Getting a DIF Breakdown with Lertap

Larry R Nelson
Curtin University, Western Australia
Document date: 15 June 2019

This document shows how Lertap 5 may be used to look for differences among groups of test takers. Given two groups, say, for example, males and females, did one group out-perform the other, getting higher test scores?

Even if the groups appeared to have similar proficiency on the subject matter covered by the test, may there nonetheless have been group differences at the item level? Was there evidence of "DIF", differential item functioning?

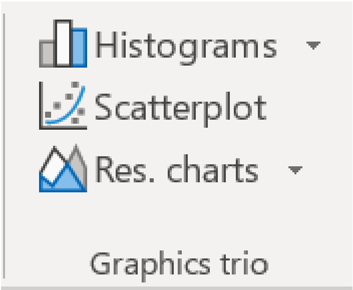
*I'll use a 49-item multiple-choice test to begin the examples, a professionally-developed instrument created by a large-scale test centre, showing how Lertap 5 may be used to answer questions such as these. Then the focus will shift to another example, a 56-item multiple choice test known to have a DIF item or two; you'll see how to "purify" the matching score and the resultant effect on DIF results. Finally, I'll also demonstrate how to use **SPSS** to get DIF results, and make mention of the capable, free, "**difR**" package.*

The discussion presupposes some familiarity with Lertap 5, an Excel-based system. More information about Lertap is available at this website.

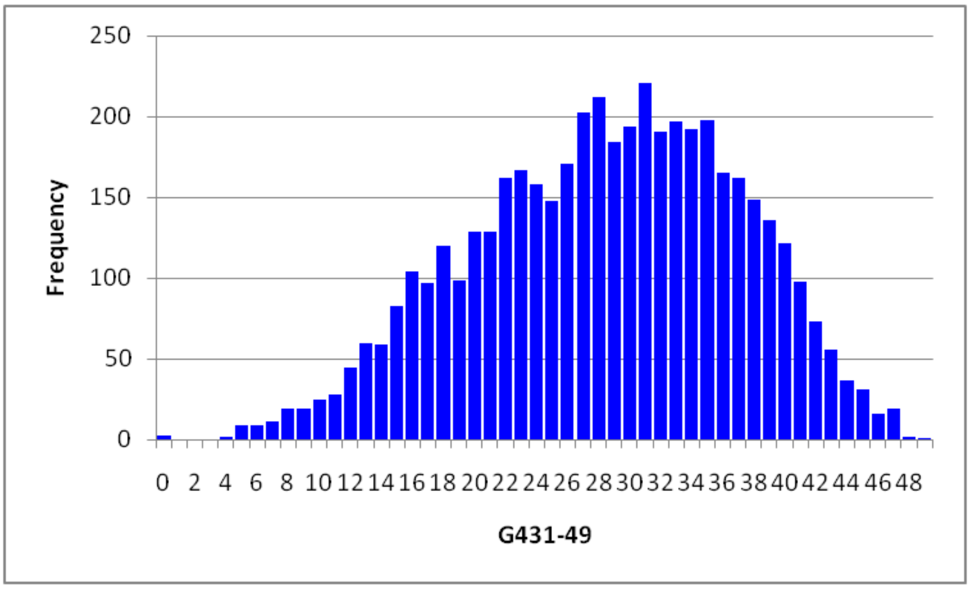
The data (copies may be downloaded from here)

More than four thousand high-school students sat a 60-item test. The test had ten trial items (not scored), and, as it unfortunately turned out, one item whose stem had a substantial error, not caught before printing. This item was also omitted from scoring, leaving 49 four-option multiple-choice items scored on a right / wrong basis, with one point for each correct answer.

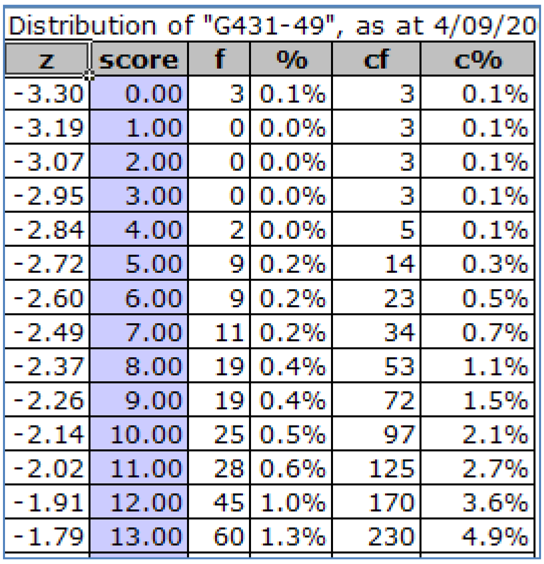
My first concern was to see how the distribution of test scores looked. I used Lertap to produce a couple of histograms, something I do while looking at the Scores worksheet, and clicking on the Histogram option found on the Lertap Excel ribbon.



The option produces a chart and a table.



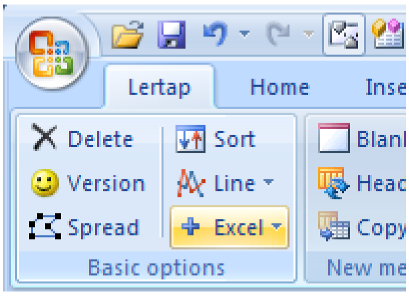
What I particularly wanted to investigate was the number of zero scores. There turned out to be three:



Distribution of "G431-49", as at 4/09/20

z	score	f	%	cf	c%
-3.30	0.00	3	0.1%	3	0.1%
-3.19	1.00	0	0.0%	3	0.1%
-3.07	2.00	0	0.0%	3	0.1%
-2.95	3.00	0	0.0%	3	0.1%
-2.84	4.00	2	0.0%	5	0.1%
-2.72	5.00	9	0.2%	14	0.3%
-2.60	6.00	9	0.2%	23	0.5%
-2.49	7.00	11	0.2%	34	0.7%
-2.37	8.00	19	0.4%	53	1.1%
-2.26	9.00	19	0.4%	72	1.5%
-2.14	10.00	25	0.5%	97	2.1%
-2.02	11.00	28	0.6%	125	2.7%
-1.91	12.00	45	1.0%	170	3.6%
-1.79	13.00	60	1.3%	230	4.9%

