USING LERTAP 5.6 TO MONITOR CHEATING ON MULTIPLE-CHOICE EXAMS

Larry R. Nelson
Curtin University of Technology
Western Australia

Paper date: 10 January 2006

Lertap 5's RSA, response similarity analysis, has been enhanced—it now allows users to look at the statistical reliability of its similarity measures. This paper explains how to use and interpret Lertap's new RSA methods.

In a recent article (Nelson, 2006), I introduced "response similarity analysis", RSA, as seen in Lertap 5, an item and test analysis system released in the year 2000.

Lertap's original RSA method was based on the "Harpp-Hogan index", also known as the Harpp-Hogan ratio. In my paper, and in Lertap, this index, or ratio, is referred to as the "H-H index".

To understand the H-H index, consider the responses of any given pair of students who have sat the same multiple-choice exam. It is of course to be expected that some of the responses given by the students will be the same; in fact, if they're top students, they might each return a perfect exam score, in which case all of their item responses will be identical.

But in the more common case students will not have perfect papers. They will get some items correct, some wrong, and, for some reason, they may omit a few items, leaving them unanswered.

The H-H index is based on two characteristics of the students' item responses: the number of exact errors in common, EEIC, and the number of different responses, D. The H-H index is expressed as a ratio of these two numbers: H-H = EEIC/D.

Two students are said to have an "exact error in common" when they both select the same distractor to an item, that is, when they choose exactly the same incorrect answer to an item.

Harpp, Hogan, & Jennings (1996) reported on their observation of the H-H index's behavior, tracking it over years of application, reporting they found it to be "*a powerful indicator of copying*". They wrote:

> *Analyses of well over 100 examinations during the past six years have shown that when this number is ~1.0 or higher, there is a powerful indication of cheating. In virtually all cases to date where the exam has ~30 or more questions, has a class average <80% and where the minimum number of EEIC is 6, this parameter has been nearly 100% accurate in finding highly suspicious pairs.*

Difficulties with the H-H index

My article (Nelson, 2006) investigated the utility of the H-H index, and found faults with it. I was unable to replicate the results seen in Harpp, Hogan, & Jennings (1996) and ended up stating that the H-H index should be used only with "great caution".

I shared my findings with Harpp and Hogan. They kindly responded, suggesting the following revised guidelines (personal email correspondence from Professor David Harpp, December 2005):

> *Analyses of well over* ***1000*** *examinations during the past* ***13*** *years have shown that when this number is* ***~1.5*** *or higher,* ***at the same time with a probability deviating from the norm by ~5 sigmas or higher*** *there is a powerful indication of cheating. In virtually all cases to date where the exam has ~30 or more questions* ***as the lowest reasonable number****, has a class average <80% and where the minimum number of EEIC is* ***at least 8****, this parameter has been nearly 100% accurate in finding highly suspicious pairs* ***who have proximate seating. Further, we recommend discarding questions where the most chosen answer exceeds the "correct" answer.***

The word "number" above refers to the H-H index. The revised guidelines recommend working with an H-H index cutoff value of 1.5, and make it clear that the H-H index must be accompanied by a probability measure.

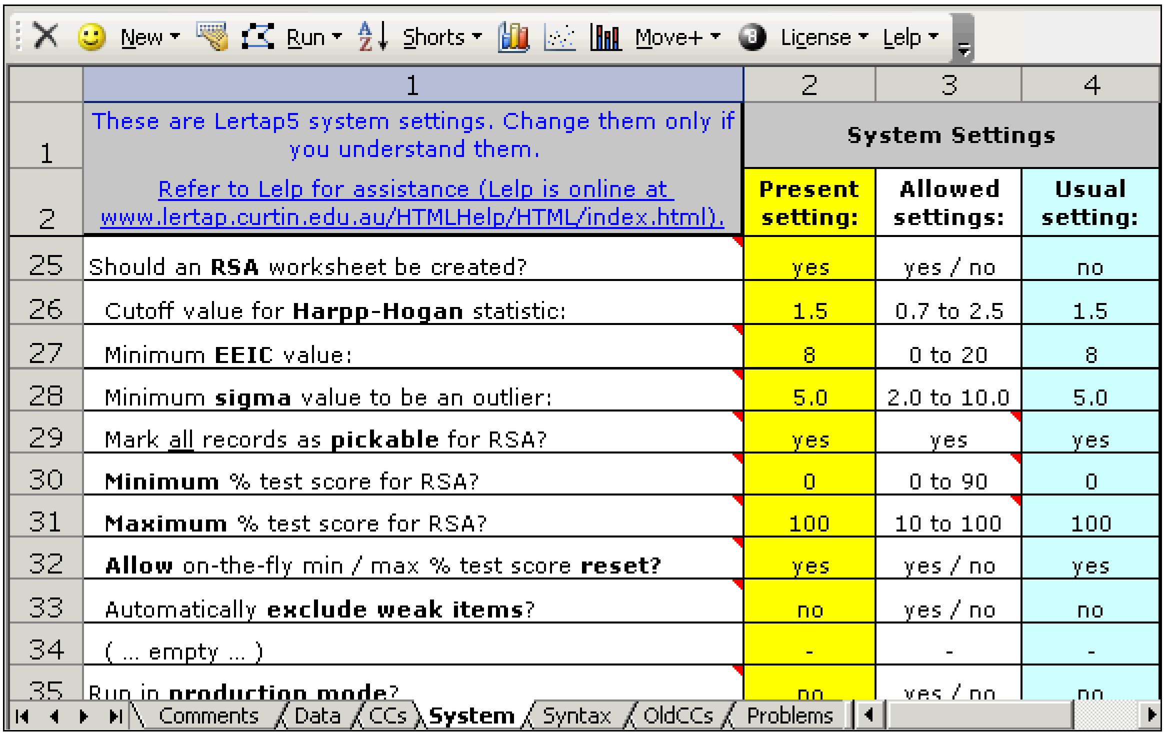
Wesolowsky's work was frequently cited in my paper (see, for example, Wesolowsky, 2000). In personal correspondence, Wesolowsky has indicated extensive experience with Harpp-Hogan methods, and he has, via personal email messages to me and via his website, cautioned that the H-H index should never be used without the "sigma" probability measure.

A revised version of Lertap: 5.6

With the revised Harpp-Hogan guidelines in hand, I have set about modifying Lertap. Lertap "5.6" is the result.

The Lertap5.xls file, that is, the Lertap 5 software system, has always included a worksheet named "System". It's this worksheet that's used to activate selected Lertap options, and it's here that the relevant new mods to Lertap are to be admired.

The partial screen snapshot below shows the options which relate to Lertap's response similarity analysis (RSA) methods:



	1	2	3	4
1	These are Lertap5 system settings. Change them only if you understand them.	System Settings		
2	Refer to Lelp for assistance (Lelp is online at www.lertap.curtin.edu.au/HTMLHelp/HTML/index.html).	**Present setting:**	**Allowed settings:**	**Usual setting:**
25	Should an **RSA** worksheet be created?	yes	yes / no	no
26	Cutoff value for **Harpp-Hogan** statistic:	1.5	0.7 to 2.5	1.5
27	Minimum **EEIC** value:	8	0 to 20	8
28	Minimum **sigma** value to be an outlier:	5.0	2.0 to 10.0	5.0
29	Mark all records as **pickable** for RSA?	yes	yes	yes
30	**Minimum** % test score for RSA?	0	0 to 90	0
31	**Maximum** % test score for RSA?	100	10 to 100	100
32	**Allow** on-the-fly min / max % test score **reset?**	yes	yes / no	yes
33	Automatically **exclude weak items**?	no	yes / no	no
34	(... empty ...)	-	-	-
35	Run in **production mode**?	no	yes / no	no

The RSA settings begin in row 25 above, and continue down through row 33. There are now no less than nine RSA options—before there were but three. Read all about them:

Should an **RSA** worksheet be created?

