exclude some items in the item analyses, simply change the *first* line (i.e., key) by leaving blank the columns that correspond to items that should be ignored. For example, assume items 5 to 15 reside in columns 10 to 25 and are to be omitted. Thus, blanks should be entered in columns 10 to 25.

2.6.3 Create Data (*.dat) File Using Notepad or Word

This option for preparing one's data for analysis is tedious and data entry errors are more commonly made, and thus is not necessarily recommended. However, if one cannot generate a comma delimited file and also an optical scanner is not available, one can generate a data (*.dat) file by doing the following.

Open Notepad or Word and create a new document.

1. The file's first line specifies the key (i.e., the list of correct answers for all items to be analyzed).

2. Each subsequent line (i.e., row) corresponds to a different candidate's responses.

3. Each character space (column) in a line corresponds to the keyed and candidates' answers to each item in the examination.

4. Answers can be coded numerically (e.g., 1 to 9). However, if the maximum number of choices for any item exceeds one digit, options must be coded as alphabetic characters (e.g., A to Z).

5. If IDs and/or a Criterion Score (reference standard) are to be entered, follow the same requirements as described for data (*.dat) files generated by an optical scanner (see Section 2.1.2)

Save the file with the optional extension of *dat (not *.doc or *.xdoc, for example)

2.6.4 Analyse Data Files Created wit Notepad or Word

Follow the same steps described for analyzing data files that are generated by a scanner (i.e., refer to and follow the instructions provided in Section 2.2).

Chapter 3

Item Analysis for Constructed Response & Multiple Selected Response Items

3.1 Introduction

This chapter describes how to use IDEAL's item analysis program (IA) and interpret the psychometrics that are derived from an analysis of candidates' responses which have been marked (scored) in a non dichotomous fashion (i.e., the candidates' scores on each item are continuous data, not simply right (1) or wrong (0)).

Non-dichotomously scored items normally include the constructed response formats of Short Answer Questions (SAQ), Modified Essay Questions (MEQ), OSCE and OSPE stations. Scores from the selected response format of Multiple Response (Pick n of N for A- and R-type) as well as Key Features also yield nonbinary data.

3.2 Similarities and Differences in Analysing Binary and Continuous Data

The reported item analyses for constructed response items is similar to those reported for the analysis of selected response items such as A-type and R-type items. For both, IA produces a frequency distribution and histogram of the total test scores, and for each item the statistics include difficulty levels, item-total test correlations, an analysis matrix based on high, middle and low-scoring groups, and item-difficulty plots.

For selected response items, a matrix displaying the percent of the low, medium and high overall performance groups selecting each option in an item is provided. For continuous data, the corresponding matrix is the same three performance groups cross-tabulated with different levels of achievement on each item. For the printed output the user has the option to specify the number of levels of achievement (i.e., performance categories) for this matrix. Achievement can be reported in terms of the actual awarded marks or as percentages, percentiles or standardized scores (z-scores). However, quintile groupings are always generated for these matrices that will be stored with the relevant banked items in the IDEAL system (i.e., regardless of the user's choice for a printed output).

IA reports the Cronbach's alpha as the test reliability coefficient for constructed response items (whereas with dichotomously scored items, the Kuder Richardson's KR20 is reported, a special case of Cronbach's alpha).

The difficulty index for each item is the average performance on the item regardless of how the item is scored. With binary data, the sum of the difficulty indices is equal to the mean of the test. For continuous data, difficulty indices require weighting by the respective maximum values of the items before summing to obtain the mean for the total test.

The item discrimination index for selected response items is the difference between the success rate on an item by candidates in the high-performing group (default is P_{73} and above) and those in the low-performing group (default is P_{27} and below). For constructed response items, the item discrimination index is the difference between the average scores (expressed as a percentage) for these same two groups.

For dichotomously scored items, the point-biserial correlation represents the level of association between performance on item and the total test. The point-biserial is a special case of the Pearson's product-moment correlation which is the appropriate coefficient to use with continuous data (and accordingly is reported by IA for constructed response items).

The sum of the item reliability indices is equal to the standard deviation of total test scores. This holds true for both dichotomously and non-dichotomously scored items. Thus, IA's reported item reliability index for both selected and constructed response items reflects the item's contribution to the standard deviation of the total test.

Missing data is normally scored as 0 (wrong) with selected response items. In calculating means, IA also ascribes zero for missing data in constructed response items. However, in order to be able to calculate various correlations, IA substitutes the item mean for missing or invalid data for the constructed response item. When item statistics are reported for each unique item score, if an item has any missing data, the estimated scores are isolated in the reports by appearing in their own special column under the heading "est.".

Guidelines for preparing and analyzing data files of the candidates' marks, and for understanding the item analysis output are discussed in the remainder of this chapter.

3.3 Prepare a Data File

Usually, constructed response items cannot be optically scored. Thus, examiners will need to first mark the candidates' responses to each question. These assigned scores should be entered into a spreadsheet using MS Excel. IA accommodates marks that are expressed as whole numbers (e.g., 5, 55, 193, etc.) or as partial marks specified with decimals (e.g., 4.5, 50.75, 187.3). There is no limit on the number of decimals used in the data set but more than two or three is normally not justified given the level of precision that can be achieved when assessing a candidate's abilities and skills, especially with constructed response items.

To begin creating a file which can be item analysed for constructed response items, Open MS Excel and create a new spreadsheet.

The first row in the spreadsheet must indicate the maximum mark possible for each item (i.e., for each SAQ, MEQ or station). Each item's maximum score is entered in a separate column. In the remaining rows of the spreadsheet, enter the candidates' awarded marks for each item, using exactly the same columns for each item as specified in the first row. A separate row is allocated for each candidate; thus, the total number of rows will be one more than the number of candidates.

If you wish to use a reference standard in lieu of the total test score as the measure for a candidate's overall ability, enter this criterion score after the candidate's score on the last item in each row.

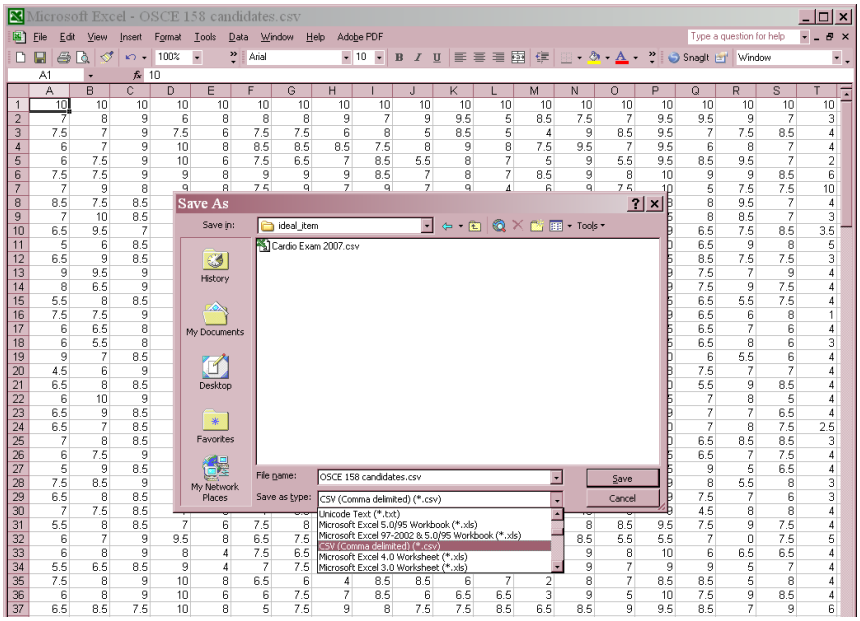
If you also wish to include the candidates' IDs, enter each ID as the last variable in each row.

After entering all the data, save the spreadsheet as a "Comma delimited" file (i.e., click "Save As" and scroll down to the choice with the extension *.cvs).

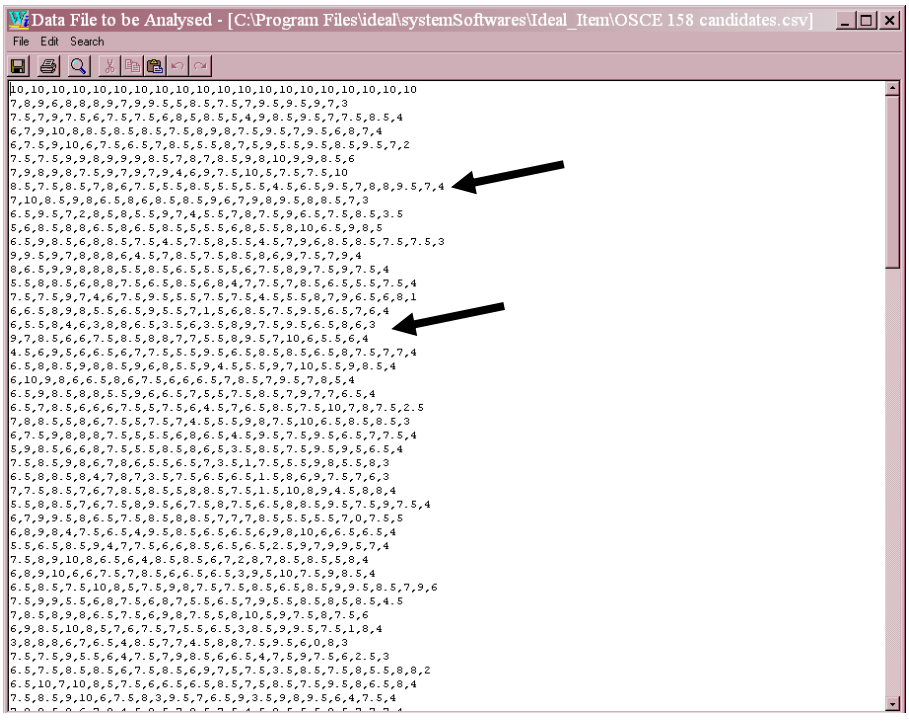
The following example (see illustration on the next page) is a spreadsheet produced by using Excel. The file contains marks for a group of candidates who were assessed by a 20 station surgical OSCE. As each of the 20 stations had a maximum score of 10, this value was entered into each column A to T in the first row.

The marks awarded to the candidates in each station were entered in the subsequent rows, beginning with row 2. Each candidate's marks are in separate rows. Awarded marks for some candidates include partial scores represented by one decimal place. No criterion score or candidate ID has been included.

The file was saved as a comma delimited (*.csv) file as illustrated below.



When the file is opened and displayed in Excel, it will appear as displayed in the above image. However, if this same comma delimited (*.csv) file is opened in IA, the data will appear as displayed in the following image.



A comma delimited file contains variables that are continuous data, separated only by commas. Thus, each row is not necessarily of the same length (note arrows). In this example, some rows have only or mostly whole numbers while other rows have several entries with decimals (making the data row appear longer). However, both of the above displayed data sets include exactly the same data from the 20 station OSCE.

3.4 Theoretical Considerations in Selecting Data to be Analysed

Theoretically, one could treat each task within each station as a separate item (or treat each question within a modified essay as a separate item). If this approach for an item analysis is adopted, the marks for each required task in the station (or MEQ) would need to be entered into the spreadsheet and the station (or MEQ) treated as the overall assessment. This statistical approach is justified if one's goal is to evaluate if each of the item's components was appropriately designed, administered and scored. If needed, one would refine some of the specific tasks in the station (or some of the questions within a MEQ) that were insufficient.

However, in terms of sharing banked assessment items, the underlying goal is different. Users normally do not search for and select just one of the tasks within a particular station (or one of the sub-questions within a MEQ). Rather users search and select a whole station or MEQ. Thus, the relevant psychometric properties need to reflect the assessment quality of the entire station or MEQ. To generate relevant statistical properties for the whole station or MEQ, the overall mark that each candidate achieved for a station (or for a MEQ) needs to be entered in the spreadsheet for analysis.

Given this perspective, the psychometrics for all constructed items stored in the IDEAL bank should reflect the properties of the overall station, MEQ or SAQ.

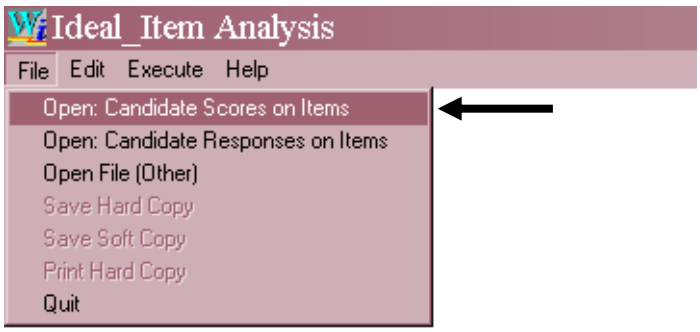
3.5 Analyse Candidates' Scores

To analyse data entered into an Excel spreadsheet and saved as a *.csv file (see Section 3.3), take the following steps and in the sequence indicated. For purposes of illustrating these steps, the 20 station OSCE displayed in Section 3.3 is used.

