

SCheck Version 8

Statistical detection of cheating (copying, collusion) in multiple choice tests or exams

or

Checking for 'Excessive Similarity' in multiple choice answers

George O. Wesolowsky - wesolows@mcmaster.ca

New features in Version 5:

- Responses can be on the range a-z, or 1-5.
- The permissible number of questions and number of students were enlarged.
- An adjustment for non-responses was introduced.
- The complete listing of similarity scores is now an option.

New features in Version 7:

- Less conservative but more accurate cutoffs for program selected similarity scores are used. The original cutoffs are still calculated.
- Because of the need for pre-calculations, the program runs about 50% slower.

New features in Version 8:

- Using approximate significance (as in version 7) or the Bonferroni bound on significance (as in pre-7 versions) is now a choice.
- An Excel readable file containing a summary of information on all detected pairs, for single or batch runs can now be produced. This is useful if a large number of tests is being processed.

Contents

1. Files used by SCheck.exe	page 2
2. Running the program interactively	page 3
3. Running the program from BATCHS.INI	page 7
4. Running multiple files simultaneously	page 8
5. Listing similarities for all pairs	page 8
6. Information in output files	page 9
7. Interpretation	page 10
8. A special adjustment for non-responses	page 12
Appendix: File Formats	page 12

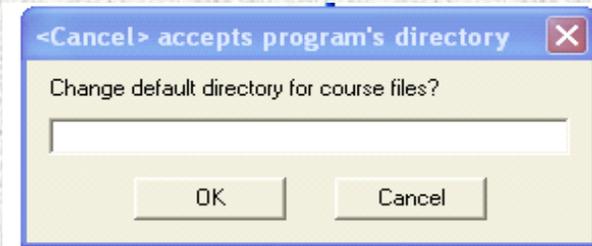
1. FILES USED BY SCheck.EXE [To page 1](#)

- The .TXT file specifies, for each student, which answers were correct and which particular wrong answers were selected. Students are identified only by sequential number.
- The .DAT file includes student ID numbers and names as well as the .TXT information.
- The .ANS file contains the list of correct answers in a single column or a comma separated list.
- The .UNM file is a file that is easily constructed from other formats because it uses raw responses and an answer key. It does not need correct answers to be identified directly for each student as does the .DAT file. A converter, UNMtoDAT.EXE, is provided to create .DAT and .ANS files from the .UNM format.

The .TXT file is essential for running SCheck. The optional .DAT file allows identification of suspected pairs by name. If a .DAT file is present, a corresponding .TXT file will automatically be created, *but only if it does not already exist*. For full functioning of the program, the .DAT file is necessary. If an ANS file is present, the correct answers will be listed in the .NAM output file and in an optional EXCEL file.

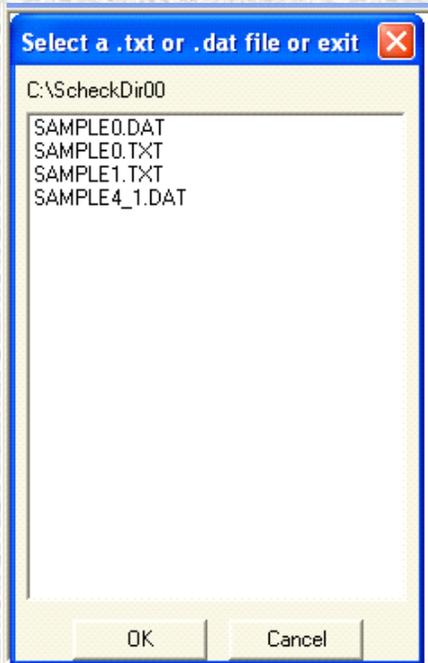
WARNING: *If only one of the .TXT and .DAT files is modified, the mismatch could result in incorrect output. If the .DAT file is modified, the old .TXT file must be deleted so that a matching one can be created by SCheck during the run.*

2. RUNNING THE PROGRAM INTERACTIVELY [To page 1](#)



The first screen allows setting the data path to other than the directory which contains the program. Example: `c:\jnque`. The Windows paste function could be used to enter a complicated path. Note the optional use of `“*”` as described in section 5.