If this option is set to "yes", Lertap will produce a worksheet called RSAdata1 whenever the "Output item scores matrix" option is selected from Lertap's Run menu. This is the core worksheet for all of Lertap's RSA calculations. If Lertap is running in "production mode", there will be one RSAdata worksheet for each subtest. Once an RSAdata worksheet has been created, the "Response similarity analysis (RSA)" option may be taken from the Run menu. It is this option which produces Lertap's RSA reports.

Cutoff value for the **Harpp-Hogan** statistic:

This refers to the H-H index. As mentioned above, Harpp and Hogan now suggest a minimum of 1.5 for this index; in personal correspondence, I have seen Wesolowsky and Harpp running with values as low as 1.3 for the index.

Minimum **EEIC** value:

EEIC means "exact errors in common". In personal correspondence, I have seen Harpp drop this number as low as 6, depending on the number of test items – with a low number of items, say 30 or 35, I know that Professor Harpp will sometimes set EEIC=6.

Minimum **sigma** value to be an outlier:

As mentioned below, sigma refers to how far a student pair's probability measure is from the mean of the distribution of probability measures. Sigma is a z-score. If the probability measures are normally distributed, a z-score of +5.0 or -5.0 more is a very rare outcome indeed—only 0.0000003 of the area under a normal distribution lies beyond a z-score of 5.0. (Experience indicates that the probability measures developed by Harpp and Hogan often come very close to having a normal distribution.) In practical terms, an exam given to three thousand students will produce about five million pairings of students; if the students have not colluded in their item responses, only about two of the student pairs can be expected to have a sigma greater than 5.0, assuming that the distribution of probability measures follows a normal distribution.

Mark all records as **pickable** for RSA?

This option is, in fact, not yet an option. It may be activated at a future date. In the present version of Lertap, students may be excluded from an RSA analysis by removing the comment (the red triangle) from their RSAdata records; students will also be excluded if their test score does not fall within the range of scores specified by the minimum % and maximum % test score values set in the System worksheet (see immediately below).

Minimum % test score for RSA?

Maximum % test score for RSA?

These two settings determine which students will be included in any RSA analysis. A minimum of 0 (zero) and maximum of 100 will see all students included. Note that experienced users of Harpp Hogan methods will often run several RSA analyses for any given test. They may start with a 0-100 range for these settings, or 30-100, and then reprocess the data with revised settings (this is discussed below).

Allow on-the-fly min / max % test score **reset**?

If this option is set to "yes", then Lertap will ask you to enter the minimum and maximum % test scores each time you select the "Response similarity analysis (RSA)" option from the Run menu. This completely over-rides the Minimum and Maximum % test score settings in the System worksheet.

Automatically **exclude weak items**?

For RSA work, "weak items" are those where the number of students selecting the item's correct answer is less than the number selecting one of the distractors, or less than the number of students who omitted the item. If this option is set to "no", then Lertap will pause every time it encounters a "weak item", asking if you'd like to exclude it from the RSA analysis. If the option is set to "yes", then weak items are automatically excluded. Excluding weak items is strongly recommended; if a test has weak items, the EEIC measure will be inflated, resulting in more "suspects pairs", that is, more student pairs whose item responses may be judged suspiciously similar (possibly implying cheating). Is it common for tests to have weak items? Yes, it is; difficult items with poorly-functioning distractors will often fall under this definition of a weak item. Note that a "weak item", in RSA terms, does not necessarily mean a bad item—bad items are, generally, those with a negative discrimination index; it is possible for an item to be weak, in RSA terms, but still have an adequate discrimination figure.

Evaluating these new guidelines/options

The Nelson (2006) article applied the old Harpp-Hogan guidelines to three data sets. The old guidelines suggested an EEIC minimum of 6, and an H-H index minimum of 1.0. Student pairs with EEIC and H-H index values above these minimum figures were said to have suspiciously-similar responses, thus qualifying for the implied label of "suspect cheaters".

My article looked at the accuracy of the old guidelines by comparing their suggested results with those reported by Wesolowsky's SCheck program (Wesolowsky, 2000), and, where possible, by referring to known exam seating patterns. I found that the old guidelines consistently produced too many false positives—more suspect cases than found by SCheck and/or by reference to known seating arrangements.

With the enhanced version of Lertap working under the revised Harpp-Hogan guidelines, I returned to the three data sets covered in my original article, and compared results.

For my first data set, from "test center A", the revised Lertap found fewer suspect cases. I originally had five suspect cases, but now, operating with the new whiz-bang version, 5.6, using the revised guidelines, the last two cases have dropped off the radar, leaving me with just three suspect pairs. This is a positive outcome—as just noted, a problem with the original approach was that it returned too many "hits". With the revised guidelines in place, the number of hits has reduced, and is more in line with expected results.

The second data set, "center B data set 1", had no less than forty (40) hits with the original guidelines in operation. Applying the new guidelines had a dramatic effect: only one (1) suspect case was uncovered, corresponding to the first pair listed in Table 2 of the Nelson (2006) paper. This was exactly in line with expected results.

The third data set involved another exam from test center B. Five (5) suspect pairs were originally found using the former guidelines; this number reduced to zero with the new guidelines in place. Once again this was exactly in line with expectations.

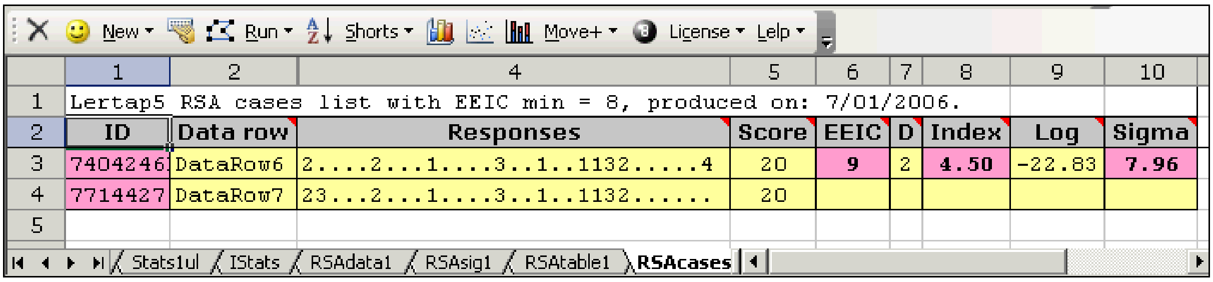
These results are pleasant. I will continue to compare results from Lertap's revised Harpp-Hogan methods with those from other programs, but, for the moment, I am satisfied that the revised approach seems solid, and represents a definite improvement.

The rest of this document looks at how the new version of Lertap works, starting with glances of the three RSA reports produced whenever the "Response similarity analysis (RSA)" option is taken from the Run menu.

Lertap's RSAcases report

Lertap 5.6 produces three RSA reports: RSAsig, RSAtable, and RSAcases. Of these, the RSAcases report will be the one of most use to the majority of users. It is this report which reveals how many suspect cases were found.