1. Click IA's icon on the desktop.

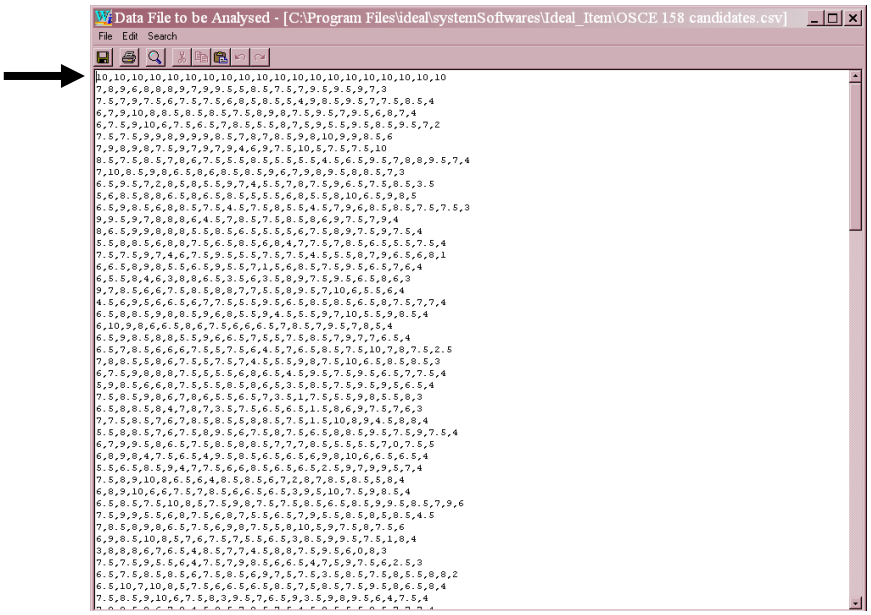
   The following illustrated window opens. An inserted message box will also appear (not illustrated), providing an overview on how to use IA. This message will disappear as you click "File" in IA's main menu.

2. Click "Open: Candidate Scores on Items" (in the drop down menu).



3. You will be shown the default directory in which normally you will find the data file that you want to use (i.e., the *.csv file); if you previously saved the file in another directory, navigate to the appropriate directory.

   Double click the desired filename and the comma delimited data file is displayed.



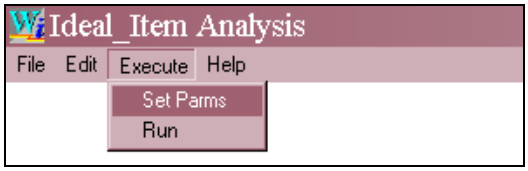
4. Confirm that the opened data file is the one you intend to analyze. If you wish to modify the data file, click Edit in the main window of IA and make any required changes to the data file.

5. In the first row, place your cursor in front of the number that represents the maximum mark for the first item (i.e., SAQ, station or MEQ) to be analyzed. If you wish to skip some initial items, appropriately locate the cursor anywhere to the right in row one.

6. If you wish to omit particular questions for any items that follow the cursor's placement, leave the relevant column in row 1 blank. That is, items that are keyed with a blank (in row 1 only) will be omitted in the item analysis.

7. In the Ideal_Item Analysis window, click File, and then Exit; however, if you modified the data file (step 5) you must click Save before exiting. After exiting you will be returned to the main menu in IA.

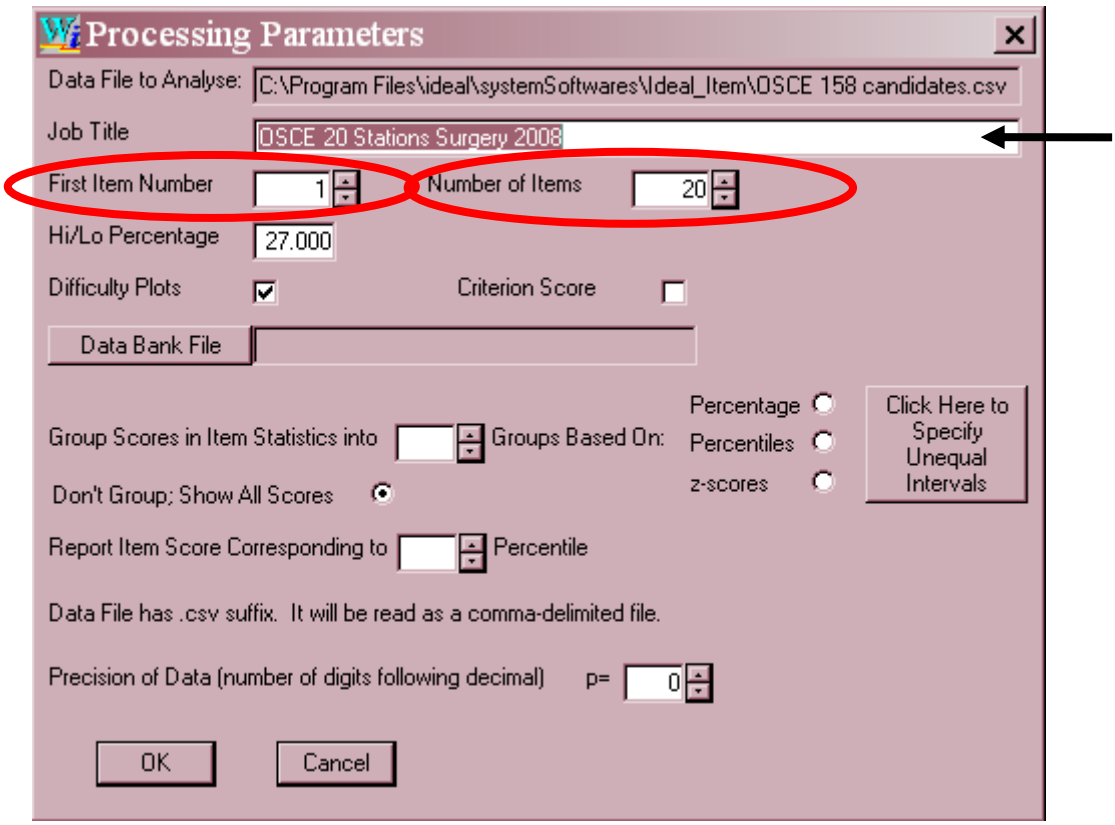
8. Click "Execute" and then "Set Parms" in the drop menu. If you click "Run" without having set the parameters, the program will prompt you to specify these parameters before the analysis can be executed.



The Processing Parameters window opens (illustrated below). In this window, the user must indicate how the item analysis should be computed. In this window, you may use the tab key to navigate from one field to the next.

If you attempt to enter a value that is out of range for a particular feature, the computer's bell will ring and the invalid value will be highlighted. If this occurs, revise the relevant entry.

9. In the open text box to the right of "Job Title", enter descriptive text that you want to appear at the beginning of the printed output.



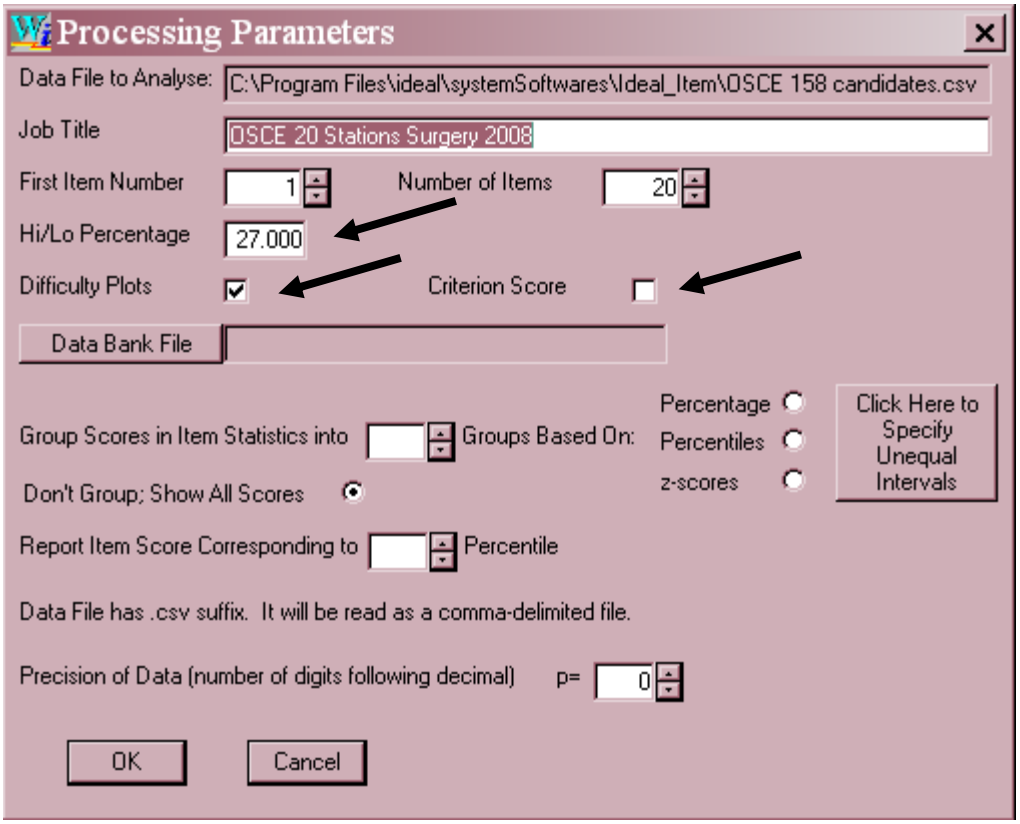
10. Enter the number of the first item to be analyzed to the right of the label "First Item Number" (see left circled area in the above illustration). This number is needed to ensure the labeling used in the output is appropriate.

    a. The default number '1' is correct if the entire test is being analyzed.

    b. If you skip some initial items (for example, items 1 to 10), enter "11" and the printed output for each analysed item is correspondingly labeled 11 to 20 (not 1 to 10).

11. To the right of the label "Number of Items", specify the total number of items in the assessment; in the example illustrated, the number is 20 (see right circled area). This total number includes all items following placement of the cursor in row 1, including any that you subsequently might wish to have the program ignore for certain analyses (i.e., include the relevant columns in the key (row 1) that are located right of the cursor and are blanked). In other words, for a test that begins with item number one, the 'Number of Items' should be the last item number in the test even if some have the keyed answer blanked.

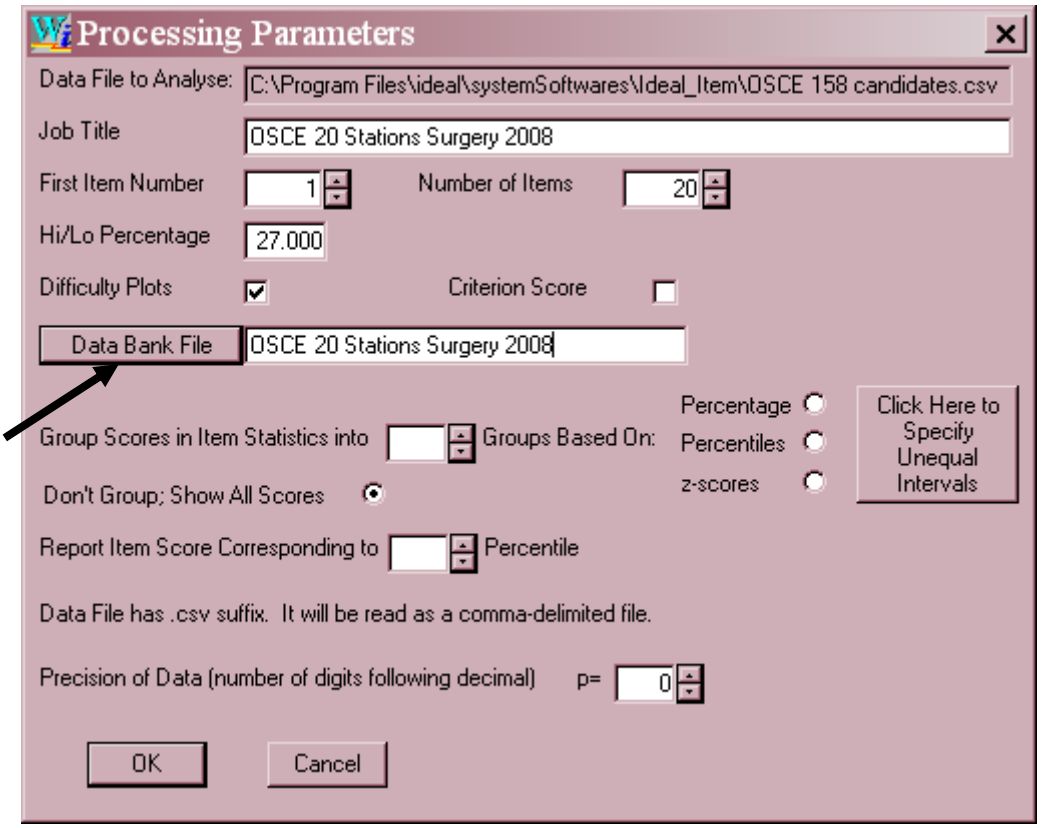
12. To the right of the label "Hi/Lo Percentage", specify the percentage of candidates that will constitute the size of the high and low performance groups (see top arrow in the following illustration). The default is 27% as this corresponds to a correct computation for the classical discrimination index. The exact percentage of candidates used may be approximated by IA due to rounding error, possibly resulting from presence of a small group of candidates and/or because of tied scores at the cut point.



