July, 2003

SCheck.exe

MULTIPLE CHOICE ANSWERS - SIMILARITY CHECKING

George O. Wesolowsky - wesolows@mcmaster.ca

Most of this information is also in the SIMIL.HEL file which pops up on the screen in a NOTEPAD window.

Add-ons to this program, which will allow more convenient processing of large numbers of course files, are available.

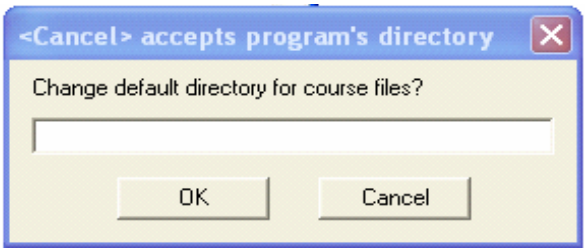
FILES USED BY THE EXCESSIVE SIMILARITY CHECKER (SCheck.exe)

- A .TXT file specifies correct answers and wrong answers without identifying the students.
- A .DAT file includes student numbers and names as well as the .TXT information.
- A .ANS file is the list of correct answers in a single column.
- A .UNM file is a file that is more easily constructed from other formats because it uses raw responses and an answer key without needing correct answers to be identified directly for each student.

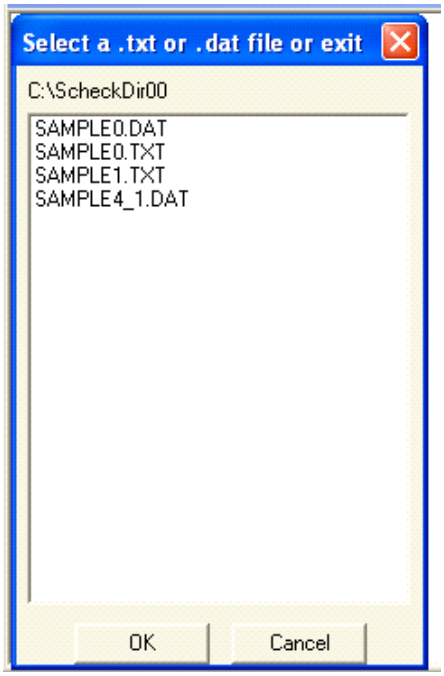
A converter, UNMtoDAT.EXE, is provided to create .DAT, and .ANS files from the .UNM format. Its use is optional if scantron outputs are converted directly into the required files by some other program.

The .TXT file is essential for running the similarity checker. The .DAT file allows identification of suspected pairs by name. If a .DAT file is present, a .TXT file will automatically be created . For full functioning of the program, the .DAT file is necessary. If a .ANS file is present, the answers will be listed in the .NAM output file and in an EXCEL file.

Running the Program

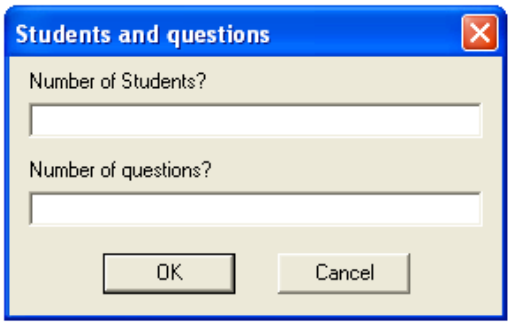


The first screen allows setting the program to a path other than the directory which contains the program. Example: c:\jnque. The Windows paste function could be used to enter a complicated path.
This option can be bypassed by clicking 'OK' or 'Cancel', or pressing <enter>.



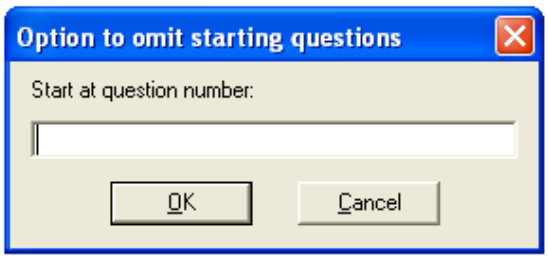
Select either a .TXT file or a .DAT file from the menu. If your file is in another directory you will get another chance. Select <Cancel> and then fill in the complete path and name of file when asked. Canceling again and entering any key will cause an exit.