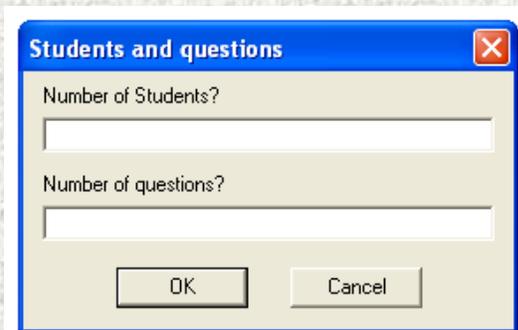
This option can be bypassed by clicking 'OK' or 'Cancel', or pressing `<enter>`.



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 the file when asked. Canceling again and entering any key will cause an exit.

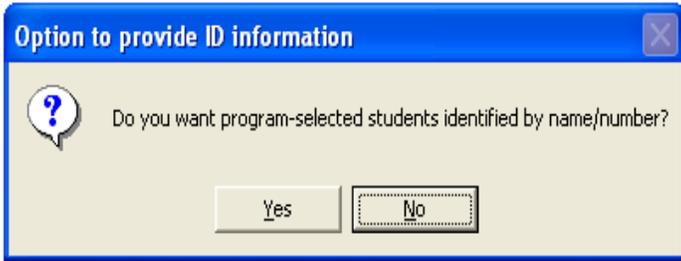
Example: `c:\myfiles\mycourse`
DO NOT include the .DAT or .TXT.



The program console screen 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.

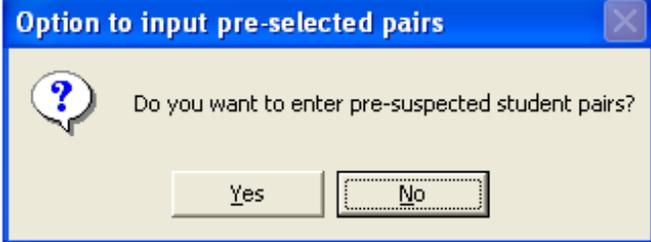
Changing the number of questions is quite useful, because you will subsequently be given the choice to start at a question other than one. This means that either the first or last question can be omitted if these are used for purposes such as version identification. `SAMPLE0.DAT` is a sample file where 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 `SCheck` unlike some other programs, can deal simultaneously with questions that are true-false, four-choice, five-choice, etc..





SCheck 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 or .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 a .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 Journal of Applied Statistics paper by Wesolowsky, the standard of evidence for pairs suspected on other grounds (proctor reports, for example) is at a lower level than pairs identified only by the 'data mining' of the program.

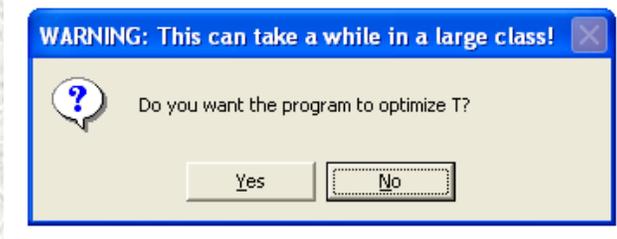
If the .DAT file is present, then a menu will allow selection of student pairs by name and number. If only the .TXT file is present, one can force the program to give a report on any pair by entering the sequential numbers for the pair. The sequential numbers are simply the positions on the list in the .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.