13. Check the appropriate box if you wish to have item difficulty by performance groups scatter plots printed (see lower left arrow in above illustration). Selecting Difficulty Plots will provide a plot of the difficulty of each item against a quintile grouping of total test scores. As with Hi/Lo Percentage, these quintile groups might not be exactly of equal size due to rounding error or tied scores.

14. Check the box beside "Criterion Score" whenever you wish to use a reference standard other than the total test score as the measure of the candidate's overall performance ability. (Default is unchecked implying the candidates' total test scores will be used; if checked, ensure the data file created in Section 3.3 includes this criterion measure, normally entered after each candidate's answers. In this example, the "Criterion Score" is unchecked (see lower right arrow in the above illustration).

15. Note the button labeled "Data Bank File". Click this button if you want to create a file that IDEAL can read for importing the relevant statistical properties to store with the appropriate items kept in IDEAL's item bank.



After clicking the button, the adjacent text box to the right will be accessible. Enter a file name. Use the same filename as assigned when the items were selected from the item bank *and printed* (i.e., use the same name that you specified for the course or module number when you printed the exam within IDEAL's "Print Exam" window). (Remember also that in that latter window you had to also check the box on the right "All items will be scored nondichotomously).

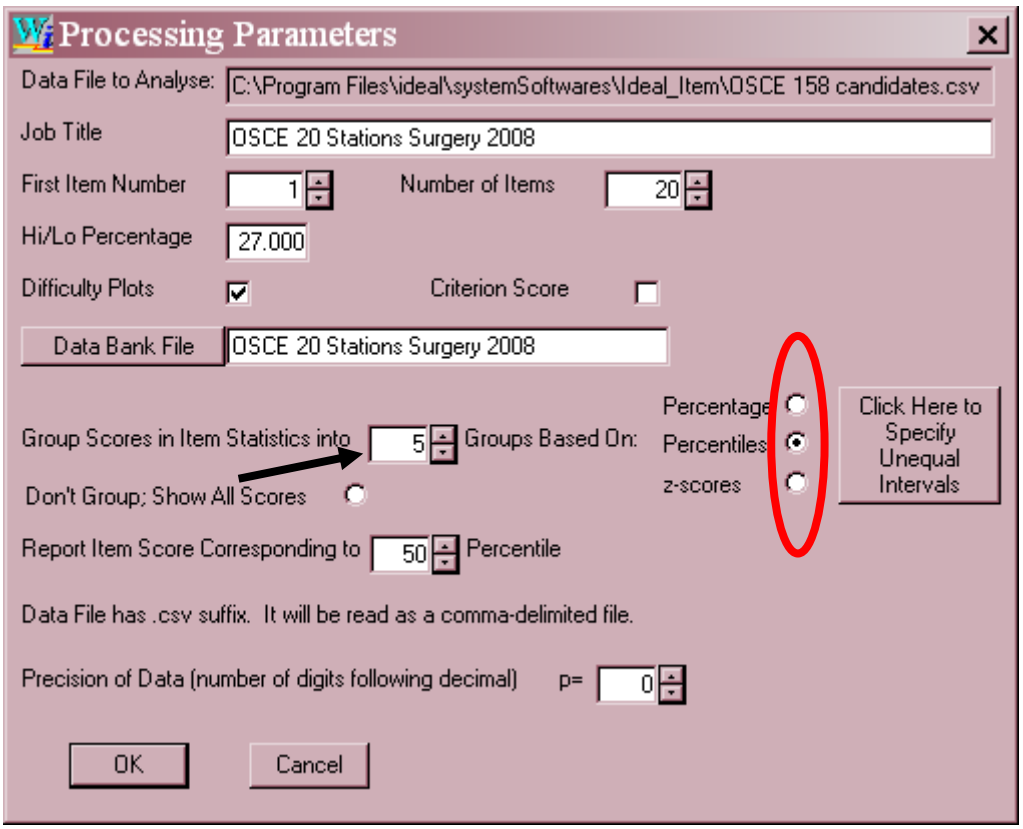
Later (i.e., after running IA and confirming the analysis is correct), you will need to click "File" (in IA's main page) and then click "Save Soft Copy". This soft copy file is saved with the same name provided for the course or module number entered in IDEAL's "Print Exam" window. This file will have the extension "*.*idb*" (which represents "*i*deal *d*ata *b*ank") whereas the file in IDEAL with the same name (Course or module number) will have the extension *.sss.

When importing the item analysis statistics using the IDEAL program, you will need to access both of these files. Using the same file name ensures you can match the files easily. The steps required for importing the item analyses are described later (Section 3.6).

Note that if a filename for the Data Bank File is omitted, a softcopy will not be created. Thus, if one's purpose is to generate a file of item statistics so that IDEAL can update the relevant psychometric properties for all banked items used in the examination, the procedures specified in this step must be completed.

16. The "Group Scores in Item Statistics" feature allows one to define up to 20 categories of performance. The following example uses five categories (note the arrow in the following image). The categories can reflect:
    - percentage ranges (of equal intervals unless otherwise specified),
    - percentile ranges (of equal number of respondents unless otherwise specified), or
    - ranges of standardized (Z-) scores (of equal distance, unless otherwise specified).

    The user specifies which of these three above options is wanted in the printed output. Click the appropriate radio button (see circled area). Note, however, that regardless of one's selection for a printed output, the softcopy file which is read by IDEAL will be based on quintiles (i.e., five equal percentile ranges).

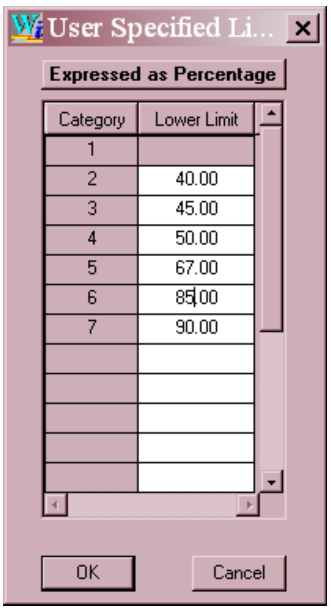


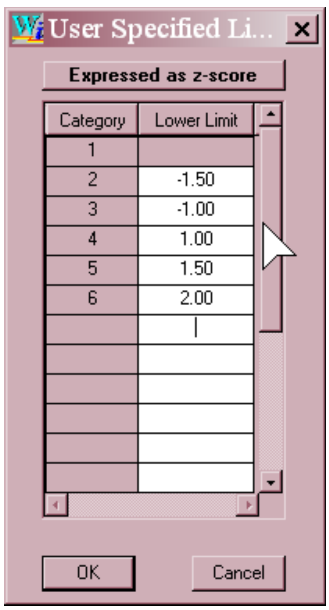
17. To access the option for defining unequal ranges for the performance categories, click the rectangular button labeled "Click Here to Specify Unequal Intervals" (located immediately right of the circled radio buttons).

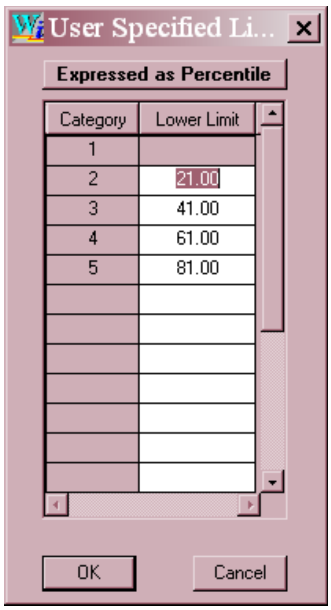
    Clicking the rectangular button opens a separate window (see illustration on next page).

After clicking the rectangular button "Click Here to Specify Unequal Intervals", one of the windows illustrated below will appear (depending on which grouping criterion and number of performance groups you previously specified). The first group will consist of subjects falling below the lower limit specified for the second group, and so forth for the remaining number of groups. Note that in all cases, the first row of the grid is greyed. That is, membership in the first group is defined as all values falling below the lower limit specified for the second group.

Put the cursor in each open category. Type in the cut-off points and then click the mouse. This ensures the entry is properly recognized by IA. If your entries appear similar to the illustrated examples then the cut-off points have been read by IA.







The percentage and z-score windows will have defaults provided. These defaults may be over-ridden. If you attempt to enter a value that is out of range for the given grouping criterion, your PC's bell will ring and the invalid value will be highlighted.

To obtain a different number of groups than the default, remember to specify the desired number in the box provided in the Processing Parameters window (designated by an arrow in the illustration on the previous page) before attempting to enter the group limits in the above window.

If a number of groups has been specified and the Unequal Interval Limits button has *not* been clicked, equal-sized intervals will be created. If values have been entered in the above window but you wish to reconsider and revert to equal-sized intervals, use the following procedure. In the Processing Parameters window:

- click the radio button for "Don't Group; Show All Scores";
- click the "Click Here to Specify Unequal Intervals" button;
- an error message will be generated but the result will be that the "Number of Groups" field will be cleared as well as the Interval Limits window;
- you can now enter a new value for the "Number of Groups" and specify the grouping criterion to be used.

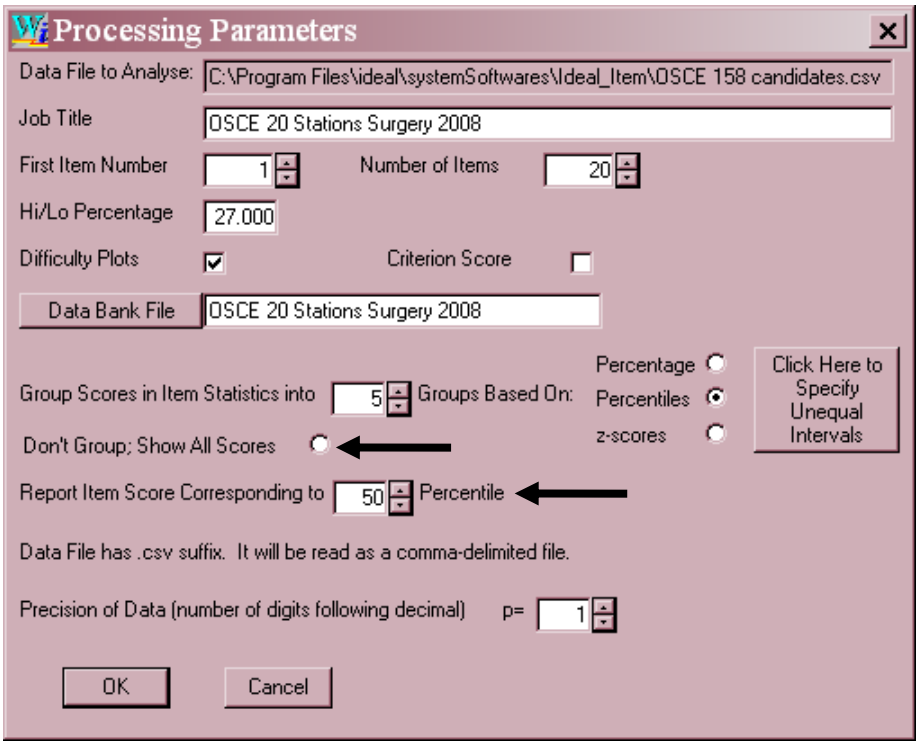
The purpose in using unequal groupings is now explained.