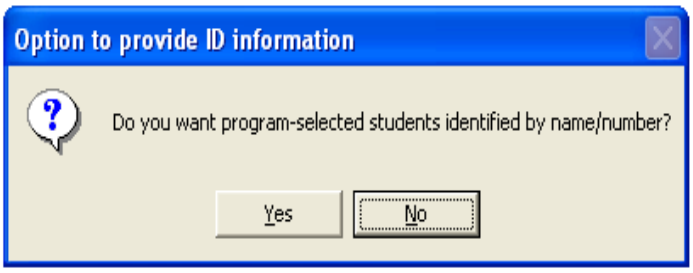
Example: c:\myfiles\mycourse
DO NOT include the .DAT or .TXT.



The program tells you the number of students and the number of questions that it found in the exam. You can change the number of students to a number less than given, but this is not generally a useful feature.

Changing the number of questions is quite useful, because you will subsequently be given the choice to start at a question other than one. This means that either the first or last question can be omitted if these are used for purposes such as version identification. SAMPLE0.DAT is a sample file where the first question was used for version identification and should be omitted. Also, if a question block is different (for example, the first 20 questions are true-false), these can be tested separately. Note, however, that this
method, unlike some others, can deal simultaneously with questions that are true-false, four-choice, five-choice, etc..

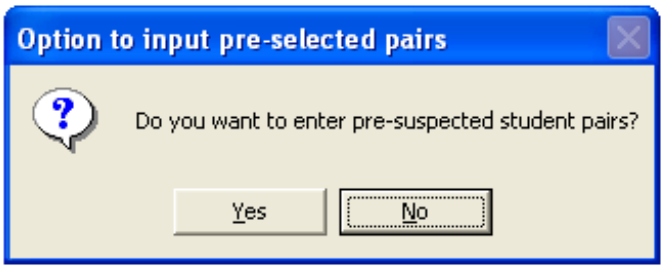




This program is designed so that files (.TXT) that do not identify students by name or number can be used. This could be useful if studies on classes are carried out by persons who do not need to know the identities of students. The .TXT files are subsets of the .DAT files, which also have student names and numbers. The program will run with either .DAT files of .TXT files, creating the latter if only the former is present.

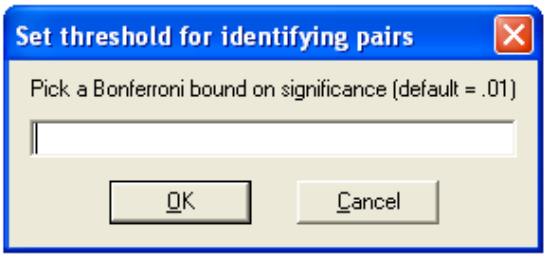
If the .DAT file is present, the program can create a detailed similarity report, with names of suspected pairs; this will be in a .NAM file. Also, there will be an offer to provide a complete list of names and marks, and a list of responses. These are optionally included in the .NAM file.

There is an option to create a tab-delimited file with an .XLS suffix. This file can be double-clicked to enter EXCEL directly. It contains the marks and responses for each student as well as some simple statistical summaries.



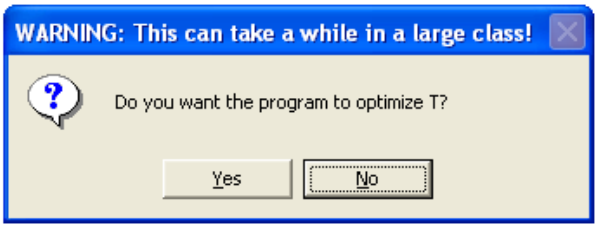
As is explained in the JAS paper by GOW, the standard of evidence for pairs suspected on other grounds (invigilator reports, for example) is at a lower level than pairs identified only by the 'data mining' of the program.

If the DAT file is present, then a menu will allow selection of student pairs by name and number. If only the TXT file is present, one can force the program to give a report on any pair by entering the sequential numbers for the pair. The sequential numbers are simply the positions on the list in the .DAT file of these individuals. One can use an editor with a line count feature to obtain these. Another method is to run the program once and ask for the complete .NAM output. Student sequential numbers are given in that output.