A sample RSAcases report is shown below. The colors used to highlight parts of the sample are important, so, in case you're looking at the sample in black and white, I need to mention that pink background coloring (Excel calls it "rose") is used below under the ID column, under EEIC, under Index, and under Sigma; otherwise the cells in rows 3 and 4 have a light yellow background color. (When are you going to buy a color printer?)



	1	2	4	5	6	7	8	9	10
1	Lertap5 RSA cases list with EEIC min = 8, produced on: 7/01/2006.								
2	ID	Data row	Responses	Score	EEIC	D	Index	Log	Sigma
3	7404246	DataRow6	2....2...1....3..1..1132.....4	20	9	2	4.50	-22.83	7.96
4	7714427	DataRow7	23...2...1....3..1..1132......	20					
5									

The sample was produced by submitting one of the Harpp, Hogan, & Jennings (1996) data sets to Lertap 5.6.

The data set involved a chemistry test of 30 items given to 106 students. Lertap's RSAcases report indicates that only one suspect pair was found.

Both of the students comprising this pair had a test score of 20, as seen under the Score column above.

The Responses column indicates how each student answered each of the 30 items. A full stop (or a "period") in the string of responses corresponds to a correct answer—note that each response string above has 20 of these.

The "2" found at the beginning of the two response strings indicates that both students chose option 2 for the first item. One of the students got the second item correct, while the other answered "3" for this item. Both students got the third, fourth, and fifth items correct, and both chose option "2" for the sixth item.

The two response strings show remarkable similarities. In fact, there are only two differences—these occur at the second item, and at the last item. Both students got 10 of the 30 items wrong; of their 10 errors, 9 were identical.

The RSAcases report summarizes these figures under its EEIC and D columns. EEIC gives the number of exact errors in common, while D indicates the total number of differences in the two response strings.

The Index column corresponds to what I have referred to as the H-H index; Harpp & Hogan, and others, often refer to this as the "ratio": it's EEIC divided by D.

The Log column is the "logarithm of PROB" seen in Harpp & Hogan (1993), while Sigma corresponds to the Greek letter sigma also found in Harpp & Hogan (1993). Sigma is computed by using the mean and standard deviation values obtained from the non-suspect distribution—what we're doing is "inserting" the Log values for suspect pairs into the distribution of Log values observed among the non-suspect pairs, and finding how many standard deviations they are from the mean.

Harpp & Hogan (1993) found that the distribution of Log often closely followed the characteristics of a Gaussian, or normal, distribution. Sigma is simply Log expressed on a standardized scale with mean zero, standard deviation one—Sigma is a "z-score". If Log is normally distributed, then the probability of obtaining any given Sigma value may be found by reference to the normal curve. In a Gaussian, or normal, distribution, Sigma values in excess of 5.00 are very, very rare, found only once in (approximately) every 3.5 million cases.

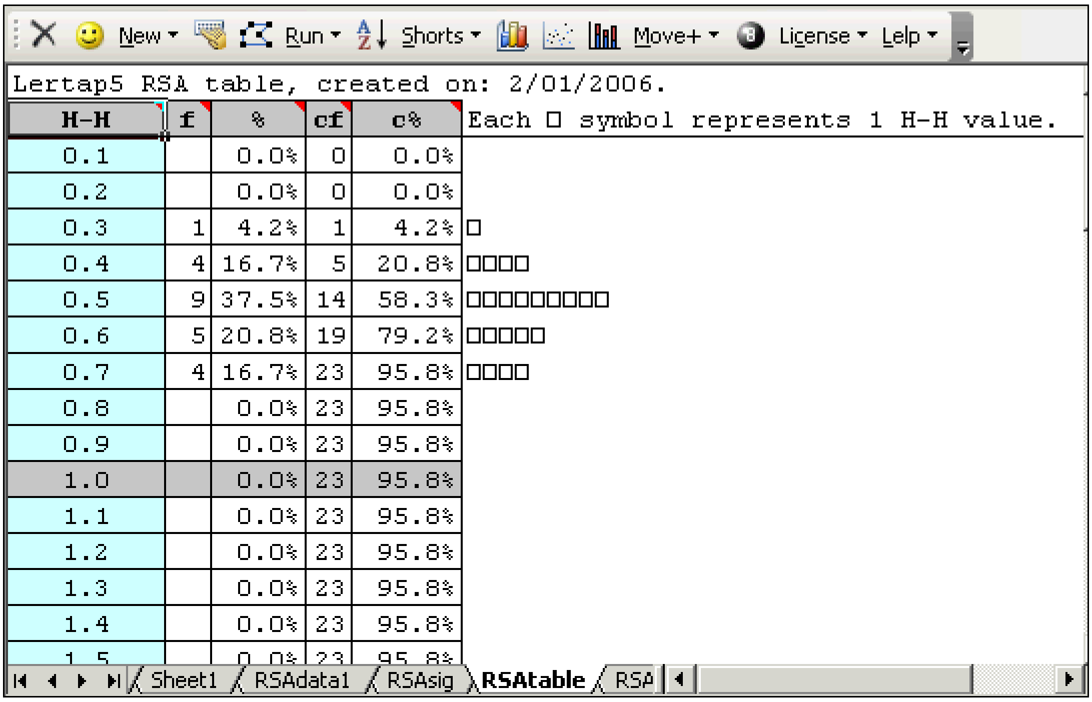
Now, having looked at the meaning of most of the RSAcases columns, I can mention how to interpret the RSAcases report.

Firstly, before the responses of any student pair may be considered "suspect", the pair must have EEIC and H-H index values at or above the cutoff levels set in Lertap's System worksheet. If they do, they'll be entered in the RSAcases report.

Secondly, if the pair's Sigma value is equal to or greater than the cutoff value seen in the System worksheet, the pair's RSAcases lines will be flagged with special coloring—if you're reading this on a color screen, or have printed this document on a color printer, you will note that the special color is applied to the ID, EEIC, Index, and Sigma columns. The color is a shade of pink, referred as "rose" in the Windows Excel color palette.

Lertap's RSAtable report

The RSAtable report is a "graph" of H-H index values, designed to resemble those seen in Figures 1, 2, and 3 of Harpp, Hogan, & Jennings (1996). If a student pair has an EEIC above the cutoff value set in the System worksheet, their H-H index value will be "plotted" in the RSAtable report, no matter its size. By far the majority of Harpp-Hogan index values will be less than 1.0.



Lertap5 RSA table, created on: 2/01/2006.