Assume that a school wishes to determine how the high to low performing groups perform on a continuously scored item, but that for a large range of failing marks on the item, no detailed breakdown is of interest. Thus, the first category might be set as 0 to 50%. The next category of just below acceptable is of more interest but the range is much smaller, e.g., 50 to 59%. The next range of interest might include those with a passing but not necessarily a good level of performance (e.g., 60 to 66%). The next range representing a good level of performance might be 67 to 79% with the remaining range of 80% and above as very good.

The same principle applies for defining practical, unequal ranges of performance using percentiles and standardized (Z-) scores. The specified ranges will be educationally meaningful if they reflect a school's policies and related decision-making with regard to identifying failures, providing needed remediation, assigning grades, awarding outstanding performance, and so forth.

18. If the user does not want to create performance groupings on the item, check the radio button immediately to the right of "Don't Group: Show All Scores". This option is normally used only if the maximum possible scores on all items are not large (e.g., 3 to 9). If the maximum scores are large (e.g., ≥21), the output for each item becomes little more than a frequency table for the high to low performers (i.e., too detailed).

    Obviously, if this option is selected, the grouping criteria noted in step 17 are irrelevant and not specified.

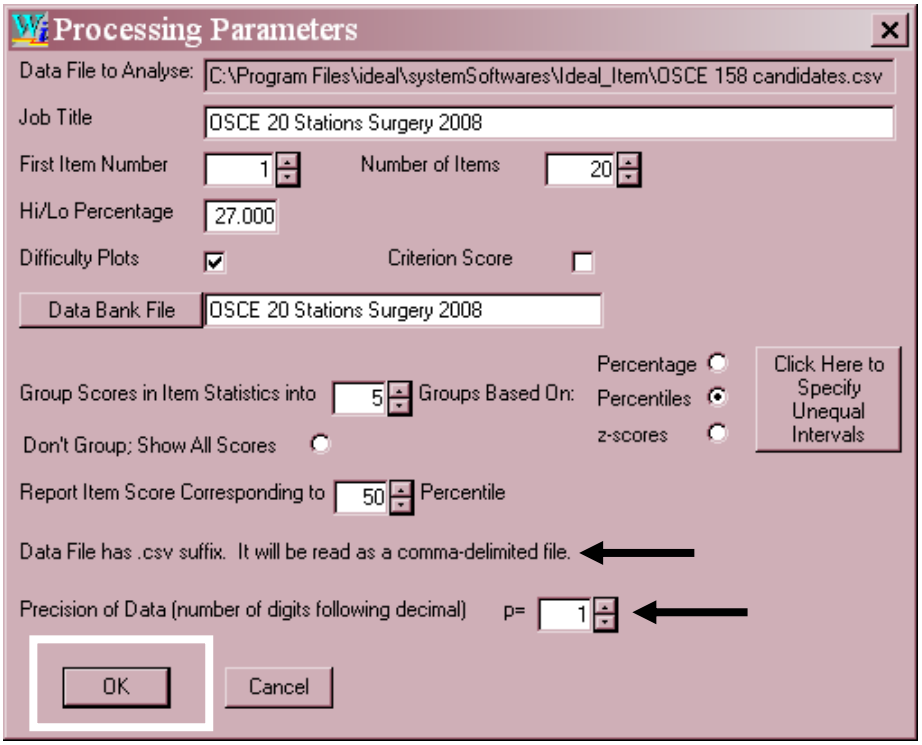


19. The feature "Report Item Score Corresponding to" is used for generating an additional line of printed output for each item --- i.e., the item score corresponding to the percentile specified in the open box; in the illustration, the 50th percentile is specified.

20. The maximum number of decimals that are used to record the candidates' scores on any items must be specified in the small open box in the lower right of the window (next to "P="). In the example data file for a 20 station OSCE, one decimal place was used in recording the candidate's scores. Thus, as illustrated, 1 is entered (see second arrow).

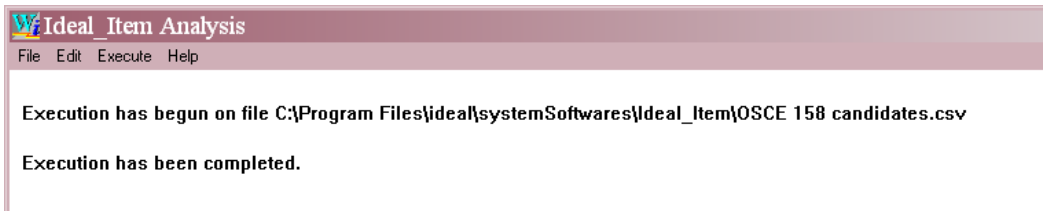
    As the file was created in Excel and saved as a comma delimited file, there is no need for any Format statement[3]. IA's recognition that the file is a comma delimited data file is displayed (see first arrow in the illustration below).

    After confirming that one's parameters have been correctly specified, click the button labeled "OK". You will be returned to the main menu in IA.



21. In the top menu of the main window, click "Execute", then click "Run".

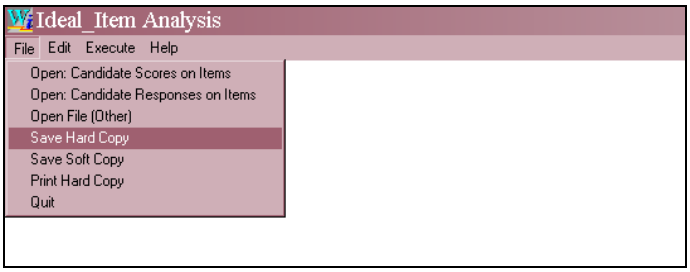
    The item analysis will be completed almost immediately with the following messages being displayed.



[3] IA also provides an option for reading continuous data sets with fixed fields, which are read with Fortran Format statements. A user who is experienced in setting up such data sets, may opt to create this type of file in lieu of a comma delimited file that is easily created using Excel. If chosen, some minor variations will be required at this point in setting the parameters. As this alternate approach is not as user-friendly, nor as easily understood by nonprogrammers, this option in IA is not discussed in this manual. For programmers the required procedures are self-explanatory as one proceeds from data entry to data analysis.

22. After the analysis is completed, confirm the analyses has run correctly. You can read the results on your PC monitor by checking "Edit" in the IA main page and then "Hard Copy" in the drop down menu.

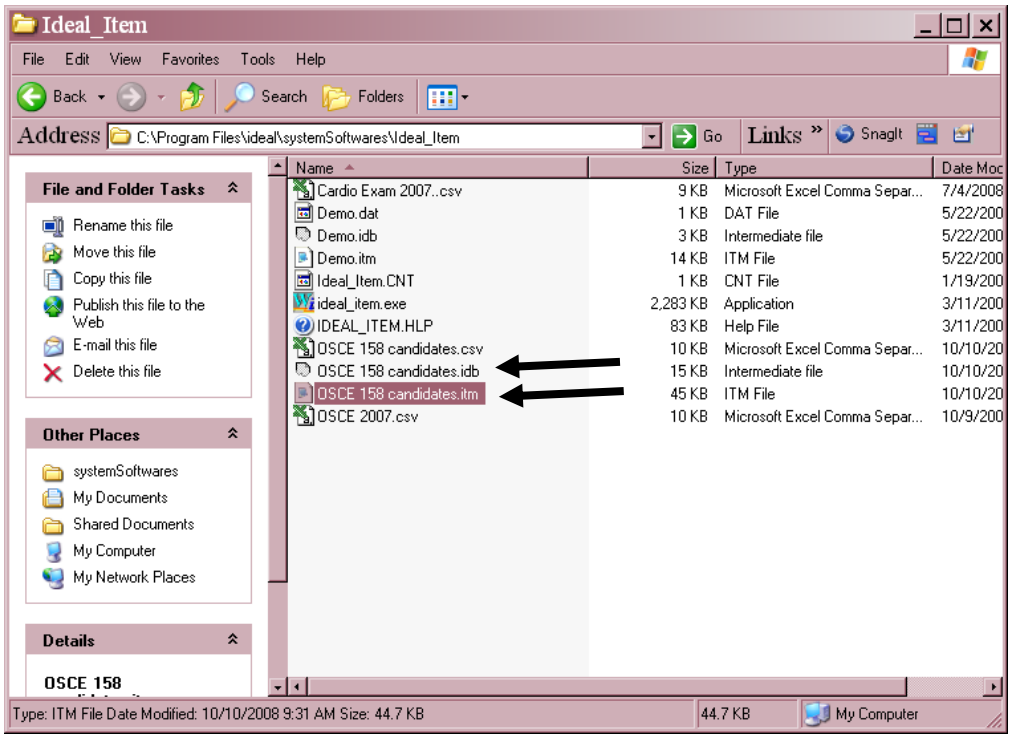
    You can save the hard copy printout file by clicking "File" and select the appropriate command in the drop-down menu. The extension for the hardcopy output is *.itm (representing item analysis).



23. Similarly, save the soft copy file. By default the filename that you specified for the Data Bank File when setting up the parameters will be displayed in the window. The extension for the softcopy file is *.idb (item data bank).

    It is recommended that you save the softcopy output file containing item analyses, using the same filename that you specified for the Data Bank File.

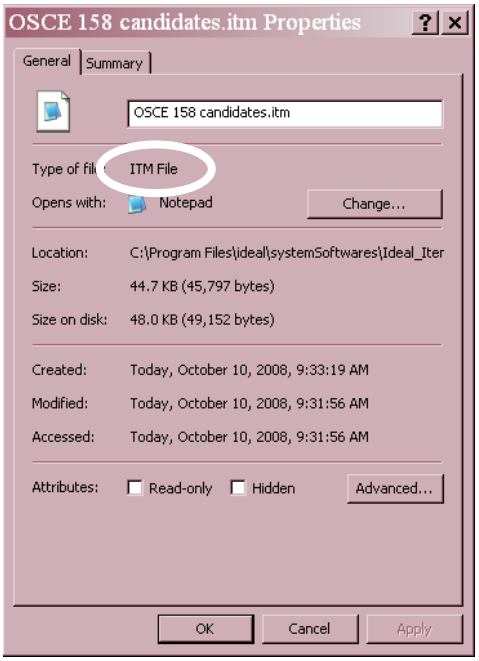
    The default directory in which both of these two files will be saved is C:\Program Files\ideal\systemSoftwares\Ideal_Item. You may wish to select another directory [such as C:\Program Files\ideal\CohortDatabase*(or another database name)*\ExamPapers].



24. Recall that if you did not provide a filename in the Data Bank File, the program will not produce this softcopy file and thus, you will be unable to save it.

    If desired you may open, view and/or print the Soft Copy (*.idb) file (although this file is produced for IDEAL to read and import relevant item psychometrics).To view, on your PC, the contents of either of the above files, navigate to and click to select the appropriate file.

25. If you used the recommended file name for both the Hard Copy and Soft Copy output files you will see that both have the same names which are distinguishable only by their extensions (*.itm for the hardcopy and *.idb for the softcopy). If your PC does not display the file's extensions, hover your mouse over each file and a pop message with the extension will be presented. Alternatively, right click on the files, select "Properties" and the subsequent window will display the extension.



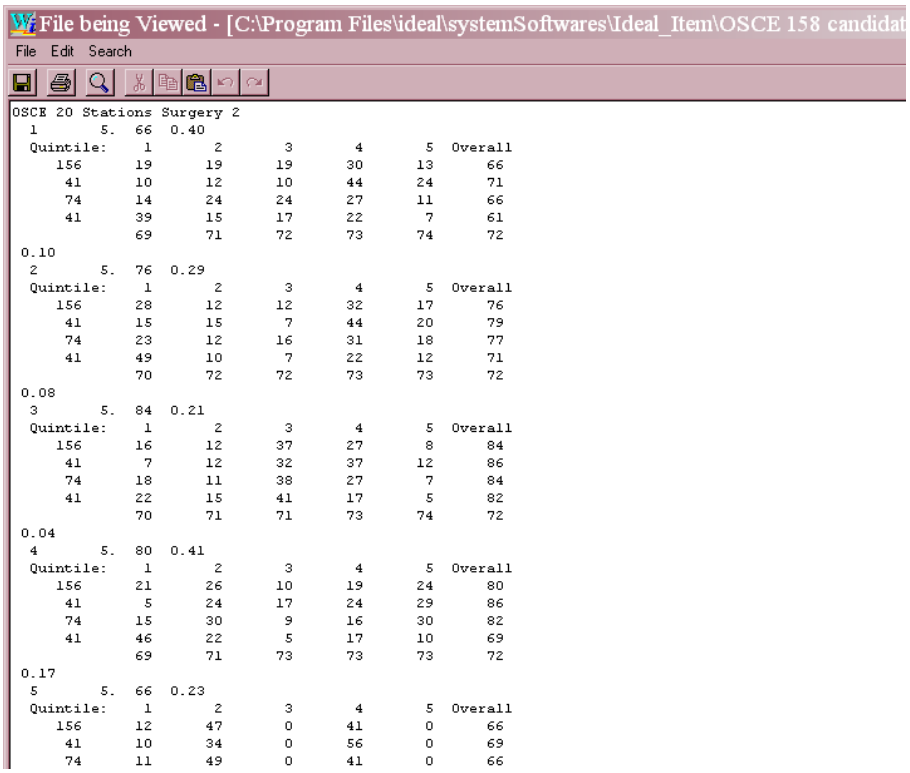
26. To print the Hard Copy output file (*.itm); click "File" and then "Print Hard Copy" in the drop-down menu.