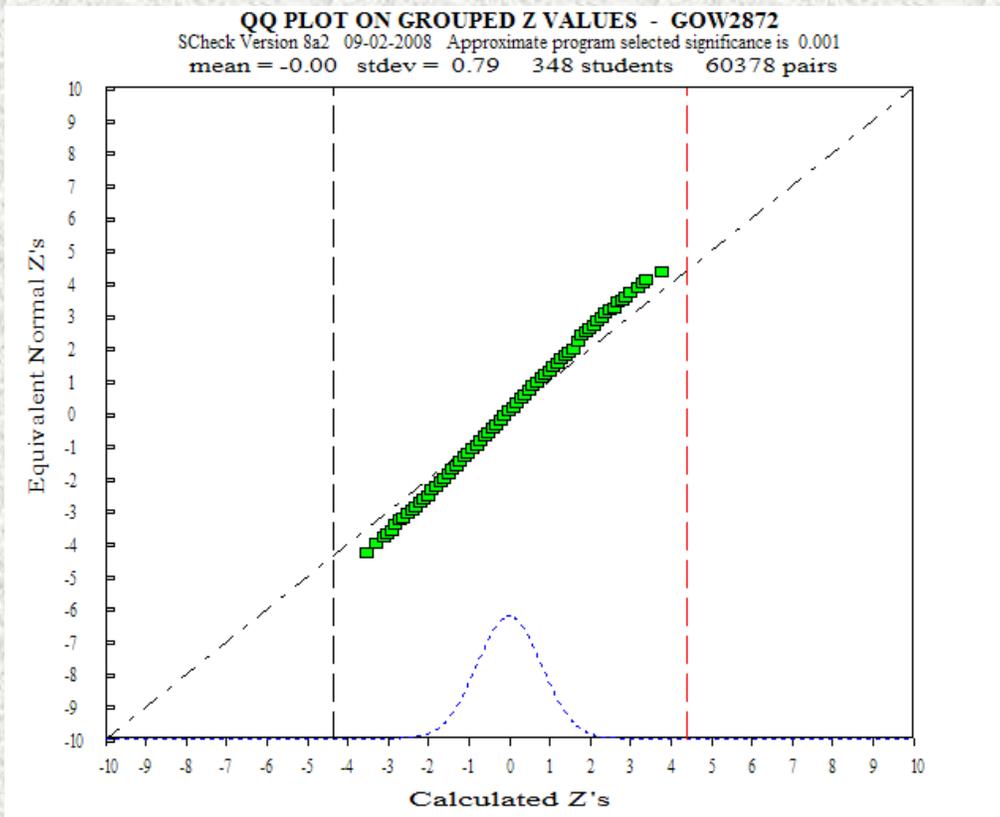
The default level of .001 means that it is estimated that .1% of completely 'innocent' classes will contain one or more false positive pairs. Setting this parameter involves the usual balancing act between Type I and Type II error.

This calculation differs from that in the original paper. This change was made so that tests with a high standard deviation of the Z_b scores in that particular class (say, nearly 1) will have a comparable false positive rate with tests with a low standard deviation (say .7 or .8). Using the Bonferroni bound treated such tests inconsistently. However, more approximations have been introduced. A normal approximation, as well as another estimated parameter are incorporated into the significance estimate.

However, entering .001B, for example, gives a Bonferroni bound significance according to the original paper.



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 dashed line** on the right is the estimated cutoff, at the specified level, for program selected pairs. 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.

The diagram above is typical of a class with no detectable cheating. For small numbers of questions (<20) the plotted points may not be a straight line because the normal approximation loses accuracy. Note, however, that for identified pairs the probabilities are calculated from the cumulative binomial distribution and not from the normal approximation. There was also a minor change in Version 7 in the method of calculating the normal approximation. Unusual patterns on the left side of the plotted points may indicate some students with many unanswered questions (the OUT file gives the non-response rate for every student). 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.

3. RUNNING THE PROGRAM FROM BATCHS.INI [To page 1](#)

If BATCHS.INI is present in the same directory as SCheck.EXE, and if the former is activated by answering the first question with a 'y', then the introductory screens of SCheck.EXE will be omitted and all parameters will be read from BATCHS.INI. Pre-requisite choices are indicated in the wording of the questions. Note that if the graph is chosen to be displayed it will pause the program until a mouse click on the screen is executed.

BATCHS.INI is a text file that can be accessed with WORDPAD or NOTEPAD. It permits a variety of choices for parameters and output. An example is:

```
Activate the batch mode?
n
Display output files?
Y
Display graph?
n
Create a BMP file for the graph?
Y
Number of students (override file count) ?
-1
Number of questions (override file count)?
-1
Start at question:
1
Identify suspects by name?
Y
Identify all students by name (must identify suspects)?
n
Produce Excel file (must identify all students)?
n
Enter pre-selected pairs?
n
Set program selected significance; append B for Bonferroni bound
.001
Use T optimization?
n
Specify path to course files (eg. c:\courses\);append ^ for summary
file
g:\myfiles^
NRT: Matches on "-" with a proportion > NRT won't count
.1
```

This file MUST be a plain text file with no extraneous symbols, and no leading blank lines. The format of this file must be preserved (as above) but the choices can be altered. Use lower case 'y' or 'n'. If the directory specified does not exist, the same directory as the one that contains the program will be used.