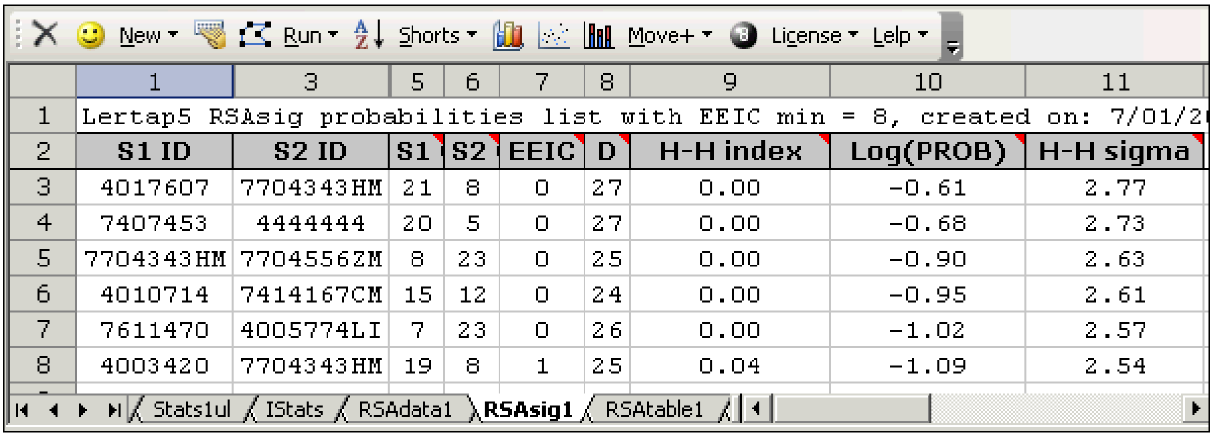
H-H	f	%	cf	c%	Each □ symbol represents 1 H-H value.
0.1		0.0%	0	0.0%	
0.2		0.0%	0	0.0%	
0.3	1	4.2%	1	4.2%	□
0.4	4	16.7%	5	20.8%	□□□□
0.5	9	37.5%	14	58.3%	□□□□□□□□□
0.6	5	20.8%	19	79.2%	□□□□□
0.7	4	16.7%	23	95.8%	□□□□
0.8		0.0%	23	95.8%	
0.9		0.0%	23	95.8%	
1.0		0.0%	23	95.8%	
1.1		0.0%	23	95.8%	
1.2		0.0%	23	95.8%	
1.3		0.0%	23	95.8%	
1.4		0.0%	23	95.8%	
1.5		0.0%	23	95.8%	

The RSAtable report was the main RSA focus in the previous version of Lertap, used when it was thought the original Harpp-Hogan guidelines would suffice for identifying possible cheaters. In Lertap 5.6 a new report, "RSAsig", has been added, in large part supplanting RSAtable.

Lertap's RSAsig report

RSAsig is quite a special report. The "sig" part of this worksheet's name refers to the fact that the worksheet features "Sigma" values—it is the distribution of these Sigma values which establishes the basic reference "curve" used as the final tool in identifying which student pairs have item responses which appear to differ significantly from the norm.

The screen snapshot below indicates RSAsig's basic format:

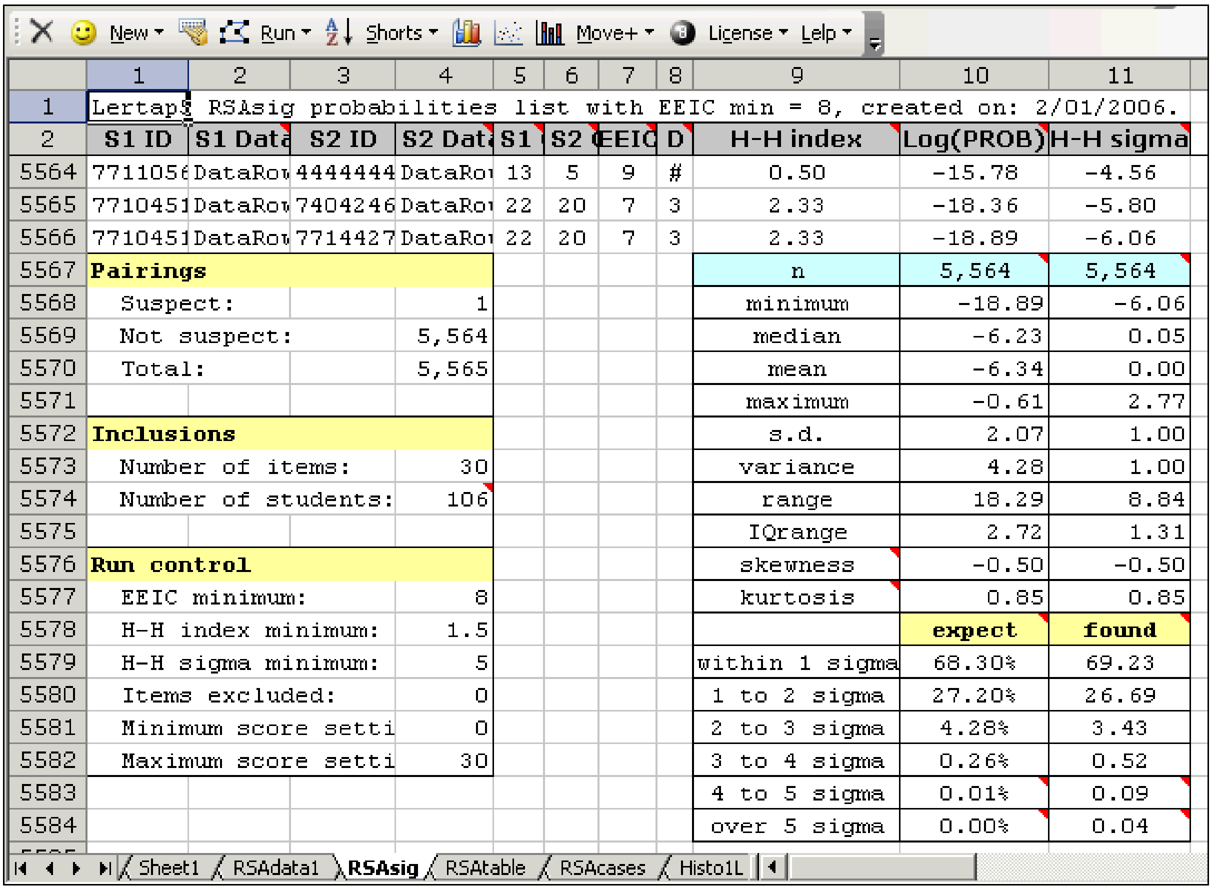


	1	3	5	6	7	8	9	10	11
1	Lertap5 RSAsig probabilities list with EEIC min = 8, created on: 7/01/2								
2	S1 ID	S2 ID	S1	S2	EEIC	D	H-H index	Log(PROB)	H-H sigma
3	4017607	7704343HM	21	8	0	27	0.00	-0.61	2.77
4	7407453	4444444	20	5	0	27	0.00	-0.68	2.73
5	7704343HM	7704556ZM	8	23	0	25	0.00	-0.90	2.63
6	4010714	7414167CM	15	12	0	24	0.00	-0.95	2.61
7	7611470	4005774LI	7	23	0	26	0.00	-1.02	2.57
8	4003420	7704343HM	19	8	1	25	0.04	-1.09	2.54

The RSAsig report is automatically sorted by Lertap so that the Sigma values, shown in the report as “H-H sigma”, are presented in decreasing order. In the report pictured above, the highest Sigma for all the student pairs whose results have been tabulated in the report was 2.77. The corresponding Log(PROB) figure was -0.61.

Log(PROB) is really the heart of this report, that is, the working part, the driving force, the yin and yang. PROB is the response probability measure developed by Harpp & Hogan, and featured in their 1996 article (Harpp, Hogan, & Jennings, 1996). As pointed out in the article, PROB values “can become very small”, occasionally so small as to cause computational difficulties on some computers. To circumvent these potential problems, the logarithm of PROB is used as the core index, expressed as Log(PROB).

Important summary information is given at the bottom of the RSAsig report in two relatively small sections. A screen snapshot is shown below (note that the width of some of the columns has been manually reduced so as to fit the sample within page boundaries; in the screen pictured above, the second and fourth columns were entirely hidden).