    If the Print Hard Copy function is greyed (i.e., inaccessible), this probably means that even though you may have previously saved this *.itm file, you also subsequently did other analyses, or you closed the application or some other procedure such that the application is unable to determine which Hard Copy to print).

    Provided you have saved the file at some previous point, click "Open File (Other)" and then select the appropriate *.itm file. Click Open. Subsequently, you will see that the Print Hard Copy is now no longer greyed and can be printed.

    The Print Hard Copy option uses a special print routine and scales the output to fit on the page. For further details see Section 2.2.4.

27. The Soft Copy file produced by the item analysis program, when open and viewed on your monitor, appears as:


File being Viewed - [C:\Program Files\ideal\systemSoftwares\Ideal_Item\OSCE 158 candidat

File Edit Search

```
OSCE 20 Stations Surgery 2
 1       5.  66  0.40
 Quintile:    1       2       3       4       5  Overall
    156      19      19      19      30      13      66
     41      10      12      10      44      24      71
     74      14      24      24      27      11      66
     41      39      15      17      22       7      61
             69      71      72      73      74      72
0.10
 2       5.  76  0.29
 Quintile:    1       2       3       4       5  Overall
    156      28      12      12      32      17      76
     41      15      15       7      44      20      79
     74      23      12      16      31      18      77
     41      49      10       7      22      12      71
             70      72      72      73      73      72
0.08
 3       5.  84  0.21
 Quintile:    1       2       3       4       5  Overall
    156      16      12      37      27       8      84
     41       7      12      32      37      12      86
     74      18      11      38      27       7      84
     41      22      15      41      17       5      82
             70      71      71      73      74      72
0.04
 4       5.  80  0.41
 Quintile:    1       2       3       4       5  Overall
    156      21      26      10      19      24      80
     41       5      24      17      24      29      86
     74      15      30       9      16      30      82
     41      46      22       5      17      10      69
             69      71      73      73      73      72
0.17
 5       5.  66  0.23
 Quintile:    1       2       3       4       5  Overall
    156      12      47       0      41       0      66
     41      10      34       0      56       0      69
     74      11      49       0      41       0      66
```

The first record (line) in the Soft Copy file contains the filename that was provided by the user.

The remaining records (lines) appear in sets, one set for each item in the test. Each set consists of seven lines and in each of these seven records, the following data are specified:

The first record in the Soft Copy file contains the Test Identifier label that was provided above.

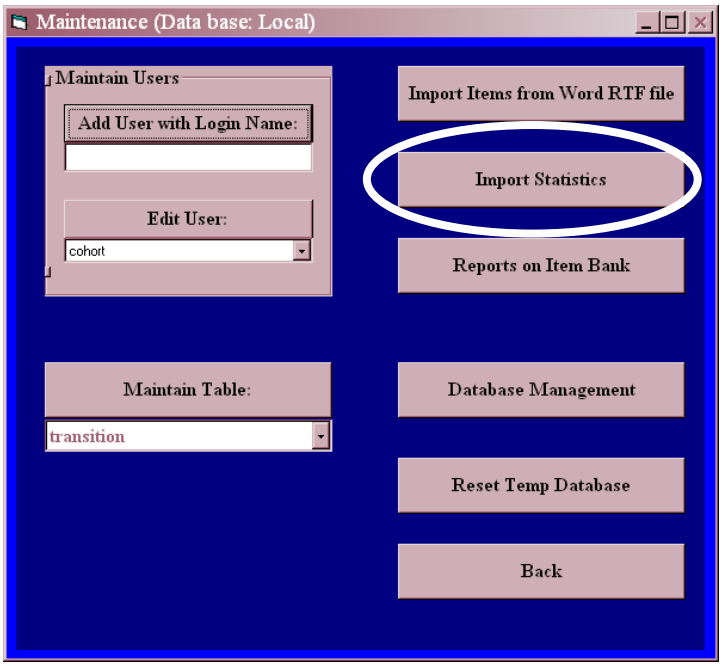
The remaining records appear in sets; one set for each item in the test. The sets consist of:

- Item Number, Maximum Value, Difficulty (expressed as a percentage) and Item-Total Correlation
- A line of column headings for the 5 rows that follow
- For Total group: Number of examinees and percentage of this group with scores in the respective intervals
- For High group: Number of examinees and percentage of this group with scores in the respective intervals
- For Middle group: Number of examinees and percentage of this group with scores in the respective intervals
- For Low group: Number of examinees and percentage of this group with scores in the respective intervals
- Test Score Mean Percentages for the examinees with scores in the respective intervals
- Discrimination index based on the overall average success on the item by the High and Low groups.

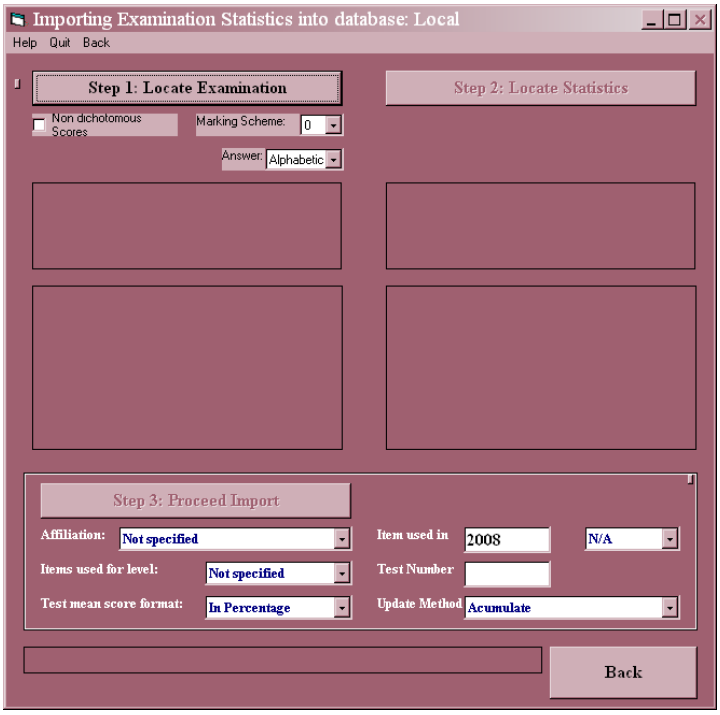
3.6 Importing Statistics into IDEAL's Item Banks

The "Import Statistics" function in IDEAL facilitates importing statistics into the item bank for a large number of questions. One imports the statistics contained in the Soft Copy Output file that has an extension *.idb. This file is generated by IDEAL's item analysis program (IA) when the Data Bank File option is selected in the "Set Parameters" window. To input these item statistics into the item bank follow these steps and in the sequence specified.

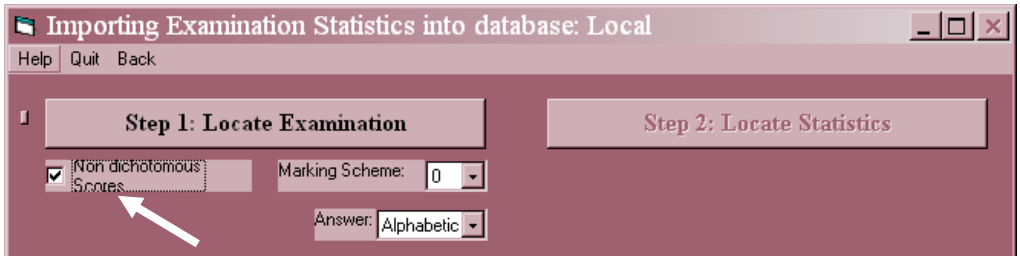
1. In the main menu of the IDEAL PC program, click the "Maintenance" button. The "Maintenance" window opens (illustrated below).



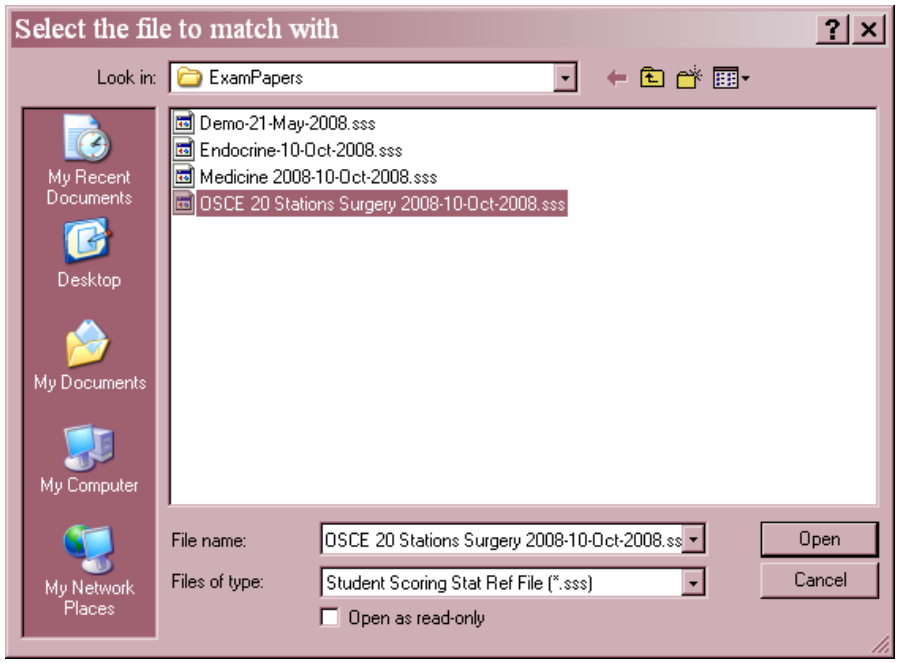
2. Next, click the "Import Statistics" button. The "Importing Examination Statistics into database" window opens (illustrated below).



3. If the item statistics for uploading are for SAQs, MEQs, OSCEs, OSPEs, Pick n of N or Key Features, place a check (√) in the open box next to "Non dichotomous Scores". You may ignore the other two options for Marking Scheme and Answer when constructed response items are involved.



4. Click the button "Step 1: Locate Examination". This will load an open file dialog box (following illustration).

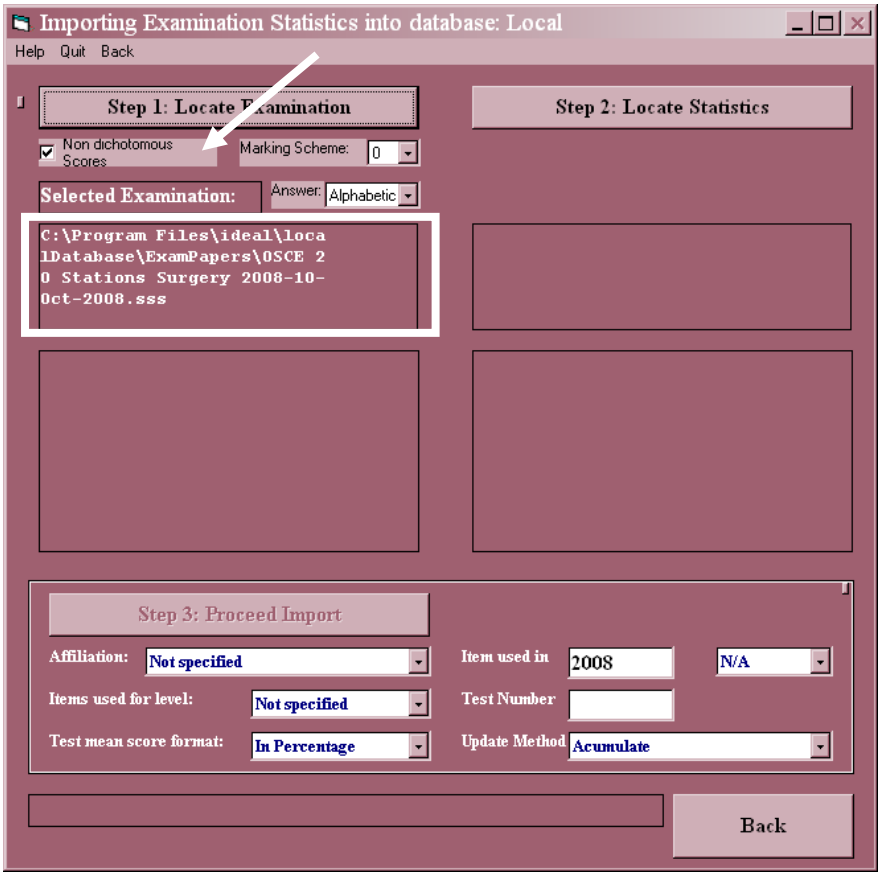


Locate and highlight the relevant file with the extension *.sss (that was created by IDEAL when you originally selected *and printed* the items in IDEAL PC).[4] The file name will include the course or module number for which the generated test was intended, along with the date when the items were printed. Click the button "Open".