To find the Data records corresponding to zero scores, I used Lertap's "Sort" option, found in the collection of icons grouped under the "Basic options" section of Lertap's Excel ribbon tab[1]:



Records 1024, 1028, and 1067 had missing data for each of the 49 test items. I deleted them from the Data worksheet, and started again (that is, once again ran

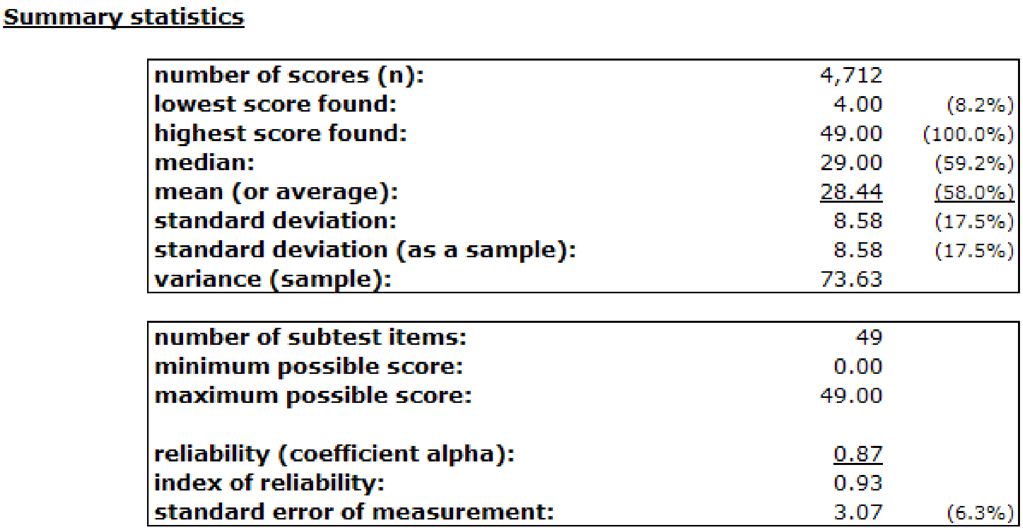
[1] This was done when using Excel 2007, now an old version.

Lertap's "Interpret" and "Elmillon" options).

The test

I looked at overall test quality by first going to the Stats1f report, then the bottom of the Stats1b report, then the Stats1ul report, and then clicked on the "Res. Charts" option to generate quintile plots.

This output is found towards the bottom of the Stats1f report:

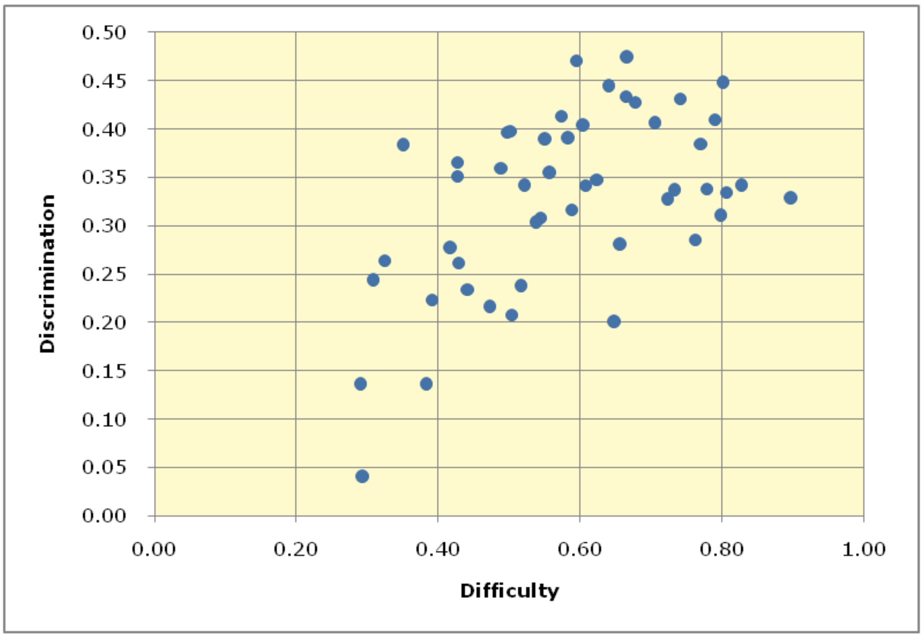


Summary statistics

number of scores (n):	4,712	
lowest score found:	4.00	(8.2%)
highest score found:	49.00	(100.0%)
median:	29.00	(59.2%)
mean (or average):	28.44	(58.0%)
standard deviation:	8.58	(17.5%)
standard deviation (as a sample):	8.58	(17.5%)
variance (sample):	73.63	

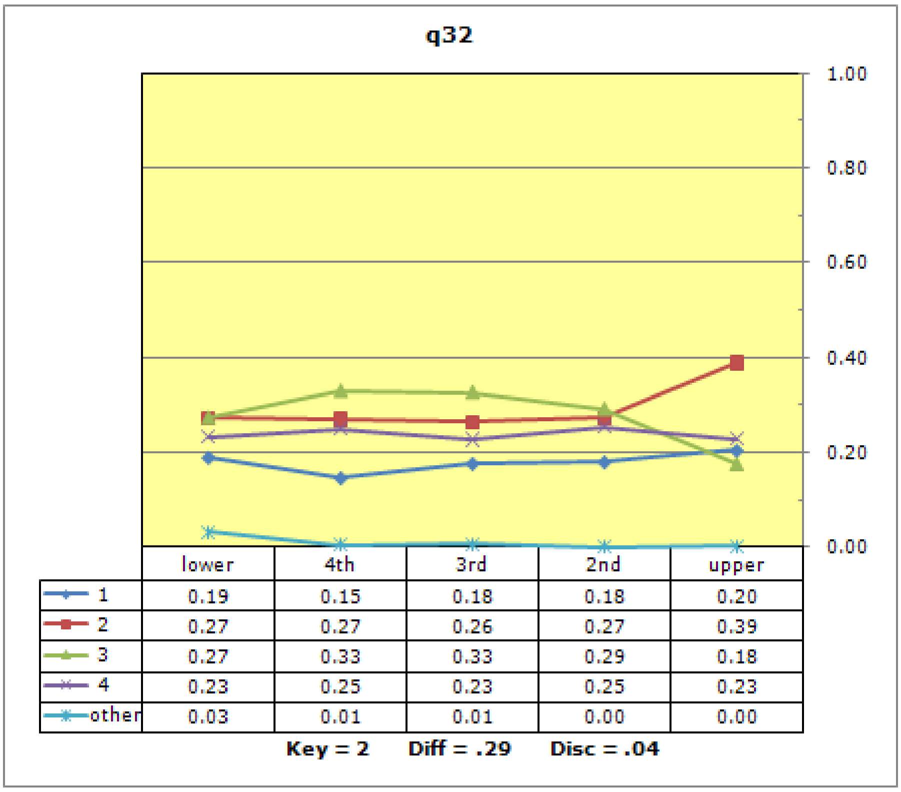
number of subtest items:	49	
minimum possible score:	0.00	
maximum possible score:	49.00	
reliability (coefficient alpha):	0.87	
index of reliability:	0.93	
standard error of measurement:	3.07	(6.3%)