	1	2	3	4	5	6	7	8	9	10	11
1	Lertap5 RSAsig probabilities list with EEIC min = 8, created on: 2/01/2006.										
2	S1 ID	S1 Data	S2 ID	S2 Data	S1	S2	EEIC	D	H-H index	Log(PROB)	H-H sigma
5564	7711056	DataRow	4444444	DataRo	13	5	9	#	0.50	-15.78	-4.56
5565	7710451	DataRow	7404246	DataRo	22	20	7	3	2.33	-18.36	-5.80
5566	7710451	DataRow	7714427	DataRo	22	20	7	3	2.33	-18.89	-6.06
5567	**Pairings**								n	5,564	5,564
5568	Suspect:			1					minimum	-18.89	-6.06
5569	Not suspect:			5,564					median	-6.23	0.05
5570	Total:			5,565					mean	-6.34	0.00
5571									maximum	-0.61	2.77
5572	**Inclusions**								s.d.	2.07	1.00
5573	Number of items:			30					variance	4.28	1.00
5574	Number of students:			106					range	18.29	8.84
5575									IQrange	2.72	1.31
5576	**Run control**								skewness	-0.50	-0.50
5577	EEIC minimum:			8					kurtosis	0.85	0.85
5578	H-H index minimum:			1.5						**expect**	**found**
5579	H-H sigma minimum:			5					within 1 sigma	68.30%	69.23
5580	Items excluded:			0					1 to 2 sigma	27.20%	26.69
5581	Minimum score setti			0					2 to 3 sigma	4.28%	3.43
5582	Maximum score setti			30					3 to 4 sigma	0.26%	0.52
5583									4 to 5 sigma	0.01%	0.09
5584									over 5 sigma	0.00%	0.04

Basic information pertaining to the RSA analysis itself is given under the Pairings, Inclusions, and Run control headings.

A total of 106 students had scores within the range set by the Minimum and Maximum score settings, which in this case were 0 to 30.

Given this number of students, how many possible student pairings would there be? The answer is 5565, a value determined by (N(N-1))/2, where N=106.

Of all these pairs, Lertap 5.6 found only one which would be suspect, given the EEIC and H-H index minimum values of 8 and 1.5.

Details summarizing the Log(PROB) and Sigma results are given to the right of the report.

In this example, we see that the minimum Sigma value found among the non-suspects was -6.06. "We do?", you say? Sure. Remember that the Sigmas have been sorted from highest to lowest. The Sigmas are found in column 11. The last Sigma in this column is -6.06.

Why do I say "found among the non-suspects"? Ah, a very crucial question indeed. All of the data rows in the RSAsig report, starting from row 3 down to where the "Pairings" row appears, pertain only to non-suspects, that is, only to those student pairs whose EEIC is less than the EEIC cutoff, or whose H-H index is less than the cutoff value for the Harpp-Hogan statistic. If a student pair has EEIC above the cutoff, and also has H-H index above its cutoff, then the students' results are found in the RSAcases report—these student pairs are our "suspects"; their results are not included in the RSAsig report.

Here it will be important to mention that it is always the lower tail of the Log and Sigma distributions which becomes the main focus of the Harpp-Hogan based RSA methods used in Lertap 5.6.

In RSAsig, the summary of *non-suspect* pairs, there will be positive and negative Sigma values. However, the Sigma values corresponding to the *suspect* pairs will never (ever) be positive. No? No. We always find our suspects in the lower, left end of the

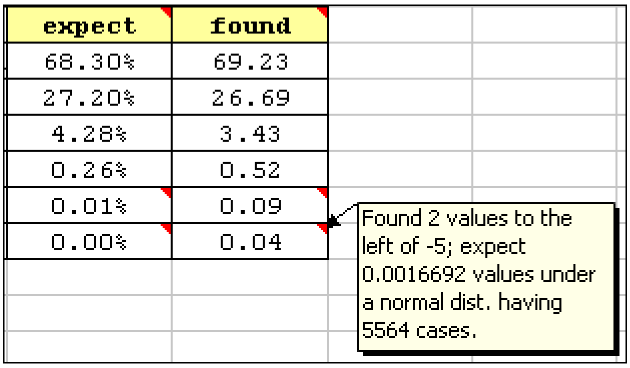
Sigma distribution, the negative end. When Sigma values appear in the RSAcases report, the summary of suspect pairs, the minus sign on Sigma is dropped for convenience, and we say that a suspect pair is, for example, "almost eight sigmas out", instead of saying "-7.96".

Now look at the cute little table at the bottom right, with the "expect" and "found" headings. This table compares actual distribution results with those of a normal distribution.

Under a normal distribution, 68.30% of all cases will lie within one standard deviation, one "sigma", of the mean value. For this data set, 69.23% of the 5564 non-suspect pairs were found in this region, slightly above the expected 68.30%.

Of particular interest are the tails of the distribution, starting from 3 sigmas and extending outwards. Under the normal curve, only 0.26% (0.0026 of all values) will be expected to the right of +3 sigma, and to the left of the -3 sigma points. In this case we've found twice the expected number of cases in the tails.

If we've got the Harpp-Hogan minimum sigma setting at 5, we may have special interest in the last row of the "expect – found" table. Here values get very small indeed. Lertap 5.6 has comments attached to the bottom cells which provide more details. For example, letting my mouse hover over the cell with 0.04 prompts the comment to appear:



expect	found
68.30%	69.23
27.20%	26.69
4.28%	3.43
0.26%	0.52
0.01%	0.09
0.00%	0.04

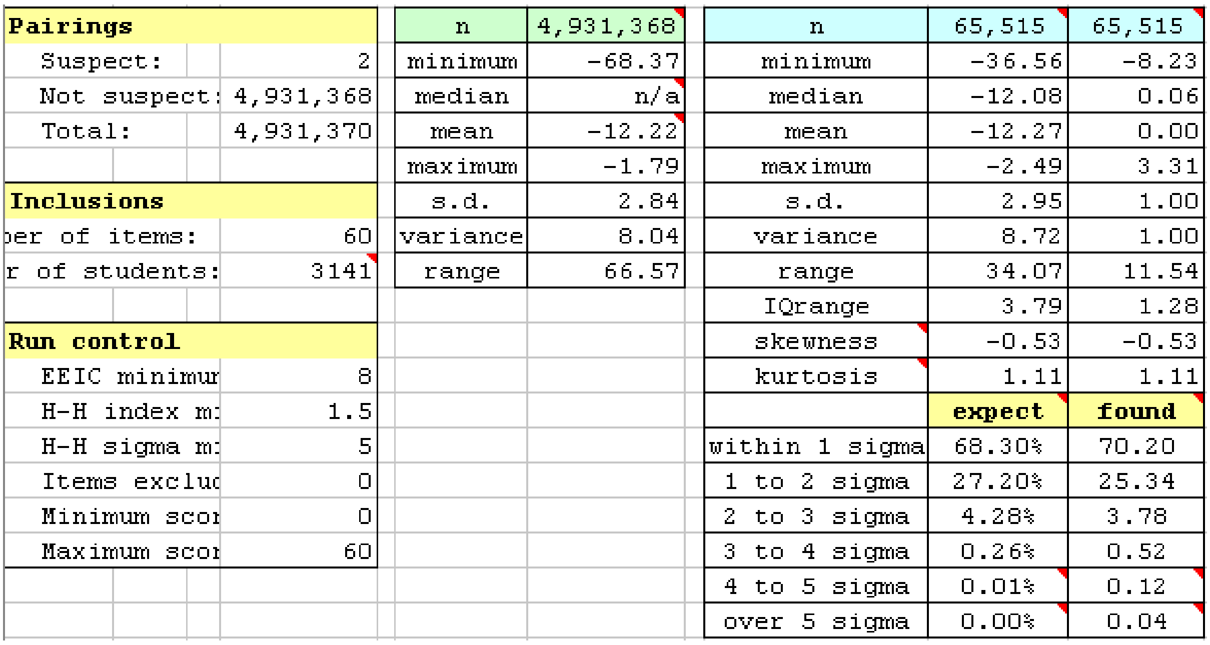
Found 2 values to the left of -5; expect 0.0016692 values under a normal dist. having 5564 cases.