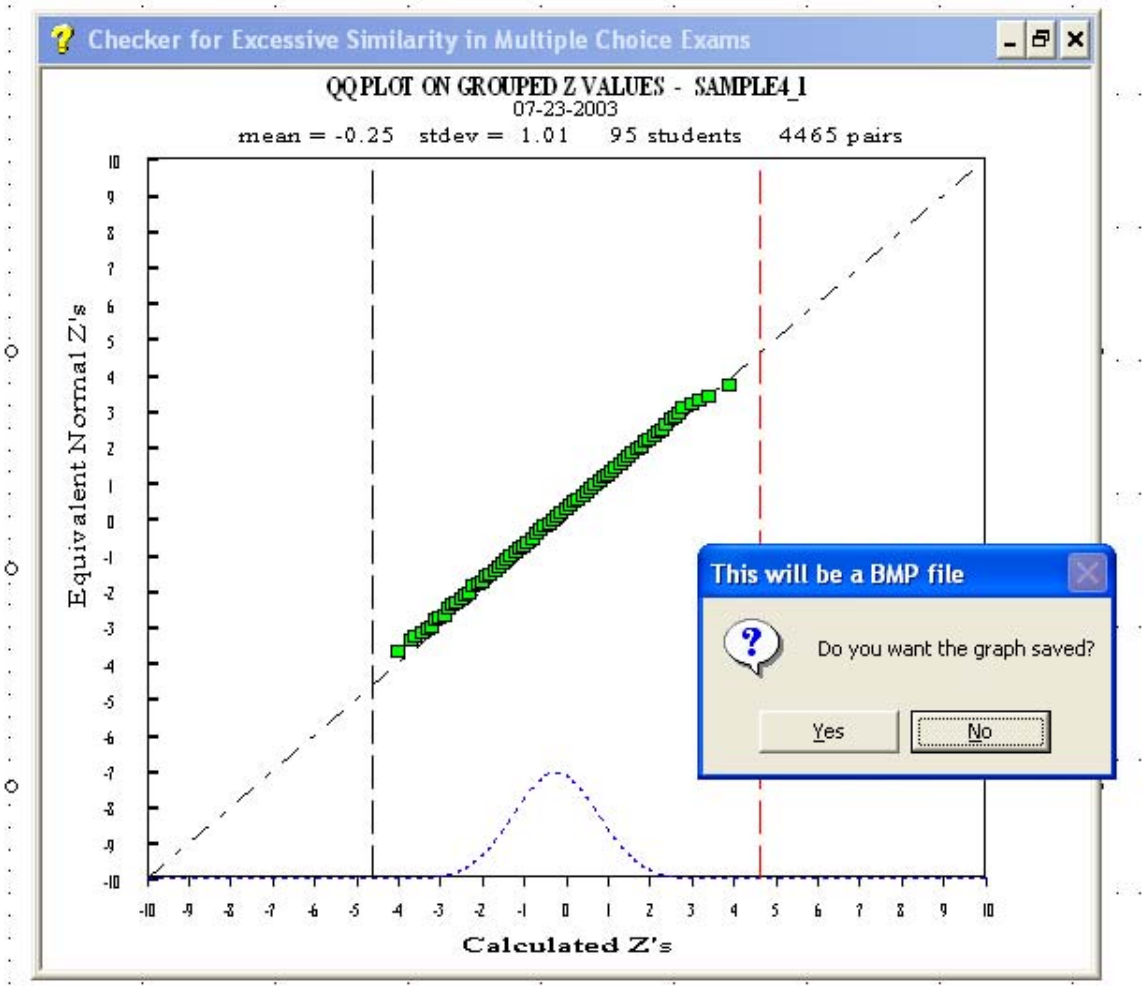


This topic is explained in the JAS paper by GOW. A default level of .1 means that fewer than 10% of classes will contain a falsely selected pair with that value of Zb or above. The actual level of significance is given for each pair and is usually much higher than the default level.

Setting a value of .1 for this bound means that, very approximately, one class in twenty will produce a false positive. Note, however, that classes with a high standard deviation of Zb (nearly 1) will have a false positive rate close to .1, while classes with a low standard deviation (say .7 or .8) will have a much smaller rate. Setting this parameter involves the usual balancing act between Type I and Type II error.



The T parameter fine-tunes the shape of the probability function describing the probability of a correct answer. It is set to .13 as a default. If you select 'yes' to optimize on T, this may consume a couple of minutes or more, depending on the size of the class and the number of questions. For more information open the Acrobat file T.pdf.



The diagram that appears on the screen after the program is run is an approximate Q-Q plot, and is a useful diagnostic. The diagram plots the equivalent normal z's against the calculated z's. If the line is approximately straight, this indicates normality. If the slope of the plotted points is steeper than the diagonal straight line then the standard deviation of the Z's is less than one. Note that the Z's are not independent. The vertical dotted line on the right is the cutoff. If the class is "clean" and the default cutoff is used, there should be a gap between the last of the points on the right and this vertical line. For small numbers of questions (<20)) the plotted points may not be a straight line because the normal approximation loses accuracy. Unusual patterns on the left side of the plotted points may indicate some students with many unanswered questions. The diagram will disappear if the mouse is clicked while the pointer is on this screen.