This scatterplot is found at the bottom of a Stats1b report:



This wasn't really my test, but it seemed to have reasonable quality – coefficient alpha was comfortable at 0.87, and the scatterplot of item discrimination and difficulty suggests that only one item was sort of an "odd-man out". Here I refer to the only item having a discrimination value less than 0.10, which turned out to be item "q32".

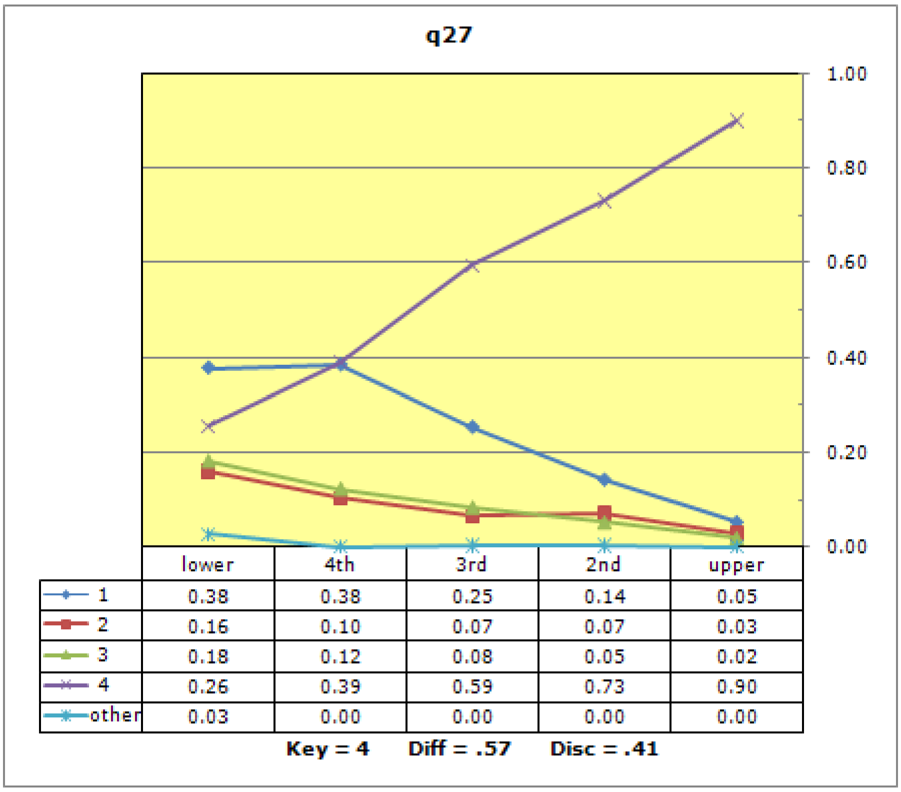
Two of the "quintile plots" are shown below:



	lower	4th	3rd	2nd	upper
1	0.19	0.15	0.18	0.18	0.20
2	0.27	0.27	0.26	0.27	0.39
3	0.27	0.33	0.33	0.29	0.18
4	0.23	0.25	0.23	0.25	0.23
other	0.03	0.01	0.01	0.00	0.00

Key = 2 Diff = .29 Disc = .04

Lertap’s standard quintile plot for q32 (above) clearly indicates a weak test item. For example, compare its plot to that for a nicely-performing item, q27:



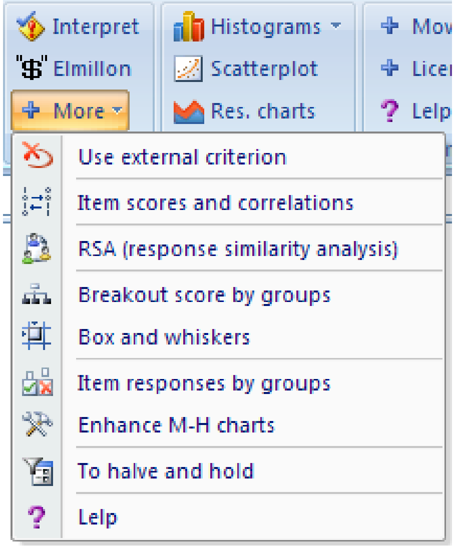
	lower	4th	3rd	2nd	upper
1	0.38	0.38	0.25	0.14	0.05
2	0.16	0.10	0.07	0.07	0.03
3	0.18	0.12	0.08	0.05	0.02
4	0.26	0.39	0.59	0.73	0.90
other	0.03	0.00	0.00	0.00	0.00