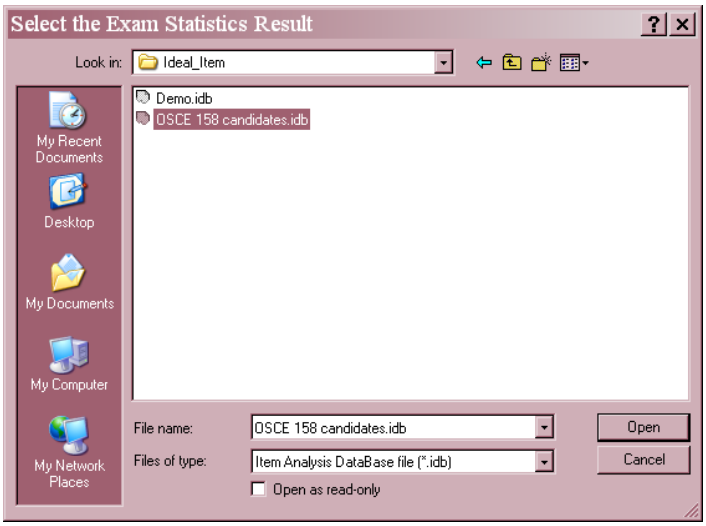
Note that IDEAL, by default, displays the above illustrated "ExamPapers" folder. If the relevant file was not saved in the ExamPapers folder, navigate to the appropriate directory.

[4] Recall that in the "Print Exam" window in IDEAL for all constructed response items you must check (√) the open box beside the label "All items will be scored nondichotomously".

The name of *.sss file that you have selected will now be displayed (as illustrated below).



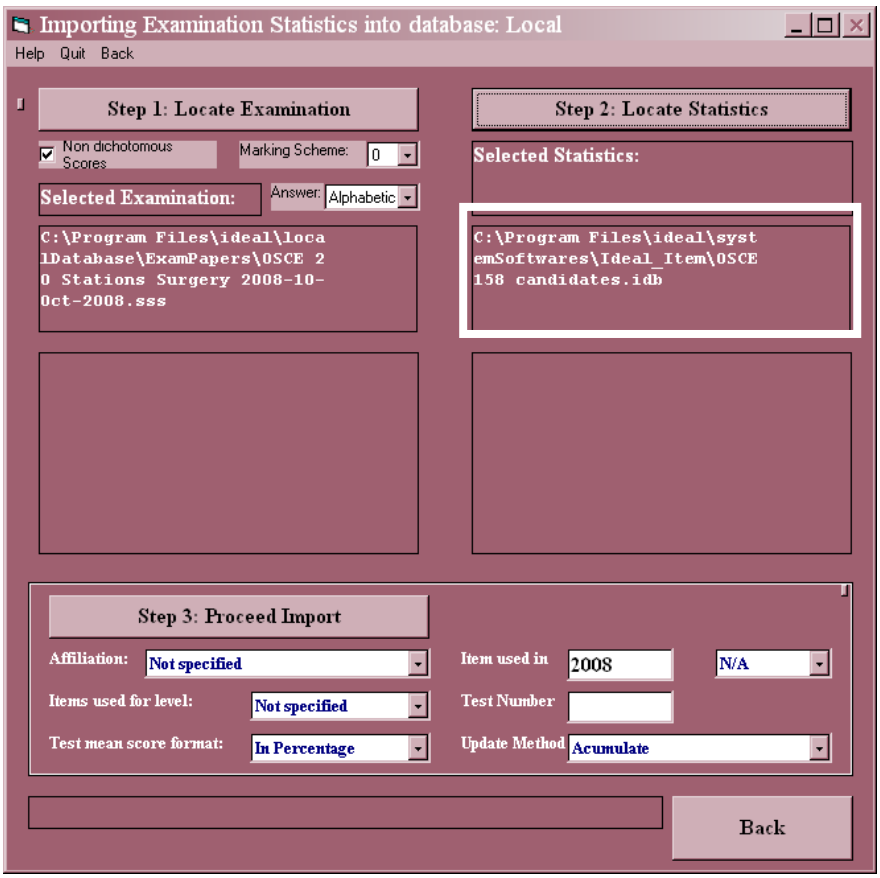
5. Next, click the "Step 2: Locate Statistics" button (on the right side of the window). This will load the following illustrated window.



Locate and select the relevant Softcopy output item statistics file (*.idb) generated by IA [presumably assigned with the same file name as that used for the *.sss file selected in step one (i.e., the name used for course description or module number when the items were printed by IDEAL)].

By default IDEAL will display the ExamPapers folder (as illustrated in the previous page. If you did not previously save the relevant *.idb file in this folder, navigate to the appropriate directory (for example, IA's default directory for saving Softcopy Output file is C:\Program Files\ideal\systemSoftwares\Ideal_Item\). Click "Open".

6. The Softcopy Output (*.idb) file you just selected will be displayed (as illustrated below).



7. Before clicking the Step 3 button, note the lower section in the above window. Always enter information related to the administered examination – i.e., your university, the year of the medical program in which the test was used, the chronological year in which the test was administered. Use the default "Percentage" for indicating how the test mean is reported. Normally use the default "Accumulate" in specifying the updating method; the alternate choice of "Update" is used only if you previously uploaded the same statistics but made some mistake --- i.e., use of "Update" will overwrite the previous errors.

   Click the button "Step 3: Proceed Import".

   IDEAL will determine if the two selected files match. If they do match in terms of their items, the import proceeds and this message appears:

   "Succeeded in importing the statistics for <total record number> records. Press OK to return to menu."

   If the uploading is successful, the statistics will be assigned to the relevant banked items. If the banked items already have one or more sets of stored psychometric properties, this latest set of item analyses will be displayed as the last used statistics. Earlier psychometrics will become the second last use, with the third and older sets archived (i.e., not displayed but retrievable if needed).

8. If the selected files do not match, the item statistics are not uploaded into the item bank. A message appears, ensuring that you will not inadvertently update the relevant banked items with psychometric properties that apply to other items.

3.7 Solving Problems in Uploading Item Statistics into IDEAL's Item Bank

If uploading the items' statistical properties was unsuccessful, determine the reasons and correct the error. Subsequently, rerun the steps for uploading items into the item bank as described in Section 3.6.

Common causes for failing to successfully upload the item analysis statistics for continuous response items are as follows.

1. One did not check the open box for "All items will be scored non dichotomously" at the time the constructed response items were selected *and printed* in IDEAL (Section 5.1 in IDEAL manual).

2. The user did not check the open box "Non dichotomous Scores" located under the Step 1 button (see illustration on previous page).

3. The items selected and printed were different than those that were item analysed.

To correct the problem noted in point 1,

run the IDEAL program and reselect the same items; print the items, being sure to check the box "All items will be score non dichotomously". Now rerun the steps for uploading items as described in Section 3.6.

To correct the problem for point 2,

repeat the steps for uploading items, being sure to check the box "Non dichotomous Scores" located under the Step 1 button.

To correct the problem noted in point 3,

there are two possible solutions to consider.

The first possible solution is to locate and use correctly matched files if you feel confident that they have been generated. You may have misidentified the relevant files for the *.sss and *.idb files. For example, if one did not use the same name for the printed items, for the Data Bank File and for the corresponding Soft Copy output statistic file, it is possible one has inadvertently specified the wrong file in Step One or Step Two during the uploading procedure.

The second possible solution is to confirm that the Softcopy item statistics file (*.idb) refers only to items that were selected and printed by IDEAL. If the administered test included items that were not stored in the item bank, it is possible that the item analysis was for all the items in the test (rather than only the items from the bank). If this has occurred, rerun IA, appropriately modify the first row in the file that contains the key and candidates' responses. That is, in row 1, (the key) enter blanks

into all the columns used for designating the items that do reside in the bank. Rerun the uploading procedure described in Section 3.6, using the latter generated *.idb file in Step Two.

There is an additional error to those listed in the previous page that can occur. This last error will not be detected by IDEAL only if uploading item statistics for constructed response items. Thus, if the following testing situation can occur in your school, take the necessary steps to ensure that the uploaded statistics are correctly matched to the banked items.

Assume the sequence of the printed items becomes modified when the test is administered (i.e., a different sequence is used in the actual test than the order the items were printed by IDEAL). In this situation, the *.sss and *.idb files will not be correctly matched even though the two files reference the same items. Notwithstanding the sequenced mismatch, IDEAL will upload the item statistics in the *.idb file if the items referenced in the *.sss file are the same. The uploaded item analyses will be incorrect for all items that were shifted.

To avoid this undetected problem occurring, do the following.

1. Always determine if the order of the printed items was changed when the selected items were administered in the actual examination. If their relative order is unchanged, there is no additional, compensatory steps required.

2. If the order of the items was modified in the examination, run IDEAL and reselect the same items; use IDEAL's feature for reordering the sequence of items and generate a test that matches the sequence used in the examination.

3. Print the re-ordered items, assigning an appropriate name for the course/module number and checking the box "All items used will be scored non dichotomously". The *.sss file generated by IDEAL will now match the *.idb file not only in terms of the same items, but also their sequence in the test.

4. Rerun the uploading item statistics procedure as described in Section 3.6.

Finally the user may notice there is a very small (quasi-hidden) button on the right hand side of the window previously illustrated. This button when checked will provide features for the user to permanently delete an item's statistics (including the item's statistics that have been archived). This button is used if one wants to eliminate any previous errors one might have made in entering item statistics. The feature should be used with care as once the statistics have been removed, the result is not reversible.

3.8 Psychometric Properties Defined

The Open Item Analysis program provided with IDEAL produces test and item statistics based on classical test theory. Output includes a frequency distribution and histogram of total test scores, item statistics such as difficulty level and biserial correlation, discrimination index, item responses by high, middle and low-scoring groups, item-difficulty plots of each item against the quintile-groupings of total test scores and a scatterplot of difficulty level by biserial correlation.

The following item (*vs* test) statistics produced by this item analyses program are stored with the respective items in the bank. Statistics kept in the institution's bank will be those derived from the item's use in one's own institution unless the item has been only used elsewhere (i.e., in another institution within the Consortium).

The head office of IDEAL will archive all institutions' respective item statistics so that any member may use the latter database for benchmarking purposes.

1. Difficulty Level:

   Average mark obtained by candidates on the item.

2. Discrimination Index:

   This index provides information about the item's ability to differentiate between high and low scoring examinees on the item.

   Discrimination Index is calculated as: $P_{high} - P_{low}$

   where P_{high} is the average mark (expressed as a proportion out of 100) on the item by examinees in the upper 27% of the total test score distribution and P_{low} is the parallel average mark by examinees in the lower 27% group.

3. Correlation Coefficient:

   Correlation between scores on the item and scores on the total test is the Pearson's product moment correlation coefficient. A positive correlation indicates that examinees who performed well on the item scored relatively higher on the whole examination. A negative value indicates that examinees who performed well on the item scored relatively lower on the whole examination (and *vice versa*).

3.9 Interpreting Item Analysis Report

The open item analysis program computes an array of test and item statistics (based on classical test theory) for constructed response item formats. Assessment committees will find these data useful aids for assigning grades, diagnosing the quality of an overall examination and/or its individual items, detecting possible limitations in marking schemes or instructional coverage, and/or a means for improving subsequent assessments.