4. RUNNING MULTIPLE FILES SIMULTANEOUSLY [To page 1](#)

The SCheck.EXE program was designed to be used by individual instructors on their own courses, and not for large scale processing. Nevertheless, it now has capabilities for such use if BATCHS.INI, and a program, SelectCh.EXE, are present.

SelectCh.EXE allows simultaneous menu selection of course files to be run sequentially. It simply calls SCheck.EXE once for each choice made in the menu. Files are selected in the usual way, by <Ctrl> clicking choices in a window or by using the <Shift> key and dragging. Output is controlled by instructions in BATCHS.INI. Selection is made from the root names of either .DAT files or .TXT files. The program is not elegant, and runs accompanied by flashing screens. This means that it monopolizes the screen while it runs. Part of the reason for this inelegance lies in technical limitations in using consoles in WINDOWS (this is a console application) and part because the program was not re-designed, just adapted for batch use. NOTE that this program requires that BATCHS.INI be present and activated. Processing is much faster if T is not optimized.

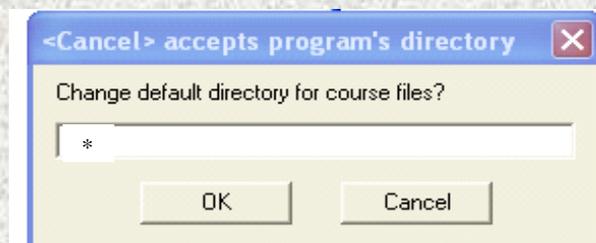
ChooseUNM.EXE is a converter utility that may be used in conjunction with UNMtoDAT.EXE. It permits many .UNM files at a time to be converted into .DAT and .ANS files.

Also, see below for how to produce an Excel summary of the information in multiple .NAM files.

5. SPECIAL FEATURES (*,^) [To page 1](#)

It is useful for some applications to list the similarities for all pairs. Note that this could result in a very large file because there are $n(n-1)/2$ pairs, where n is the number of students.

If the program is run in the interactive mode, then placing an asterisk (*) anywhere in the default directory slot (whether or not a directory is specified) will activate the listing of the similarities for each pair.



If the program is run with BATCHS.INI, then the asterisk is appended to the directory name.

When running multiple tests, using the symbol ^ in the same way as above will produce a summary of the information in NAM files. This information will be placed in an Excel-readable file.

For example:

Specify path to course files (eg. c:\courses)
c:\jnque* or c:\jnque^

Note that this will work even if the directory specified does not exist, and hence the program default directory is used.

6. INFORMATION IN OUTPUT FILES [To page 1](#)

The program will write a .OUT file which contains various forms of analysis. It will also optionally produce .NAM and .XLS files.

The .OUT file does not use pagination, but only simple text output. This makes it convenient to cut and paste the output into a program such as MINITAB for further analysis. This file presents all the components estimated in the Scheck model.

If suspicious pairs are detected, a detailed analysis includes pre-selected pair significance, program selected pair significance, expected Zb score for each pair in the absence of collusion, the Harpp-Hogan statistic, and estimated probabilities of a match on each question in the absence of collusion.

A table gives the proportion of correct responses for each question, as well as the distribution of wrong answers. This is useful in detecting if the answer key had an incorrect answer; one should investigate if a low proportion of correct answers is accompanied by a substantial majority choosing a particular wrong answer.

The proportion right for each student is given along with the model's prediction for this proportion (which must be identical). The model's performance in predicting the proportion right on each question is also detailed.

Statistics include point biserial correlations, Cronbach's alpha, means, medians and quartiles. The frequency distributions and calculations necessary for plotting the graph are also given.

The .NAM file is optional and identifies students by name if this information was present in the .DAT file. The .XLS file provides a mark report that is directly readable by Excel.

If a complete listing of similarities is selected by the asterisk (*) method, then a Wordpad .WRI file is created. This file will contain Z and Zb scores, as well as cumulative probabilities for each pair.

7. INTERPRETATION [To page 1](#)

The program will open output files of type .OUT and .NAM (optional) and display them in NOTEPAD windows.