Key = 4 Diff = .57 Disc = .41

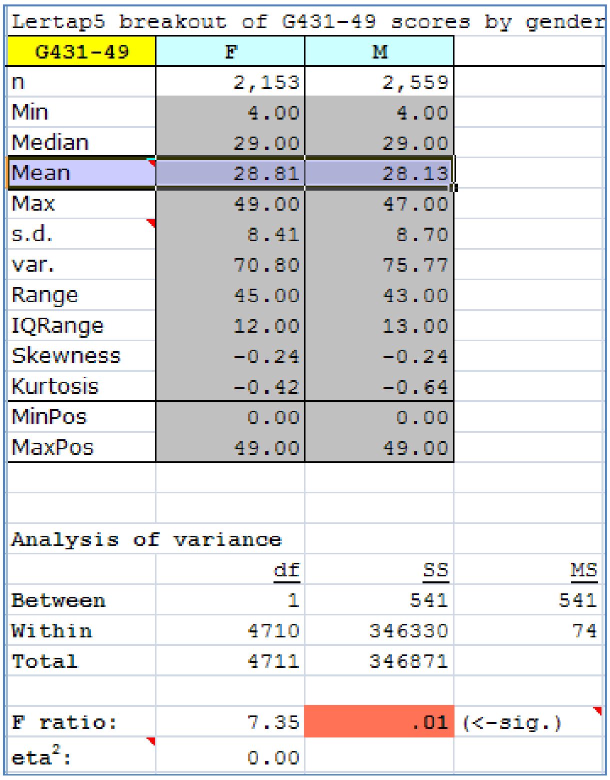
There’s more about using quintile plots here.

The groups

Gender information was coded in column 4 of the Data worksheet, in this case with "M" for males, and "F" for females. (Group codes must start with a letter; see the "Notes" in red ink close to the top of this page, or the "Caveat" near the top of this page.)



Were there gender differences on the test? I used the "Breakout score by groups" option as one means of answering the question. This option prompted Lertap to create two new reports, or worksheets. I copied the following from the "Breaks1" report:



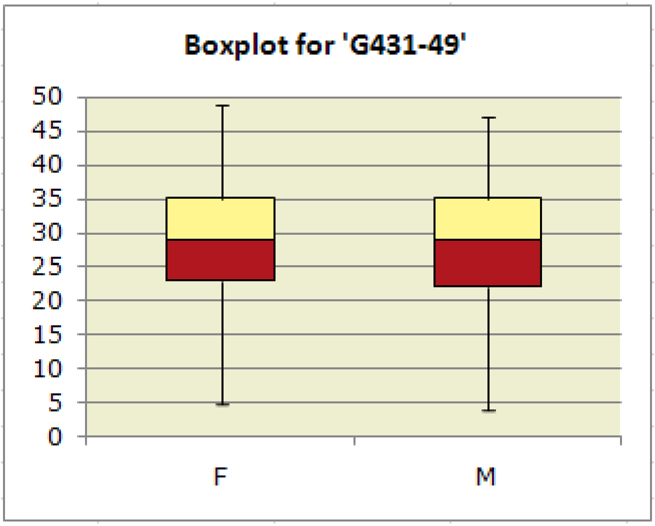
Lertap5 breakout of G431-49 scores by gender

G431-49	F	M	
n	2,153	2,559	
Min	4.00	4.00	
Median	29.00	29.00	
Mean	28.81	28.13	
Max	49.00	47.00	
s.d.	8.41	8.70	
var.	70.80	75.77	
Range	45.00	43.00	
IQRange	12.00	13.00	
Skewness	-0.24	-0.24	
Kurtosis	-0.42	-0.64	
MinPos	0.00	0.00	
MaxPos	49.00	49.00	
Analysis of variance			
	df	SS	MS
Between	1	541	541
Within	4710	346330	74
Total	4711	346871	
F ratio:	7.35	.01	(<-sig.)
eta^2:	0.00		

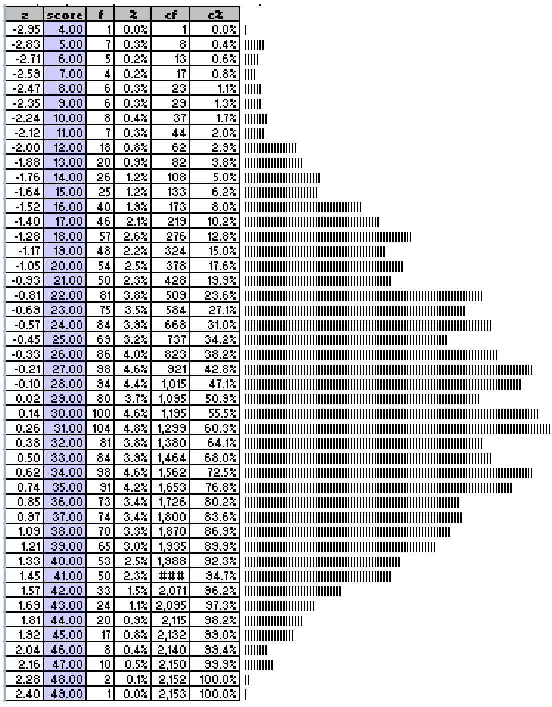
Not surprisingly, there was a statistically significant result; were I to hypothesise that the population means were equal, I'd reject the hypothesis at the .01 level. But this is not my interest. With more than 2,000 people in each sample, statistical significance would not be my focus, not at all.

The eta^2 index of effect size, a measure of practical significance, indicates that the difference in sample means was nothing to crow about (that is to say, was not at all substantial, hardly even measurable from a practical standpoint).

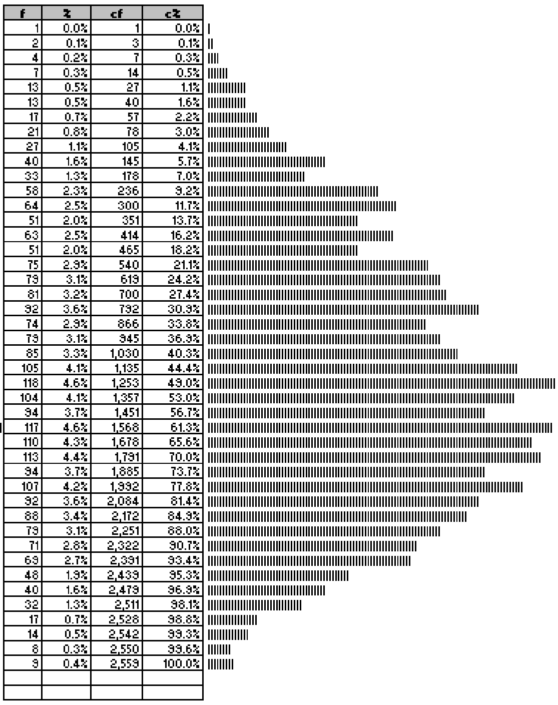
Next, I took the "Box and whiskers" option from the Run menu– I wanted a picture of the group results. I found that the boxes and whiskers were about the same for F and M:



I like Lertap's simple histograms. I went back to the Breaks1 report, and then clicked on Lertap's "Histograms" option. Twice I clicked, once for the girls' results, once for the boys'. With a little cutting and pasting, and a scale reduction, I made the following, females on the left, males on the right. The results, as presented below, are hard to read, but I think the gestalt is visible (note that low scores are at the top).



z	score	f	%	cf	c%
-2.95	4.00	1	0.0%	1	0.0%
-2.83	5.00	7	0.3%	8	0.4%
-2.71	6.00	5	0.2%	13	0.6%
-2.59	7.00	4	0.2%	17	0.8%
-2.47	8.00	6	0.3%	23	1.1%
-2.35	9.00	6	0.3%	29	1.3%
-2.24	10.00	8	0.4%	37	1.7%
-2.12	11.00	7	0.3%	44	2.0%
-2.00	12.00	18	0.8%	62	2.9%
-1.88	13.00	20	0.9%	82	3.8%
-1.76	14.00	26	1.2%	108	5.0%
-1.64	15.00	25	1.2%	133	6.2%
-1.52	16.00	40	1.9%	173	8.0%
-1.40	17.00	46	2.1%	219	10.2%
-1.28	18.00	57	2.6%	276	12.8%
-1.17	19.00	48	2.2%	324	15.0%
-1.05	20.00	54	2.5%	378	17.6%
-0.93	21.00	50	2.3%	428	19.9%
-0.81	22.00	81	3.8%	509	23.6%
-0.69	23.00	75	3.5%	584	27.1%
-0.57	24.00	84	3.9%	668	31.0%
-0.45	25.00	69	3.2%	737	34.2%
-0.33	26.00	86	4.0%	823	38.2%
-0.21	27.00	98	4.6%	921	42.8%
-0.10	28.00	94	4.4%	1,015	47.1%
0.02	29.00	80	3.7%	1,095	50.9%
0.14	30.00	100	4.6%	1,195	55.5%
0.26	31.00	104	4.8%	1,299	60.3%
0.38	32.00	81	3.8%	1,380	64.1%
0.50	33.00	84	3.9%	1,464	68.0%
0.62	34.00	98	4.6%	1,562	72.5%
0.74	35.00	91	4.2%	1,653	76.8%
0.85	36.00	73	3.4%	1,726	80.2%
0.97	37.00	74	3.4%	1,800	83.6%
1.09	38.00	70	3.3%	1,870	86.9%
1.21	39.00	65	3.0%	1,935	89.9%
1.33	40.00	53	2.5%	1,988	92.3%
1.45	41.00	50	2.3%	###	94.7%
1.57	42.00	33	1.5%	2,071	96.2%
1.69	43.00	24	1.1%	2,095	97.3%
1.81	44.00	20	0.9%	2,115	98.2%
1.92	45.00	17	0.8%	2,132	99.0%
2.04	46.00	8	0.4%	2,140	99.4%
2.16	47.00	10	0.5%	2,150	99.9%
2.28	48.00	2	0.1%	2,152	100.0%
2.40	49.00	1	0.0%	2,153	100.0%

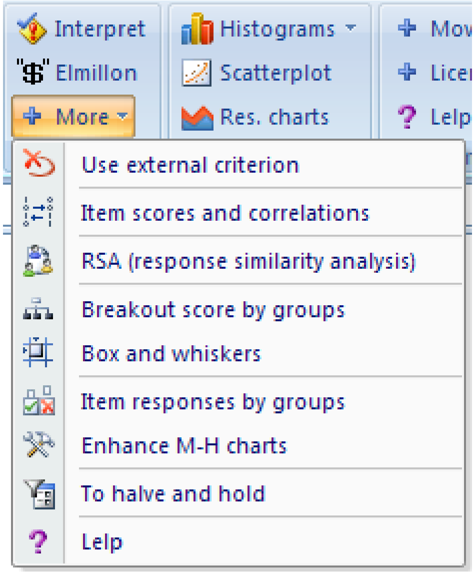


f	%	cf	c%
1	0.0%	1	0.0%
2	0.1%	3	0.1%
4	0.2%	7	0.3%
7	0.3%	14	0.5%
13	0.5%	27	1.1%
13	0.5%	40	1.6%
17	0.7%	57	2.2%
21	0.8%	78	3.0%
27	1.1%	105	4.1%
40	1.6%	145	5.7%
33	1.3%	178	7.0%
58	2.3%	236	9.2%
64	2.5%	300	11.7%
51	2.0%	351	13.7%
63	2.5%	414	16.2%
51	2.0%	465	18.2%
75	2.9%	540	21.1%
79	3.1%	619	24.2%
81	3.2%	700	27.4%
92	3.6%	792	30.9%
74	2.9%	866	33.8%
79	3.1%	945	36.9%
85	3.3%	1,030	40.3%
105	4.1%	1,135	44.4%
118	4.6%	1,253	49.0%
104	4.1%	1,357	53.0%
94	3.7%	1,451	56.7%
117	4.6%	1,568	61.3%
110	4.3%	1,678	65.6%
113	4.4%	1,791	70.0%
94	3.7%	1,885	73.7%
107	4.2%	1,992	77.8%
92	3.6%	2,084	81.4%
88	3.4%	2,172	84.9%
79	3.1%	2,251	88.0%
71	2.8%	2,322	90.7%
69	2.7%	2,391	93.4%
48	1.9%	2,439	95.3%
40	1.6%	2,479	96.9%
32	1.3%	2,511	98.1%
17	0.7%	2,528	98.8%
14	0.5%	2,542	99.3%
8	0.3%	2,550	99.6%
9	0.4%	2,559	100.0%

Were there gender differences? The mean scores are not exactly identical, but they're really close: expressed as a percentage, the girls' mean of 58.7% compares to 57.4% for the males, not much in it at all. There *were* gender differences in this sample, but they're very very minor. Very.