The Consortium has agreed that these item statistics are appropriate psychometric properties to store with banked items in the International Database for Enhanced Assessments and Learning (IDEAL).

The format and sequence of the following explanatory material follows the sequence of a typical report produced by the open item analysis program for non-binary scored items.

3.9.1 Run Parameters

The initial output presents the user's specifications for the item analysis.

```
                        OSCE Final Year Three: April 2005

Run parameters:
     First Item Number=    1
     Number of Items  =   20
     Hi/Lo Percentage =  27.00%
     Difficulty Plots =    F
   DataBase Identifier= OSCE1-156
   Item scores will be combined into 5 groups
     based on equal percentages (subject to ties).


Read MAXIMUM VALUES and ITEM SCORES using format: (T4, 20F3.0)
```

Purpose: Making the effort to enter an explanatory title will facilitate later references to the item analysis report as this title is printed as the initial output. Also, it is particularly important to pay attention to the correctness of the listed performance groupings and the format statement which are printed as the last specifications in this leading section.

3.9.2 Maximum Item Values

The output prints in sequence the maximum score for each item. In the following example, a twenty station OSCE has a maximum score of 10 for each station.

```
MAXIMUM ITEM SCORE VALUES

10.  10.  10.  10.  10.  10.  10.  10.  10.  10.

10.  10.  10.  10.  10.  10.  10.  10.  10.  10.
```

Purpose: Continuously scored items frequently have different maximum values for each item. This listing enables the user to check that the correct maximum value was entered for each item.

3.9.3 Frequency Table of Students' Total Scores on the Examination

Frequencies (number of examinees) that obtained each specified score on the total test are printed, along with the corresponding Z-score, the percentile associated with the score and the cumulative percentage of examinees obtaining a score at or below the listed value.

The two cutting points that were used to divide the class into high, middle and low groups are indicated after the frequency table. Unless otherwise specified by the user, 27 % of the examinees are placed into each of the high and low groups with the remaining 46% in the middle. The 27% figure is an approximation and depends on the size of the group of candidates and the presence of tied scores at the cutting points.

FREQUENCY DISTRIBUTION AND HISTOGRAM OF SUM OF ITEM SCORES

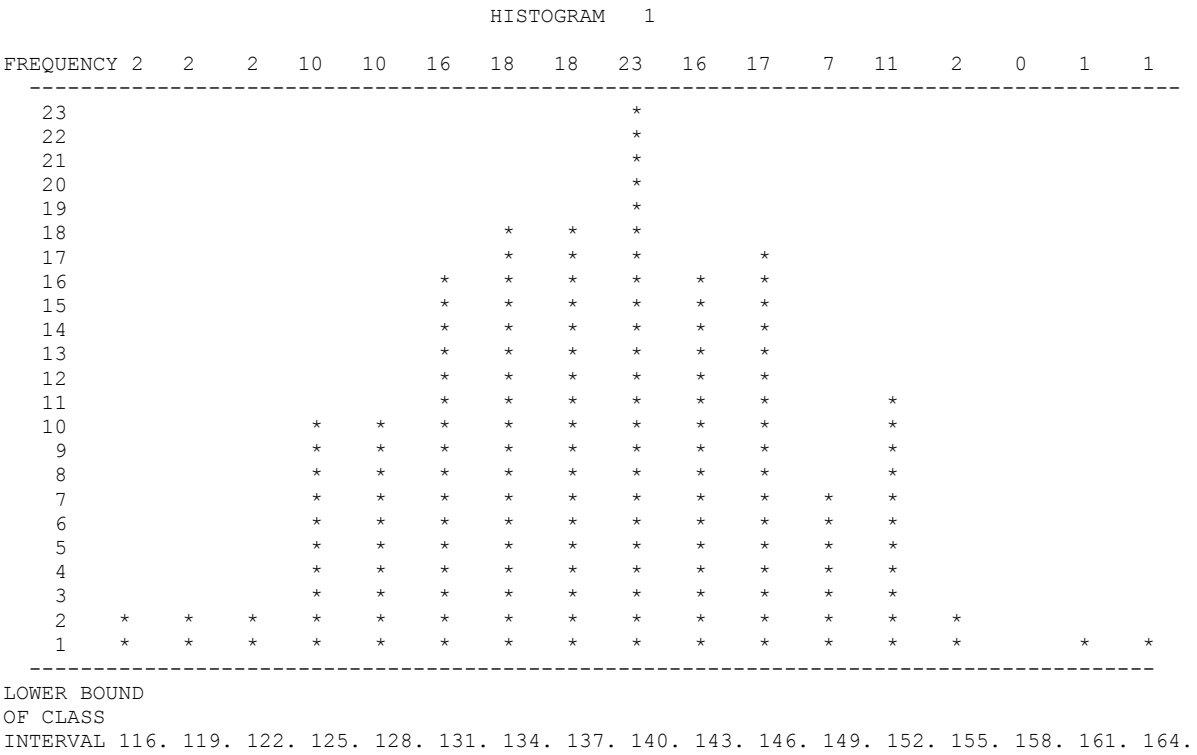
SCORE	FREQ	Z-NORML	%-ILE	CUM-%
117.	1	-2.726	0.32	0.64
118.	1	-2.341	0.96	1.28
120.	1	-2.144	1.60	1.92
121.	1	-2.006	2.24	2.56
122.	1	-1.898	2.88	3.21
124.	1	-1.809	3.53	3.85
125.	4	-1.633	5.13	6.41
126.	3	-1.449	7.37	8.33
127.	3	-1.323	9.29	10.26
128.	3	-1.215	11.22	12.18
129.	3	-1.120	13.14	14.10
130.	4	-1.020	15.38	16.67
.				
.				
.				
.				
154.	2	1.852	96.79	97.44
155.	2	2.070	98.08	98.72
161.	1	2.341	99.04	99.36
164.	1	2.726	99.68	100.00

The cutting points used to divide the class into HIGH, MIDDLE and LOW groups for the item statistics presented below are: 133.5 and 144.5

Purpose: Among the several uses that a table of this kind provides, one has an increased convenience for determining cutoff points for assigning grades, while another is the easy identification of the cumulative proportion of candidates that has met one's criterion reference standard.

3.9.4 Histogram of Candidates' Performances

A histogram representing the plotted student scores on the examination is provided (illustrated below).



Purpose: If the user finds the previous tabled frequencies too detailed, a graphic representation will normally help with visually estimating if individuals or groups are sufficiently different to merit different grades or if atypical cases merit exceptional consideration.

3.9.5 Overall Test Statistics

Following the histogram, the report specifies the overall test's mean, variance, standard deviation, reliability and standard error of measurement.

TEST STATISTICS

NUMBER OF EXAMINEES	156
TOTAL SCORE:	
MEAN	139.16
VARIANCE	76.62
STANDARD DEVIATION	8.75
ALPHA	0.5327
S.E. OF MEASUREMENT	5.9840

Purpose: These statistics for the overall test serve multiple purposes, not the least of which provides the user with data summarizing the test's overall difficulty (mean), variability across examinees (variance and standard deviation), inter item (station) consistency (alpha coefficient) and a helpful statistic for estimating which examinees' performances are sufficiently different to consider then as real differences (standard error of measurement).

3.9.6 Definitions

Immediately preceding the analysis of each item, a suitable reference source is cited for further explanations of each of the computed item statistical properties. This is followed by the definition for each abbreviated statistical term that is used throughout the remaining report.

```
G L O S S A R Y

DIF:   Difficulty: overall average performance (expressed in %)on the item
CORR:  Correlation between item and total test score
CR_R:  Corrected CORR between item & total score excluding present item
CON:   95% confidence interval for CR_R
IRI:   Item reliability index: CORR times the item's standard deviation
N:     No. of examinees in group
INV:     No. of examinees without valid score on item (i.e., negative or
       exceeding the maximum possible)
Est.:  Data for those assigned item's mean because of invalid scores
HIGH:  P73 and above
LOW:   P27 and below

TEST SCORE MEAN %:      Mean (in %) on total test for individuals achieving
                        the indicated score range on this item
DISCRIMINATING POWER:   The difference between the proportions of the HIGH
                        and LOW groups having this score
STANDARD ERROR OF D.P.: Standard error of discriminating power
```

Purpose: IDEAL does not store all possibly relevant item performance properties. Given the above, the user may find some additional properties that s/he also considers useful.

3.9.7 Item Analyses

The open item analysis program accommodates up to 20 performance groupings (each printed as a separate column for each item). One additional column also is printed for specifying the high, medium and low groups' performances for the overall item.

The following illustrated examples are taken from a statistical report for two stations (item 10 and item 13) in an OSCE that had 20 stations in total.

```
ITEM  10: DIF=0.688, CORR=  0.341, CR_R=  0.202 (95% CON=  0.046, 0.348), IRI=0.424
    GROUP   N   INV    range:       1       2       3       4       5  Overall
    TOTAL  156    0             0.00    0.01    0.15    0.60    0.24    0.69
     HIGH   45    0             0.00    0.00    0.02    0.60    0.38    0.75
     MID    69    0             0.00    0.01    0.19    0.55    0.25    0.68
     LOW    42    0             0.00    0.02    0.24    0.67    0.07    0.63
    TEST SCORE MEAN %:             0      68      70      72      74      70
    DISCRIMINATING POWER:       0.00   -0.02   -0.22   -0.07    0.31    0.12
    STANDARD ERROR OF D.P.:     0.00    0.02    0.07    0.10    0.09    0.10


ITEM  13: DIF=0.501, CORR=  0.286, CR_R=  0.090 (95% CON= -0.068, 0.244), IRI=0.493
    GROUP   N   INV    range:       1       2       3       4       5  Overall
    TOTAL  156    0             0.03    0.21    0.44    0.23    0.08    0.50
     HIGH   45    0             0.02    0.07    0.49    0.31    0.11    0.56
     MID    69    0             0.03    0.22    0.45    0.23    0.07    0.49
     LOW    42    0             0.05    0.36    0.38    0.14    0.07    0.45
    TEST SCORE MEAN %:            69      70      72      73      72      70
    DISCRIMINATING POWER:      -0.03   -0.29    0.11    0.17    0.04    0.11
    STANDARD ERROR OF D.P.:     0.04    0.09    0.11    0.09    0.06    0.11
```

The first line specifies the (i) difficulty level (average performance, expressed as a proportion out of 100) for all examinees in this station, (ii) uncorrected and corrected correlation between the candidates' marks in this station and in all 20 stations, (iii) 95%

confidence interval for the corrected correlated correlation, and, (iv) the item reliability index.

The second line provides column labels for the subsequent matrix of data. In this example, the user requested 5 equal percentage intervals (see Section 2.7.1). Note that in the above two examples the word "range" is specified, indicating five equal percentage ranges: 1 = (0-20%), 2 = (21-40%), 3 = (41-60%), 4 = (61-80%), 5 = (81-100%). The word "range" in line 2 would be replaced with "raw", "n_tile", or "Z-grp" if an alternate grouping criterion had been specified by the user. Similarly, the number of performance group columns printed could be increased (to 20) or reduced (to 2), depending on the user's specification. In all cases, one extra column is also provided specifying the overall and high to low group averages in the station.

In the next four rows of the item analysis matrix, the proportions of examinees in each of the 5 performance groupings on the station (item) are specified. In the first of these four rows, the total class performances are cited. For example, in station 10, out of a class of 156 candidates, none obtained a mark of 20% or less, 1% of the candidates received marks between 21 and 40%, 15% obtained marks of 41 to 60%, 60% obtained marks of 61 to 80%, and 24% of the candidates achieved marks of 81% or above. The overall average was 69% (i.e., the same as the specified difficulty level).

In a similar fashion, the proportions of the top performing students with marks in the 0-20%, 21-40%,…81% plus groups are listed in the next line. That is, 2%, 60% and 38% of the students in the top group on the total OSCE had marks in the ranges of 41-60%, 61-80%, and 81-100% respectively. The following two lines specify the corresponding data for the middle and low performing groupings. The next row lists the total test averages (expressed as proportions out of 100) achieved by those within each of the five performance groups on the item. The last two rows in the matrix provide the discrimination power and associated standard errors for each of the groups. The station's overall discrimination power is the mean difference between the top and bottom student groups (i.e., .75-.63=.12). That is, this OSCE station had a mean difference of 12% between the bottom and top 27^{th} percentile groups. These discrimination values are noted in the last column which specifies the overall corresponding psychometrics for the five performance groups on this item.