SAMPLE:

```

** pair = 9 12 ** Harpp-Hogan stat = #wr.mat/#diff = 2.25
#####
Zb = 5.320 'equivalent' z from the BVP model
Significance of Zb on a pre-selected pair = 5.2E-8
Approximate significance of program selected pair = 1.1E-7
Signif. bound (Bonferroni) on program selected pairs = 1.3E-4
#matches = 26 | 34 (mu,s)=( 10.697, 2.653)
prop. right for 9 = 0.324 prop. right for 12 = 0.324
Quest. range = [ 1 34 ] NRT = 1.00 #students = 72
-----
..a.ecaeea dcace.dcab ..cdd..b.b .dc.
-----
.a.ecaeea dcace.dcab ...cc.eb.b ...c
-----
estimated match probabilities:
0.318 0.366 0.254 0.288 0.360 0.350 0.326 0.245 0.252 0.255
0.225 0.288 0.317 0.307 0.251 0.343 0.257 0.273 0.248 0.604
0.639 0.330 0.346 0.241 0.403 0.441 0.262 0.236 0.370 0.250
0.230 0.285 0.293 0.245

```

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 suspicious. This statistic, however, is used in conjunction with another Harpp-Hogan statistic, called SIGMA. It is not reliable by itself. See the references for interpretation. It is presented here only as "a partial second opinion".

The probability that this pair, if innocent, will have a match on the first question is .318, on the second question .366, and so on. These are calculated according to Equation 5 in

the paper. The probability of the total number of matches on the 34 questions follows a compound binomial (BVP¹) and not an ordinary binomial distribution because the probability of each event (match on a question) is different. Incidentally, the sum of these probabilities is 10.697, or *mu*, the expected number of matches. The estimated standard deviation $s = 2.653$ of the number of matches is the square root of the sum (.318(1-.318) + .366(1-.366) +...). This *mu* and *s* are from Equations 7 and 8.

The significance of Zb on a pre-selected pair is given as 5.2 E-8. This means that the probability of 26 or more matches out of 34 (or a Zb greater than 5.32) is 5.2/10⁸ if the pair was 'innocent' but singled out for some reason or suspicion before the similarity check was run. This probability is calculated using the compound binomial recursive method in the Appendix of the article. The Zb is a standardized normal deviate chosen to give the same right-tail probability; it is merely a convenience.

The significance bound (Bonferroni) on program selected pairs is 1.3 E-4. This comes from $5.2 \times 10^{-8} \times 72 \times (72 - 1) / 2 = 1.3 \times 10^{-4}$ because the class size is 72. Please see the first three paragraphs on page 916 of the article for the justification. The interpretation is as follows. If no students were pre-selected and students were suspected entirely on the basis of Scheck, and if Zb = 5.32 were the cutoff, the upper bound on the probability that the class would yield at least one false positive would be 1.3E-4. To put it another way, false positives would occur in fewer than about 1.3 in 10000 classes (on the average).

Using the bound on the significance is unduly conservative if the empirical distribution of the indexes for the class actually has a distribution narrower than a standard deviation of 1. This is caused by inter-correlation of similarity scores. Version 7 addresses this problem. Using a normal distribution with a standard deviation of .82 (see below) and not 1, the program estimates the actual significance for program selected pairs. For this pair this probability is 1.1E-7.

Scheck produces various summaries. If any pairs are identified, they are tabled in the .OUT file. In the example below, 6 pairs were identified at the .01 significance cutoff. Approximate significances are given for each pair. The '*' symbol means that the program selected significance was set to the a priori significance.

Summary of Significances of Identified Pairs

pair	Z	A Priori Signif.	Bonferroni Signif.	Approx. Prog. Selected Signif.
9, 10	6.384	8.6E-11	2.2E-7	8.6E-11*
9, 12	5.320	5.2E-8	1.3E-4	1.1E-7
10, 12	6.384	8.6E-11	2.2E-7	8.6E-11*
18, 37	6.942	1.9E-12	4.9E-9	1.9E-12*
35, 68	4.530	2.9E-6	7.5E-3	4.2E-5
59, 71	3.893	5.0E-5	1.3E-1	2.6E-3

¹ BVP: Binomial with varying probabilities.