The groups at an item level (download all results here)

On the overall test there were no meaningful gender differences. But, what if I were to drill deeper, down to the level of results for individual items? Might there be gender differences there?

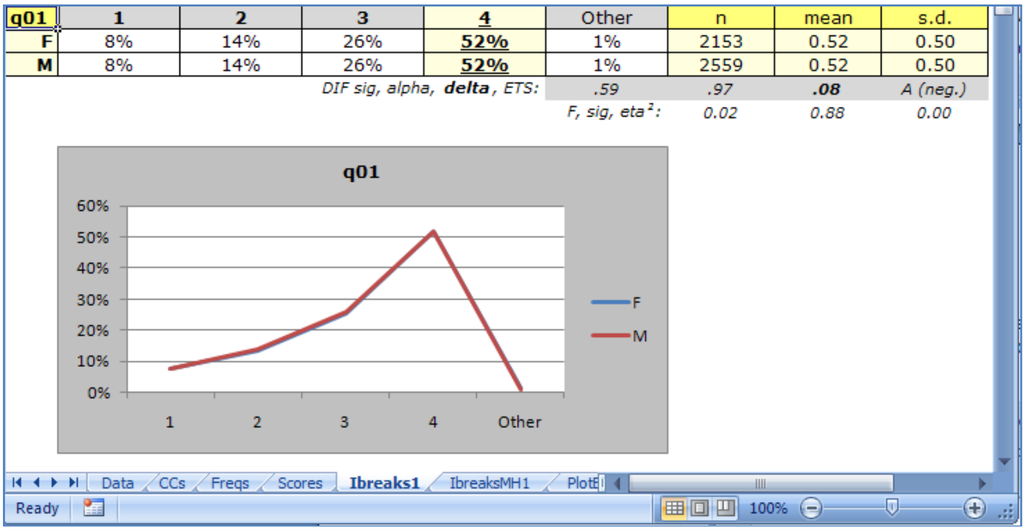


To answer this, I click on the "Item responses by groups" option.

When you try this option, you'll note that Lertap can present quite a number of questions, asking for instructions, especially when there are just two groups. This reflects Lertap's support for users who may be looking for "DIF", differential item functioning.

In this case I *was* interested in a DIF analysis, and I answered Lertap's questions accordingly.

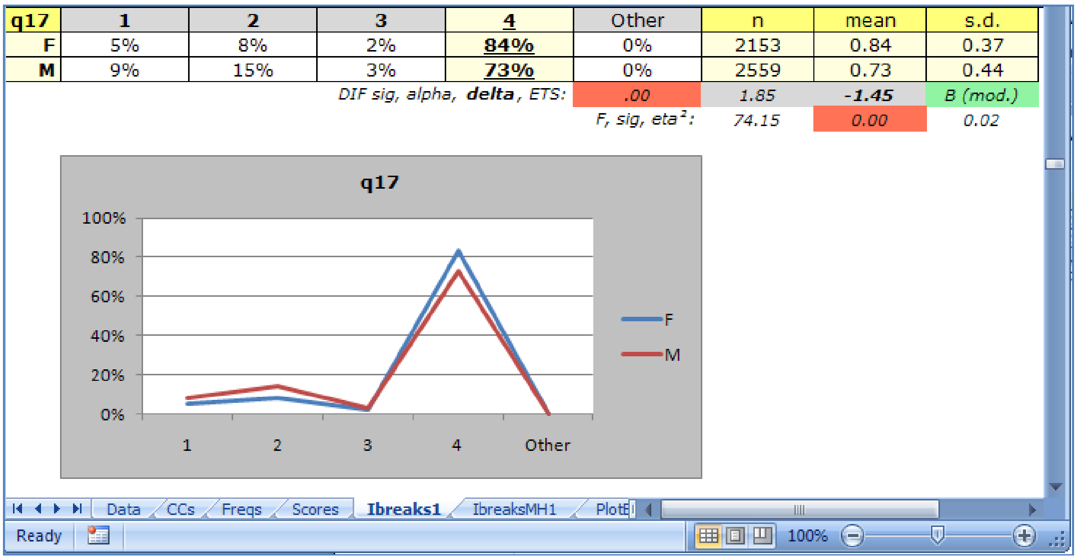
I get two new reports, "Ibreaks1", and "IbreaksMH1".



q01	1	2	3	4	Other	n	mean	s.d.
F	8%	14%	26%	**52%**	1%	2153	0.52	0.50
M	8%	14%	26%	**52%**	1%	2559	0.52	0.50
			DIF sig, alpha, ***delta***, *ETS:*		*.59*	*.97*	***.08***	*A (neg.)*
					F, sig, eta²:	*0.02*	*0.88*	*0.00*

"Ibreaks" has the small table and graph shown above. On the item labelled q01, there seem to be no differences at all between the two groups when results are looked at in this manner. The two trace lines, one for F, one for M, overlap each other, making it seem there's just a single trace (but see footnote 2).

The item labelled q17 displayed a different picture:



q17	1	2	3	4	Other	n	mean	s.d.
F	5%	8%	2%	**84%**	0%	2153	0.84	0.37
M	9%	15%	3%	**73%**	0%	2559	0.73	0.44
			DIF sig, alpha, ***delta***, *ETS:*		*.00*	*1.85*	***-1.45***	*B (mod.)*
					F, sig, eta²:	*74.15*	*0.00*	*0.02*

Now I've got a difference between group means, 0.84 versus 0.73. When it comes to item response percentages, this would be viewed as a fairly substantial difference, suggesting that there may indeed be something about this item which may, for some reason, favour the girls.