Purpose: These data help to diagnose if the station has achieved its intended assessment purpose, the adequacy of the marking scheme or markers and/or how well the skills assessed by this station were taught. For example, in station 10 the top group on the total OSCE is increasingly represented as the marks for this station increase. Conversely, the lowest groups are in greater numbers as marks for this station decrease. Given also that (i) the overall difficulty was adequate (.69), (ii) the station's marks are reasonably consistent with the total examination (CR_R=.202), and (iii) the station had contributed to variation in the OSCE (i.e., an item reliability index of .424), Station 10 apparently performed well in discriminating among candidates of varying clinical skills (at least as defined by the overall OSCE of 20 stations). Further, it would appear that clinical instructors taught this skill adequately, and students had sufficiently practiced the skill as most performed at a level that is often regarded as acceptable. However, if the assessed skill in this station was regarded as essential, the clinical instructor may well be hoping that the item was non-discriminatory with all students achieving at the highest level.

In Station 13, although the item has similar discrimination power as Stations 2 and 3, 10% of the top group, one-quarter of the average performers and over 40% of the bottom group were below what is usually regarded as a minimum standard. The test committee setting this OSCE is thereby informed that either the markers or marking scheme were too demanding or that the clinical instruction for this skill was not adequate (e.g., instruction limited to a demonstration and thus the skill was only observed by some candidates, some candidates had insufficient access to appropriate patients, inadequate time was scheduled for some candidates to practice the skill).

3.9.8 Difficulty Plots

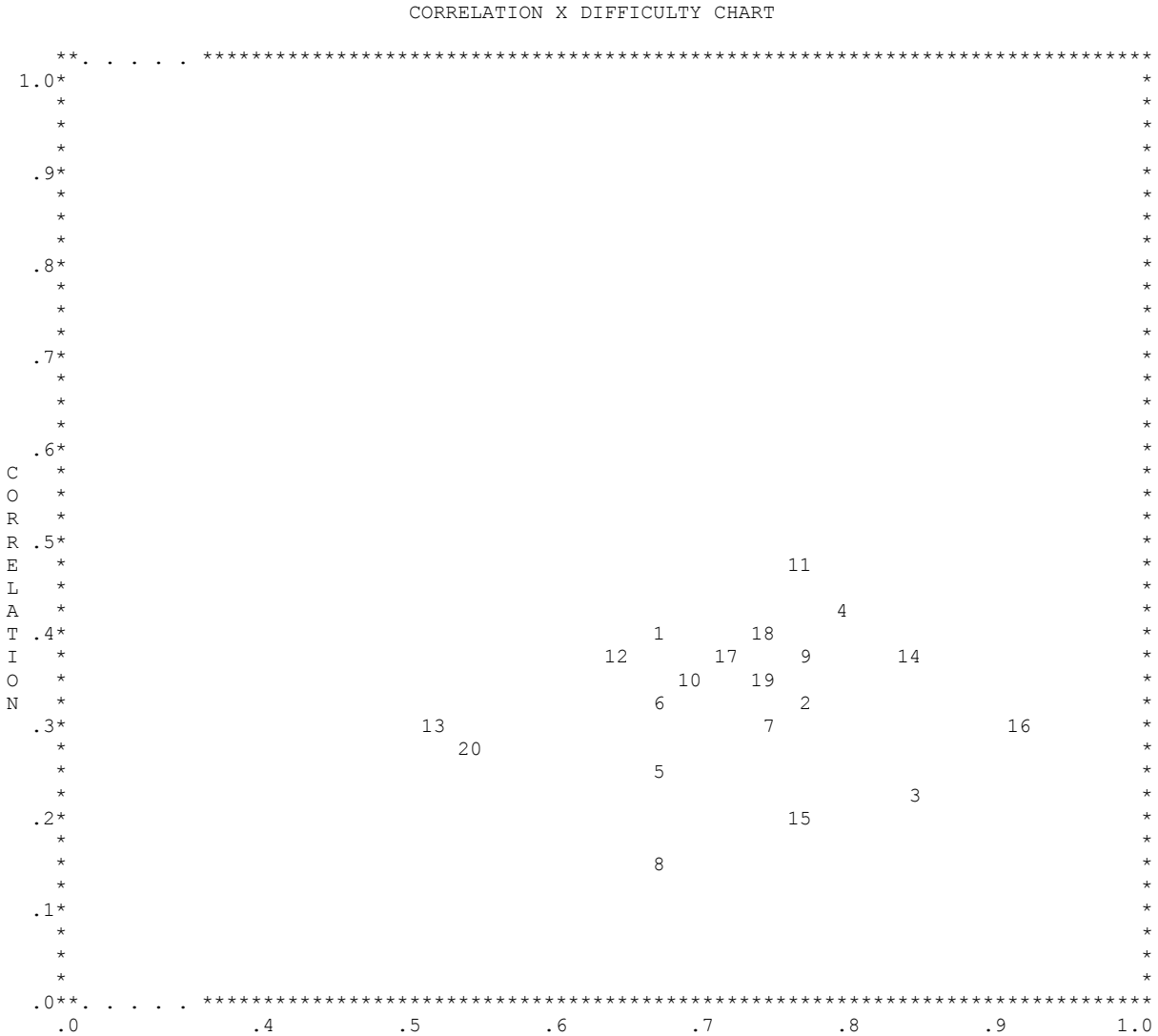
If the user requests difficulty plots, these will be included in the printout (see the following examples). Similar to the default used for generating the soft copy psychometric file (which is read by IDEAL), the performance groupings on the item are always quintiles (and used as the vertical axis). Difficulty level is always used as the horizontal axis.

```
PLOT OF TEST SCORE GROUPS (Quintiles) BY ITEM DIFFICULTY

The sizes of the quintile groups from 'Highest' to 'Lowest' are:

                                    ITEM 2
Quintile:
 Highest-                                          x

      4th-                                       x

      3rd-                                        x

      2nd-                                      x

  Lowest-                                     x
           *----*----*----*----*----*----*----*----*----*----*
      DIFF= .0   .1   .2   .3  .4   .5  .6   .7   .8   .9  1.0




                                    ITEM 3
Quintile:
 Highest-                                              x

      4th-                                          x

      3rd-                                        x

      2nd-                                      x

  Lowest-                       x
           *----*----*----*----*----*----*----*----*----*----*
      DIFF= .0   .1   .2   .3  .4   .5  .6   .7   .8   .9  1.0



                                    ITEM 13
Quintile:
 Highest-              x

      4th-      x

      3rd-       x

      2nd-            x

  Lowest-           x
           *----*----*----*----*----*----*----*----*----*----*
      DIFF= .0   .1   .2   .3  .4   .5  .6   .7   .8   .9  1.0
```

Purpose: Although one can deduce much the same information from the previously discussed item analyses matrices, these plots enable the user to visually and quickly determine if the item is performing as required. For example, in item 2, the difference in average performance of the top quintile on this item is only marginally better than the bottom quintile group, with all candidates performing in the range of 70 to 79%. In contrast, in item 3, the plot reveals that the item had much more discrimination power especially when contrasting the top and bottom fifths of the class (i.e., first quintile ($Q_{1)}$ = .43 and Q_5 = .9). Except for the bottom fifth of the candidates, the item was adequately answered or performed. Clearly, remediation needs to be considered for the bottom group, especially if the skill or knowledge being assessed relates to core material.

In the final example (item 13), it is likely the item (or station) has inherent problems and/or the instructional coverage for the assessed knowledge or skill is in question. Two of the top three groups perform more poorly than the bottom two groups. Closer examination of the item (station), the scoring scheme or markers, or the teaching of the assessed content/or skills may reveal why this unexpected distribution occurred.

Finally, if the user requested difficulty plots, an overall difficulty by correlation plot is also printed. In the following illustrated plot, the relationship of the items' difficulty levels are not strongly related to their consistency with the total test (OSCE).



3.9.9 Interpreting Results Based on Performance Groupings

IDEAL's open item analysis reports are non traditional as they are based on a continuum of the candidates' performances on a constructed response item, rather than the binary measured responses of candidates in selecting one choice from a given set of options within a selected response item.

In reference to the candidates' performances measured as a continuum for each item, the user has options how these will be represented when analyzed, i.e., in (i) raw scores, (ii) raw score ranges, (iii) percentage ranges, (iv) percentile ranges, and (v) Z-score ranges. For options 2 to 5, the range of measured performances can be grouped into equal or unequal intervals. By default 5 equal percentile ranges are the default (i.e., quintiles).

Open item analysis reports based on raw scores, raw score ranges and percent ranges are more easily understood as the marks are regarded as the absolute performances of the total class (as well as the top, middle and bottom candidates when defined by the total test mark or by a user supplied reference standard).

However, in using percentiles (or Z-scores), one may find that in real life assessments, the reported item analysis matrix may have more candidates being captured in the lower most percentile groups (than found with raw scores or percent groupings). The reason is that percentiles (and Z-scores) are determined using relative performance, not absolute performance, criteria.

In the example of an OSCE station with a maximum of 10, very few (or even no) candidates may be awarded a mark of only 0 to 2 (or the equivalent to being placed in the 0–20% range). However, it is possible that these same candidates with marks 0 to 2 would lie within the $P_{1\text{-}20}$ range (i.e., whenever 80% of the group performed better).

Whether or not the candidates with marks of 0 to 2 would also be captured in the lowest Z-score range will be determined by both the overall average performance and the standard deviation of marks that are awarded to candidates for the station. When within item variation is relatively small and an overall average is very good performance, it is more likely that the lowest Z-score category will include these candidates with very poor raw scores (or even those with marks of 3 to 4). Conversely, candidates with high scores (e.g., 8-10) will be included in the highest Z-score group only if the overall candidates' marks vary little and the assessed skill is very challenging (i.e., more difficult).

In deciding which of these various performance formats might be more appropriate to use in reporting item analyses for constructed response items, the following guidelines may be useful.

(1) Use of raw scores and percentages are appropriate if one can assume the candidates' assigned marks (e.g., for a MEQ, OSCE, or SAQ) are a fair, reliable, and an accurate representation of the candidates' knowledge or skills. These criteria are usually met when the marking scheme is well formulated, inter and intra marker reliability is good, and the assessment material has good content validity.

(2) Use of percentiles are more appropriate, if one only assumes that the assigned marks correctly order the candidates along the knowledge or skill continuum (i.e., (i) the marks do not necessarily represent the true amount of ability, and (ii) the apparent degree of differences among candidates is not necessarily accurate). These conditions can exist if different essay markers are used with some being lenient and others more demanding as they mark the essays, and where some markers assign a wide range of scores while others cluster their marks near to the passing grade. In a similar manner, if the OSCE station inadequately simulates real clinical practice, the candidates may be advantaged or disadvantage in terms of how they are able to demonstrate their true level of skill as rated by the observers.)

(3) Use of Z-scores is appropriate if one can assume the standardized distances between assigned marks for candidates is relatively acceptable (i.e., one can assume that not only the rank order of the candidates is correct but also that the indicated distances between candidates is also meaningful). Despite this however, one has less confidence in how well the marks represent the candidates' true level of ability. Assessment conditions such as this are illustrated by an OSCE station that uses a specific, easily followed marking scheme, and thus the markers are able to consistently apply the suggested marking scheme. Notwithstanding this latter advantage, the station as a simulation still imposes artificial restrictions on how well or poorly the candidates can represent their clinical skill as would be performed in an actual practice setting.

In summary, the first interpretive point to remember is the following. In using percentiles or Z-score ranges as performance groupings, the groups are defined using relative criteria (i.e., normatively), and are not based on the candidates' absolute marks (or the equivalent in the form of a percentage).

The second point to consider is that creative use of unequal ranges for raw score, percentage, percentile or Z-scores ranges can yield item analysis matrices which include candidates in almost all performance ranges. However, there is no theoretical reason that all performance levels should capture at least some candidates. This is particularly the case whenever criterion reference assessments are being conducted. The use of unequal performance ranges can be practically useful, however, such as matching one's chosen standard setting criterion, aiding decisions in relation to the assignment of grades, and helping to determine if exceptional candidates may require remediation or deserve special recognition.

Relevant References

Case, S M & Swanson, D B. *Constructing Written Test Questions For the Basic and Clinical Sciences*, 3rd Edition, Philadelphia: National Board of Medical Examiners, 2001 [available online at http://www.nbme.org/about/itemwriting2.asp].

Haladyna, T M. *Developing and Validating Multiple-Choice Test Items*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1994.

Jacobs, L C & Chase, C I. *Developing and Using Tests Effectively: A Guide for Faculty*, San Francisco: Jossey-Bass Inc., 1992.

Osterlind, S J. *Constructing Test Items*, 2nd Edition, Boston: Kluwer Academic Publishers, 1998.