There will probably be many users who won't spend a great amount of time with the RSAsig report. If administrators are simply interested in detecting the extent of possible cheating, they will likely just look at the RSAcases report and may leave things at that; the more pink entries in RSAcases, the more evidence of possible cheating. But others will often want to open RSAsig, and give it a real squiz. For example, above we've set the minimum value of Sigma at 5, a setting based on what we know to be the "odds" of a normal curve. Ordinarily we'd expect much less than one Sigma value to fall to the left of -5; here we've found two (according to the comment seen above). If the percentage found in the tails of the Sigma distribution were to depart markedly from the expected values, we might do well to alter the interpretation of the odds, and consider setting the minimum Sigma higher.

Astute readers (that is, of course, all readers), may have detected two Sigma values for the non-suspects which are above the cutoff sigma of 5. There's a -5.80 value, and a -6.06 value (refer to column 11 of rows 5564 and 5565 of the screen snapshot above). If we'd lowered the minimum EEIC figure for this data set to 7, these two pairs would have become suspects, and they would have been "pinked" (or "rosed") in the RSAcases report. Why? Look at their EEIC and H-H index values; they'd both be equal to or greater than the cutoffs if we lowered EEIC min to 7.

As it turns out, these pairs of students did cheat. They were seated next to each other, and later fessed up (admitted to sharing answers). Lertap 5.6 has failed us in this case, a fact which brings home a message: these methods are not bullet-proof. Users may recognize a need to play with cutoff values at times, especially when the number of test items is, say, less than 40.

Finally, there's an inherent limit to the size of the RSAsig report in Lertap 5.6: as this document went to press, an Excel worksheet could have no more that 65,536 rows. Using the (N(N-1))/2 expression mentioned above, it turns out that this corresponds to about 360 students. Whenever the number of students exceeds this approximate figure, Lertap's RSAsig report changes to the format seen below:



Pairings	
Suspect:	2
Not suspect:	4,931,368
Total:	4,931,370
Inclusions	
ber of items:	60
r of students:	3141
Run control	
EEIC minimur	8
H-H index m:	1.5
H-H sigma m:	5
Items excluc	0
Minimum scor	0
Maximum scor	60

n	4,931,368
minimum	-68.37
median	n/a
mean	-12.22
maximum	-1.79
s.d.	2.84
variance	8.04
range	66.57

n	65,515	65,515
minimum	-36.56	-8.23
median	-12.08	0.06
mean	-12.27	0.00
maximum	-2.49	3.31
s.d.	2.95	1.00
variance	8.72	1.00
range	34.07	11.54
IQrange	3.79	1.28
skewness	-0.53	-0.53
kurtosis	1.11	1.11
	expect	**found**
within 1 sigma	68.30%	70.20
1 to 2 sigma	27.20%	25.34
2 to 3 sigma	4.28%	3.78
3 to 4 sigma	0.26%	0.52
4 to 5 sigma	0.01%	0.12
over 5 sigma	0.00%	0.04

The right-most table above is based on 65,515 non-suspect student pairs, but there were, in this case of 3141 students, almost five million such pairs. Whenever there are more non-suspect pairs than can be listed in the RSAsig worksheet, the small center table is added to the RSAsig report. The data in this little table pertain to all of the non-suspect pairs, to almost five million in this case[1]. The data in the original table, the larger one to the right, cover only those non-suspect cases actually listed in the RSAsig worksheet; in this example, that covers 65,515 pairs. We could look at the data in the right-most table as the results from a sample of all non-suspect pairs, and then compare the mean and s.d. (standard deviation) figures from the two tables to roughly assess how representative the sample is of the population.

Note that the Sigma values seen in the RSAcases report are computed using the mean and s.d. figures from the entire non-suspect distribution (-12.22 and 2.84 above). These Sigma values will always be negative, but, when carried forward into RSAcases, the minus sign is dropped.

Using Lertap's RSA options

The settings which control how Lertap 5.6 goes about its RSA business are all managed in the System worksheet, as discussed above, near the beginning of this opus.

A core part of Lertap RSA is the initial identification of suspect pairs. There are two settings which control this: the "Cutoff values for the Harpp-Hogan statistic", and the "Minimum EEIC value". When the item responses of any given pair of students meet or exceed both of these values, the pair is placed in the "suspect" category. Otherwise the pair join the non-suspect group. It is common for the number of suspect pairs to be much, much less than the number of non-suspects.

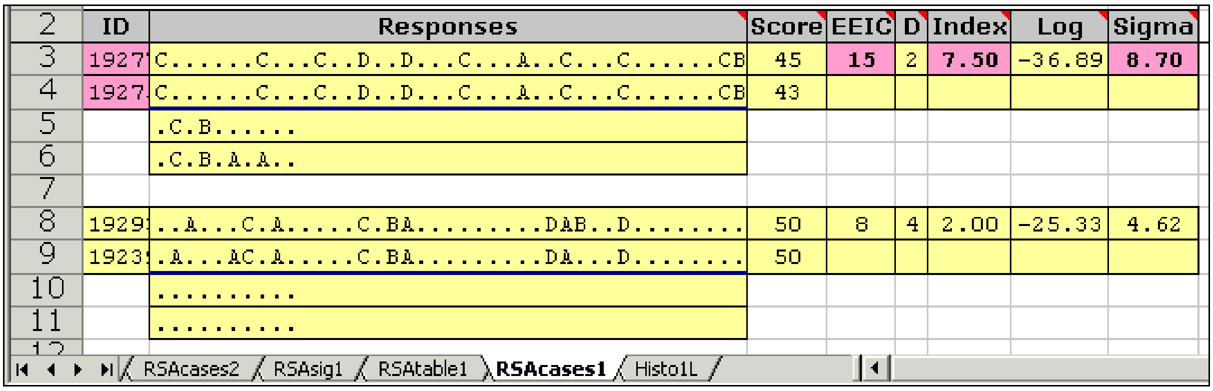
[1] It would be possible to compute the right-most table using all non-suspect pairs, but the computational time would double; it took about an hour to process the 3141 cases on a laptop running Windows XP with a Pentium processor said to be running at 1.8 GHz, with 1 G of RAM on board.

The RSAcases report contains information for all the suspect pairs. When the Sigma value for a suspect pair meets or exceeds the "Minimum sigma value to be an outlier", special coloring is applied to the pair's RSAcases rows—they "go pink", or "become rosed". It is these rose-colored pairs who may become the focus of further investigation. If it can be shown that the pair had access to cheating opportunities, then some subsequent action might be taken against them. What might constitute a "cheating opportunity"? Adjacent seating would be the primary example.