8. A SPECIAL ADJUSTMENT: NRT [To page 1](#)

The model used by SCheck assumes that non-responses to questions are simply one of the wrong choices. If there is adequate time for the test and if there is no penalty for wrong answers, non-responses should be extremely rare, and this assumption does not affect the analysis. Unfortunately there are instances where this assumption breaks down. For example, there are some tests where many students miss large blocks of questions (usually at the end of the test), simply because of lack of time. Matches on such non-responses should then obviously not cause suspicion of cheating.

The NRT (non-response threshold) is a not entirely successful attempt to deal with this problem. If there is some question with more than 10% of non-responses, a window is activated which gives the option of entering the NRT. If the NRT is .2, for example, then if there is a non-response match by a pair on a question with more than 20% non-responses, this match is not counted, and its probabilities are not calculated. The NRT can be set between 0 and 1. When NRT=1 this option does not alter the standard analysis and when NRT=0, no matches on non-responses are counted. NRT can also be activated from the batchs.ini file.

The best solution to this problem is to avoid tests which cause it. There is as yet little evidence pertaining to the reliability of this adjustment.

APPENDIX: FILE FORMATS [To page 1](#)

Example of a .DAT file with responses in the 1-5 range:

```
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.
```

The first field is the ID (up to seven characters), the second is the name, the third can be initials, and the fourth is the block of responses.

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 , , , 321412435412345231242123123
7129221 ,KRULL , , 322224512223-22234224343314
9735555 ,BORAGERDD ,SG,32345*423112322321233212232
9456779 ,SAURON , ,123421442221233212334444555
9707777 ,SHIVAGO ,D ,322212323323321232443453455
6788888 ,KREUSUS ,O ,123424343324343434342424441
```

None of the above examples are from real classes, but they will "run" on the software (for illustration). The number of questions must be 400 or less.

The number of students in a single run of the program can be up to 100,000 or more. The program then becomes quite slow and processing this many examinees at once is not desirable for other reasons as well. If a large scale examination is administered at many centers, it is much faster and preferable, to analyze logical 'districts' separately. Setting the level of aggregation (number of separate writing locations that are combined) is a somewhat complicated but important issue. For example, cut-off thresholds should be adjusted according to the level of aggregation.

Note: Versions 4 and above of Scheck permit *abcde* choices in the input files **instead** of 12345, and also permit R-type responses on the range of *a-z*. Upper case letters are converted to lower case automatically. Do not mix letters and numbers in the same file. If both a 1 and an *a* are present in the same file as choices, they represent **different** choices. In other words, a different character means a different choice. The exception is that upper case characters are equivalent to their lower case characters.

An example of a permissible format is:

```
answers , , , DGAFAFAGFDTFAGDFTGDAGFGAGDAGD
154677 ,KRULL , , DGGGGFTAGGGD-GGGDFGGFDFDDAF
328999 ,BORAGERDD ,SG,DGDFT*FGDAAGDGGDGAGDDGAGGDG
345666 ,SAURON , , AGDFGAFFGGGAGDDGAGDDFFFFTTT
989936 ,SHIVAGO ,D.G. ,DGGGAGDGGDDGAGDGFFDFTDFTT
A67888 ,KREUSUS ,O ,AGDFGFDFDDGFDFDFDFDFGFGFFFA
```

which is equivalent to:

```
answers , , , dgafagfdtfagdftgdagfgagdagd
000154677,KRULL , , dggggftagggd-gggdfggfddaf
000328999,BORAGERDD,SG,dgdft*fgdaagdggdgagddgaggdg
000345666,SAURON , , agdfgaffgggagddgagddffffttt
000989936,SHIVAGO,D.G. ,dgggagdgddgddgagdgffdfdtftt
000A67888,KREUSUS,O ,agdfgfdiddgfdfdfdfdgfgfffa
```

Either one will be converted by UNMtoDAT.exe to the .DAT file:

```
154677,KRULL,,...gggftagggd-gg..f.gfdf.daf  
328999,BORAGERDD,SG,..d.t*.gda...ggdg..ddgaggdg  
345666,SAURON,,a.d.ga.fgggagdd.agddffffttt  
989936,SHIVAGO,D.G.,..gg..dgddgd.ga..gf.dft.ftt  
A67888,KREUSUS,O,a.d.gfdfddgf..df.fd..f.ffff
```

Leading and trailing blanks will be trimmed by the program. However, a .DAT file with capital letters and blanks around the names is acceptable.

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.