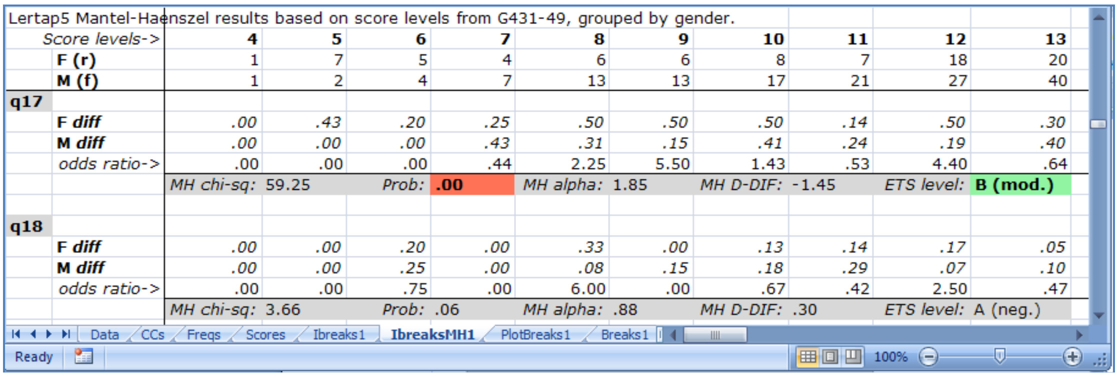
The two lines of statistics which appear below the table, with colours, are used to answer questions commonly found in data analysis of this sort.

The *F, sig, eta^2* results are entirely analogous to those seen earlier in the analysis of variance table of the "Breaks1" report. Are the differences between the sample means statistically significant? Yes, the F ratio of 74.15 is significant at the 0.00 level. But once again I tread lightly here; given the large sample sizes, even very small differences are likely to be statistically significant – what captivates my interest is the difference in percentage correct: 84% versus 73%.

If I want my test items to show no favour, to have equal challenge for each gender, how much difference will I tolerate?

To be honest, I don’t know. This is a question for the test developers, not me.

What I do know is that Lertap has more information for me to study. It turns out that the *B (mod.)* outcome seen above (in green) is a flag[2], waving away, urging me to have a look at the second report, “IbreaksMH1”. I turn to it.



Lertap5 Mantel-Haenszel results based on score levels from G431-49, grouped by gender.											
	Score levels->	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
	F (r)	1	7	5	4	6	6	8	7	18	20
	M (f)	1	2	4	7	13	13	17	21	27	40
q17											
	F diff	*.00*	*.43*	*.20*	*.25*	*.50*	*.50*	*.50*	*.14*	*.50*	*.30*
	M diff	*.00*	*.00*	*.00*	*.43*	*.31*	*.15*	*.41*	*.24*	*.19*	*.40*
	odds ratio->	.00	.00	.00	.44	2.25	5.50	1.43	.53	4.40	.64
		MH chi-sq: 59.25		*Prob:*	**.00**	*MH alpha:* 1.85		*MH D-DIF:* -1.45		*ETS level:*	**B (mod.)**
q18											
	F diff	*.00*	*.00*	*.20*	*.00*	*.33*	*.00*	*.13*	*.14*	*.17*	*.05*
	M diff	*.00*	*.00*	*.25*	*.00*	*.08*	*.15*	*.18*	*.29*	*.07*	*.10*
	odds ratio->	.00	.00	.75	.00	6.00	.00	.67	.42	2.50	.47
		MH chi-sq: 3.66		*Prob:*	.06	*MH alpha:* .88		*MH D-DIF:* .30		*ETS level:*	A (neg.)

The “MH” part of this report’s label stands for Mantel-Haenszel, named after authors who years ago developed a statistical method now frequently used for representing the extent to which the responses of two groups might differ. The method is based on the “odds ratio”, an index of how one group, say the “reference” group, might be favoured to get an item correct when compared to the second group, often called the “focal” group. An odds ratio greater than 1.00 indicates that the odds favour the reference group (they’re more likely to get the item right); a ratio less than 1.00 favours the focal group.

Lertap’s IbreaksMH tables of item results are broken into score levels, which are seen at the top of the report. These levels generally begin with the lowest score found, 4 in this case, and range out to the highest score found, which in this case was 49. In the tables showing above, I can see that one person in the F group had a score of 4, as did one person in the M group.

The ***diff*** figures (for relative difficulty) indicate the proportion of people in each group, at each score level, who correctly answered an item. For q17, 43% of the 7 female students with a score of 5 correctly answered the item; of the 2 males with a score of 5, none answered q17 correctly.

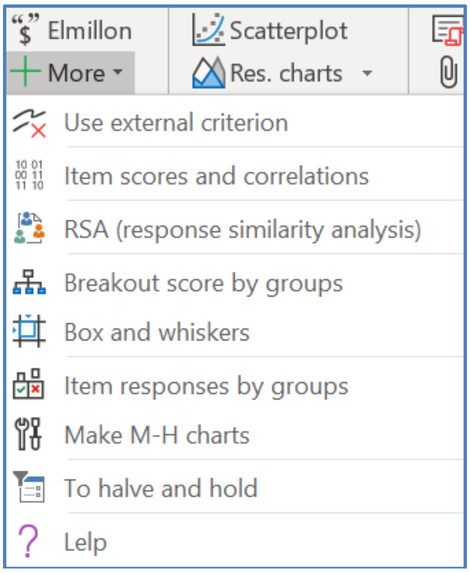
These tables are usually quite wide, with one column required to capture all the action at each score level. (Actual results may be seen in this download.)

The statistics which appear for each item, immediately below the row of odds ratios, are those often used in an M-H analysis. You’ll find more about them here.

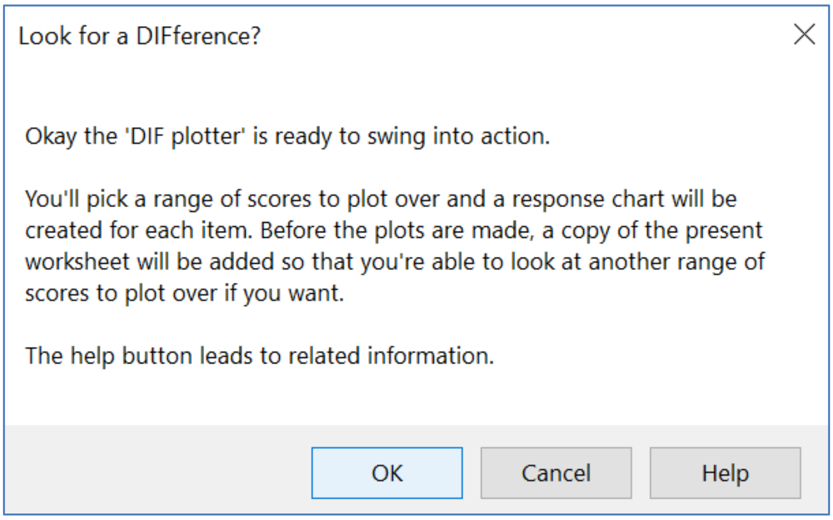
It is possible, and fairly easy, to get what amount to empirical DIF plots based on the statistics found in the IbreaksMH reports. They can be quite useful.