The RSAsig report contains information for the non-suspect pairs; the mean and s.d. of the Log(PROB) values in RSAsig are used to compute the Sigmas seen in both RSAcases and RSAsig. It is important to note that the figures seen in RSAsig exclude the suspect pairs. Harpp-Hogan methods involve establishing a distribution of probabilities for non-suspects (RSAsig). Once we have this, we "insert" the Log(PROB) values for the suspects into the distribution to see if they fall beyond the cutoff point—the RSAcases report summarizes the results of this process.

Another important point is that the entire process of separating suspects from non-suspects is controlled by other settings. The "Minimum % test score" and "Maximum % test score" settings are often altered by "serious" Harpp-Hoganers so as to tune the non-suspect distribution to a particular range of scores.

Consider, for example, the RSAcases report shown below (for the benefit of those viewing the sample in black and white, the top pair, rows 3 through 6, has pink coloring under the ID, EEIC, Index, and Sigma columns, while the second pair does not):



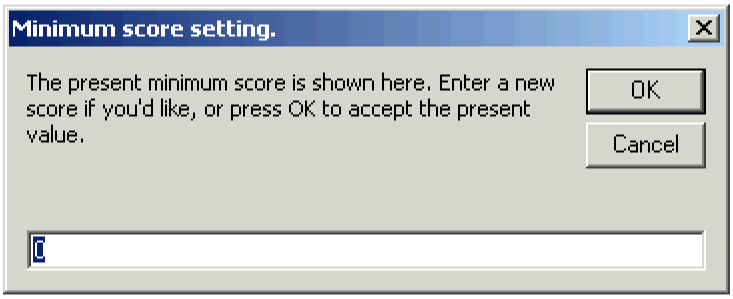
2	ID	Responses	Score	EEIC	D	Index	Log	Sigma
3	1927	C......C...C..D..D...C...A..C...C......CB	45	15	2	7.50	-36.89	8.70
4	1927	C......C...C..D..D...C...A..C...C......CB	43					
5		.C.B......						
6		.C.B.A.A..						
7								
8	1929	..A...C.A.....C.BA.........DAB..D........	50	8	4	2.00	-25.33	4.62
9	1923	.A...AC.A.....C.BA.........DA...D........	50					
10		....						
11		....						

The results seen in this RSAcases report are from a 60-item test. Lertap 5.6 has identified two suspect pairs, and has "rosed" the first pair as its Sigma value is beyond the 5.00 setting used in the analysis (note: not all of the 60 item responses are shown above).

We might wonder about what would happen to the second pair's Sigma value if we asked Lertap to tune the non-suspect distribution so that it focused on, say, a score range of 48 to 52.

Who, you muse, would ever think of having such a wonder? Those who are familiar with the literature, of course. This wonder range includes 50, the score earned by both members of the second pair.

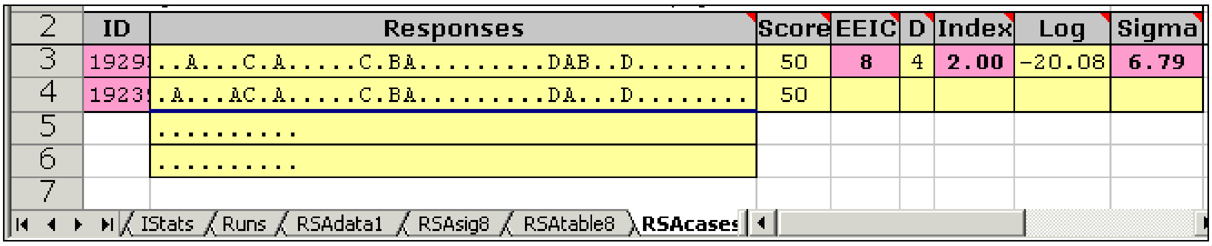
In order to have our wonder, the "Allow on-the-fly min / max % test score reset?" option has to be set to Yes. Once it is, Lertap will pause at the start of an RSA run in order to collect Min and Max score values for use in its RSA analyses, presenting two dialog boxes like the one seen below:



Minimum score setting.

The present minimum score is shown here. Enter a new score if you'd like, or press OK to accept the present value.

Okay? Here's RSAcases when Lertap was run with Min=48, Max=52, the answer to our wonder (black and white readers note: pink backgrounds are seen under the ID, EEIC, Index, and Sigma columns):



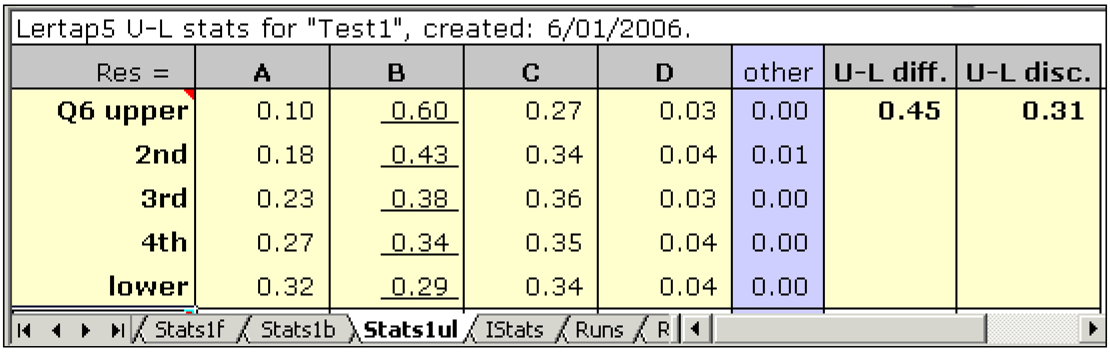
2	ID	Responses	Score	EEIC	D	Index	Log	Sigma
3	1929	..A...C.A.....C.BA.........DAB..D........	50	8	4	2.00	-20.08	6.79
4	1923	.A...AC.A.....C.BA.........DA...D........	50					
5		....						
6		....						
7								

The results for this pair of students have now "gone pink"—their Sigma now exceeds the cutoff of 5.00.

What's happened here, how can this be (wonder of wonders)?

In the process of fine-tuning the scores range, honing in on the 48-52 area, Lertap has adjusted its item-response statistics so that they're based on the cohort of students whose test scores fall within the new range. There were some 511 students in this range.

Can I expect item statistics to change much if I look at different score levels? Yes, I can. It's pretty easy to get an idea of how item response statistics change over score levels by looking at Lertap's Statsul report. Here's a sample:



Lertap5 U-L stats for "Test1", created: 6/01/2006.

Res =	A	B	C	D	other	U-L diff.	U-L disc.
Q6 upper	0.10	0.60	0.27	0.03	0.00	0.45	0.31
2nd	0.18	0.43	0.34	0.04	0.01		
3rd	0.23	0.38	0.36	0.03	0.00		
4th	0.27	0.34	0.35	0.04	0.00		
lower	0.32	0.29	0.34	0.04	0.00		

The little report above gives item Q6's response statistics by score quintiles, starting with the upper 20%, going to the lower 20%. The correct answer to Q6 was B. Notice how the item responses changed over the five score groupings—in the upper group, 60% got the item right, dropping to just 29% in the lower group. And note how the popularity of distractor A changes over the groups: only 10% of the upper group were fooled by this distractor, climbing to 32% in the lower group.