Note that the diagram is automatically written to the clip-board and hence is available for insertion into a word processing document. However, a mouse-click on the diagram will activate an option screen that will allow saving the diagram to a BMP file.

INTERPRETATION

The program will open output files .out and .nam (optional) and list them in NOTEPAD windows. These contain various forms of analysis.

SAMPLE:

```
_____________________________________________________________
** pair = 17  93 ** Harpp-Hogan stat = #wr.mat/#diff =   1.875
Zb =  5.032   'equivalent' z from the BVP model
 Significance of Zb on a pre-selected pair = 2.42E-07
Significance bound (Bonferroni)
                  on program selected pairs =  1.11E-02
 #matches = 42 | 50   (mu,s)=(  25.485   3.310)
prop. right for 17 = 0.640        prop. right for 93 = 0.560
 Quest. range = [ 1  50 ]       #students = 303
-------------------------------------------------------------
   STUDENT  17   6003317  AHAOR
.2........ 1.2..2.... ..1...1... .3.3.21544 .2..31.44.
-------------------------------------------------------------
   STUDENT  93   6445908  HITOP
.2...1..1. 1.2...2... ..1.5.24.. .3.3.21544 .2..31.54.
-------------------------------------------------------------
_____________________________________________________________
```

n = #students
m = #matches in answers

Zb is the standardized normal statistic equivalent derived from the number of matches, student performance, and question difficulty . It measures the degree of similarity between the answers of two students. Positive values mean above average similarity.

The Harpp-Hogan statistic is an empirically justified statistic. It is the ratio of exact wrong matches over the number of differences. Values > 1 are very suspicious. This statistic, However, is used in conjunction with another Harpp-Hogan statistic, called SIGMA. It is not reliable by itself. See the Harpp-Hogan papers for interpretation. It is presented here only as "a second opinion".

Significance = Prob(number of matches is >= m) = Prob(Z >= Zb)= probability that a pre-selected pair will be falsely accused if the Zb observed were to be used as a cutoff. This is the relevant significance if there is some prior reason for suspecting the pair. In this version of the program, this probability is calculated using probabilities for 'Bernoulli trials with varying success probabilities' (BVP), and not the normal approximation.

Bonferroni bound = upper bound on the probability that a class will have a falsely accused pair(s) if Zb is used as a cutoff. This is relevant if there was no prior reason to

suspect the pair. This bound is from the Bonferroni inequality. It is also calculated using BVP probabilities.

FILE FORMAT

Example of a .DAT file:

```
9706600 ,               ,  ,3...1.1...2.........5..1..
9799221 ,               ,  ,3222....222.2222..224..33..
9735555 ,AARDVARK       ,SD,3.2.1.112..............4.2
9719999 ,AHURA-MAZDA    ,S ,12....1....12..211.2......5
9707777 ,APOLONIUS      ,D ,3222..112.2.2.22...2..1.21.
9717777 ,ASMODEUS       ,Z ,1.2.1.11..2..2.211.2...331.
```

Missing names and initials are represented by spaces, field length does not matter. Correct answers are dots. The number is the position of the incorrect response. Duplicate answers is coded '*'. No answer is coded '-'

Example of corresponding .TXT file:

```
3...1.1...2.........5..1..
3222....222.2222..224..33..
3.2.1.112..............4.2
12....1....12..211.2......5
3222..112.2.2.22...2..1.21.
1.2.1.11..2..2.211.2...331.
```

The program will give an error message if it finds violations of the above format.

An .ANS file simply contains a column of answers.

Example of an .UNM file:

```
answers ,               ,  ,32141243541234523124212312 3
7129221 ,KRULL          ,  ,322224512223-22234224343314
9735555 ,BORAGERDD      ,SG,32345*423112322321233212232
9456779 ,SAURON       ,  ,12342144222123321233444455 5
9707777 ,SHIVAGO        ,D ,32221232332332123244345345 5
6788888 ,KREUSUS     ,O ,12342434332434343424244 41
```

Note: none of the above examples from real classes, but they will "run" on the software (for illustration).

Limitations: the number of students must be less than 3000, the number of questions must be 100 or less.

This program is based on the method described in:
George O. Wesolowsky, 'Detecting Excessive Similarity in Answers on Multiple Choice Exams', Journal of Applied Statistics, Vol. 27, No. 7, 2000, pp. 909-921.