Harpp-Hogan probability statistics are based on item response frequencies. Response frequencies can be expected to vary by score levels, as we've just seen. When I told Lertap to work with a score range of 48-52, it internally recalculated the item response frequencies for the corresponding student cohort, and used these new frequencies for its probability calculations.

Lertap dutifully found new Log(PROB) values for our suspect pair, and for all the non-suspect pairs. Then it worked out the corresponding Sigma values. At the end of the day, our suspect pair's Sigma turned pink. When this pair's item responses were compared with those from other student pairs having about the same test score, the similarity of their responses stood out in a marked (pink) manner. We might check our seating assignment records and, finding the two students to have been seat neighbors during the exam, call them in to the office for a wee chat.

I can imagine that some head scratching may be going on about now, some chair-fidgeting, some "gee I think I better have a break, maybe the fish are biting" thoughts cropping up.

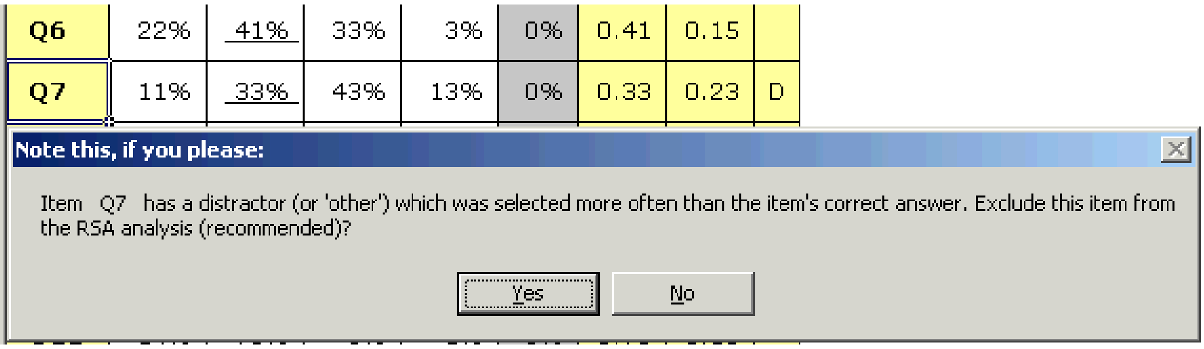
Does this mean I should always run my RSA analyses by score groups? And, if so, how many groups should I use?

What I suggest is an initial run over all scores. The EEIC and H-H index values do not depend on test score—they're invariant. No matter what score range is under consideration, a give student pair will always have the same EEIC and H-H index, and it's these two statistics which are used in the initial identification of suspects.

Then, given the RSAcases report for all scores, clean your reading glasses, get a fresh cuppa of something, and give the report a good squiz. If there are some Sigma values which are not pink, consider another RSA run with a re-defined score range, and have another go as I have done here (you might want to have two goes, or three).

Finally, let's consider one more really relevant matter: weak items. Lertap 5.6 has an option to "Automatically exclude weak items?". Weak items are one whose response frequencies might serve to muck up the RSA analyses; such items are usually ones where an item distractor was selected more often than the item's correct answer. These items can artificially increase the EEIC measure, resulting in an inflated number of suspect pairs; in turn, this may adversely affect the probability calculations.

Lertap's default setting for this option is No. You can, of course change it to Yes. Leaving it at No will give you an idea of how many weak items there were as you start an RSA analysis. This is so because Lertap will pause to display a dialog box such as this one:



Q6	22%	41%	33%	3%	0%	0.41	0.15	
Q7	11%	33%	43%	13%	0%	0.33	0.23	D

Note this, if you please:

Item Q7 has a distractor (or 'other') which was selected more often than the item's correct answer. Exclude this item from the RSA analysis (recommended)?

The correct answer to Q7 was the second option, taken by 33% of all students. However, a larger percentage, 43%, chose the third option, a distractor. This difference is not excessive, and I note that each of the other distractors, fooling 11% and 13%, were working—I might retain this item, clicking on No. Often times the differences are much more marked, resulting in a Yes click.

If there are several weak items, excluding them can substantially impact on the RSA results. For example, above I showed what happened to a suspect pair as I fine-tuned the score range used in the RSA analysis. Lertap told me that the test had three weak items; had I eliminated them, the suspect pair's EEIC figure would have dropped below the minimum value of 8, and they'd move over to join the non-suspect group.

Practical suggestions

So many comments, so many options, so many factors which seem to affect results—what's best to do? If you're new to this topic, I would suggest letting Lertap run with its default RSA settings, as seen way up above in the System worksheet. If you do, then, as you start an RSA run, Lertap will give you the chance to change the minimum and maximum score settings, and to have weak items excluded from the analysis.

I would, initially at least, leave the original minimum and maximum scores settings at 0 and 100, and answer Yes to each of the "exclude this item" queries which Lertap might display.

Look at the RSAcases report which results. It'll give you a quick idea of how many suspects there are, and rose-color those suspects whose item responses were truly far from the norm. If there are some suspect pairs with Sigma values close to but below the cutoff, consider another RSA run, changing the minimum and maximum score settings so as to select a narrower score range, one which is centered on the scores of the suspect students. For example, if the two students comprising the suspect pair had test scores of, say, 42 and 51, do another RSA run, telling Lertap to narrow its focus to, say, the 41 to 52 score range.

Of course, you'll always want to keep in mind that Lertap's results do not categorically say that students have cheated. You'll want to refer to seating charts to see if suspect pairs were within cooee[2], or if some sort of suspicious behavior may have been reported by the exam invigilators. A Lertap RSA analysis is a screening tool, one which will point out student item responses which have unexpected similarities. When it can also be shown that the students pointed to by Lertap actually did have the opportunity to share answers, well, the cards may then be stacked, suggesting some subsequent action.

There will always be the chance of forgetting about all of this, tossing Lertap out the window, never applying it again. What's that, throw away Lertap? Yes. Eliminate the opportunity for students to cheat. Have scrambled exams; random seat assignments during exams; lots of invigilators; and heaps of space separating the students.

If I throw away Lertap, will I get a refund on my purchase? Well, no, but you know you wouldn't toss it out—what would you then use to get all the quintile plots, eigenvalue scree plots, histograms and other wonders made by Lertap?

References

Harpp, D.N. & Hogan, J.J. (1993). Crime in the classroom – detection and prevention of cheating on multiple-choice exams. *Journal of Chemical Education,* 70(4), 306-311.

(URL not presently available; try working from the next URL below.)

Harpp, D.N., Hogan, J.J., & Jennings, J.S. (1996). Crime in the classroom – Part II, an update. *Journal of Chemical Education,* 73(4), 349-351.

http://jchemed.chem.wisc.edu/Journal/Issues/1996/Apr/abs349.html

Nelson, L.R. (2000). *Item analysis for tests and surveys using Lertap 5.* Perth, Western Australia: Faculty of Education, Language Studies, and Social Work, Curtin University of Technology.

http://www.lertap.curtin.edu.au

[2] In Australia, saying that two people are "within cooee" means that they're within shouting distance.

Nelson, L.R. (2006). Using selected indices to monitor cheating on multiple-choice exams. To be published in 2006 in the *Thai Journal of Educational Research and Measurement (ISSN 1685-6740):* 4(1), xx-xx. (In press.)

http://www.lertap.curtin.edu.au/Documentation/JERM2006.doc

Wesolowsky, G.O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics,* 27(7), 909-921.

http://www.business.mcmaster.ca/msis/profs/wesolo/wesolo.htm