[2] *Note that this flag may appear even when there seems to be no difference in the plot lines* – in such cases the “DIF plots” mentioned below are likely to show why – there may indeed be differences at certain score levels.

The option which generates such plots is “Make M-H charts” shown below[3]:



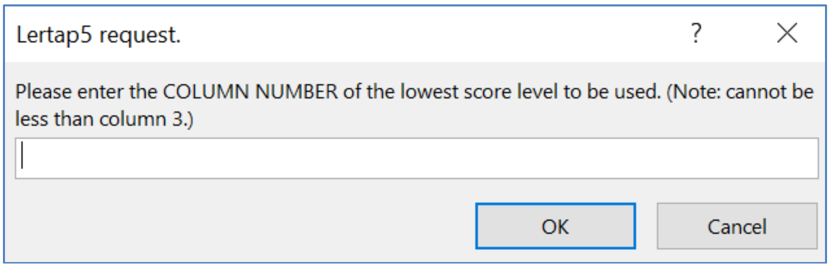
The following dialogue boxes appear when the option is taken:



Okay the 'DIF plotter' is ready to swing into action.

You'll pick a range of scores to plot over and a response chart will be created for each item. Before the plots are made, a copy of the present worksheet will be added so that you're able to look at another range of scores to plot over if you want.

The help button leads to related information.



Please enter the COLUMN NUMBER of the lowest score level to be used. (Note: cannot be less than column 3.)

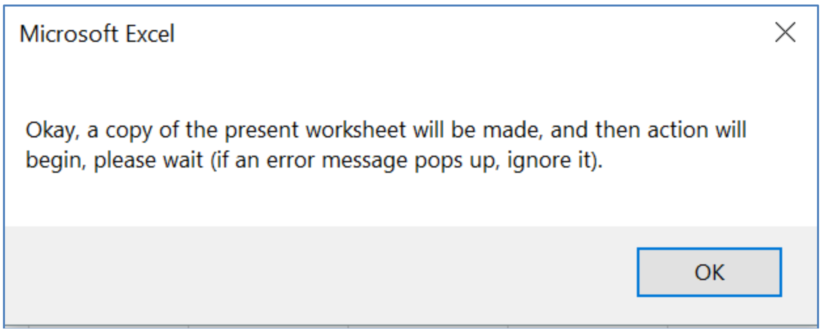
Experience will help answer this question, and the one which follows asking for the column number of the highest score to be used. It’s usually best to avoid the lowest scores, and also the highest ones – this is so as there will generally be a relatively small number of students at these score levels, and the “DIF Plots” will tend to look jagged.

[3] It may be labelled “Enhance M-H charts”, depending on the version of Lertap 5

In this case, the score histogram shown above indicates that the score tails are, say, below a score of 12, and above a score of 43. These score values will help answer the column number questions.

Using the top rows of the IbreaksMH report, the score level of 12 is found in column number 11, and the score level of 43 is found in column 42.

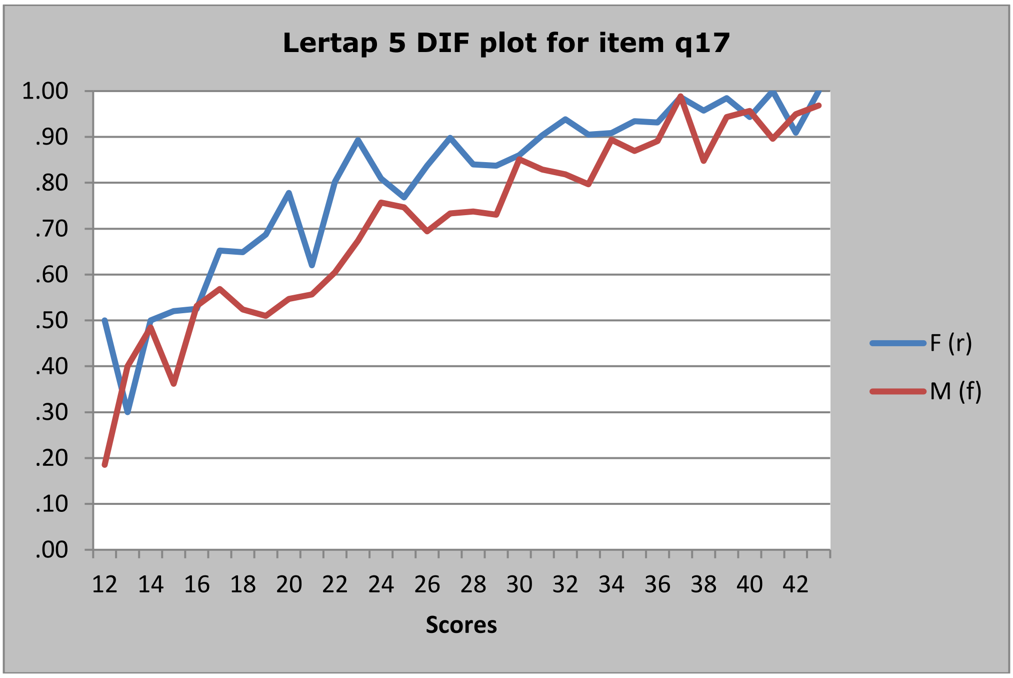
Once the two column number questions have been answered, Lertap gets Excel to present the following information:



Okay, a copy of the present worksheet will be made, and then action will begin, please wait (if an error message pops up, ignore it).

The copy is made so that it will be possible to later plot over another range of test scores if wanted.

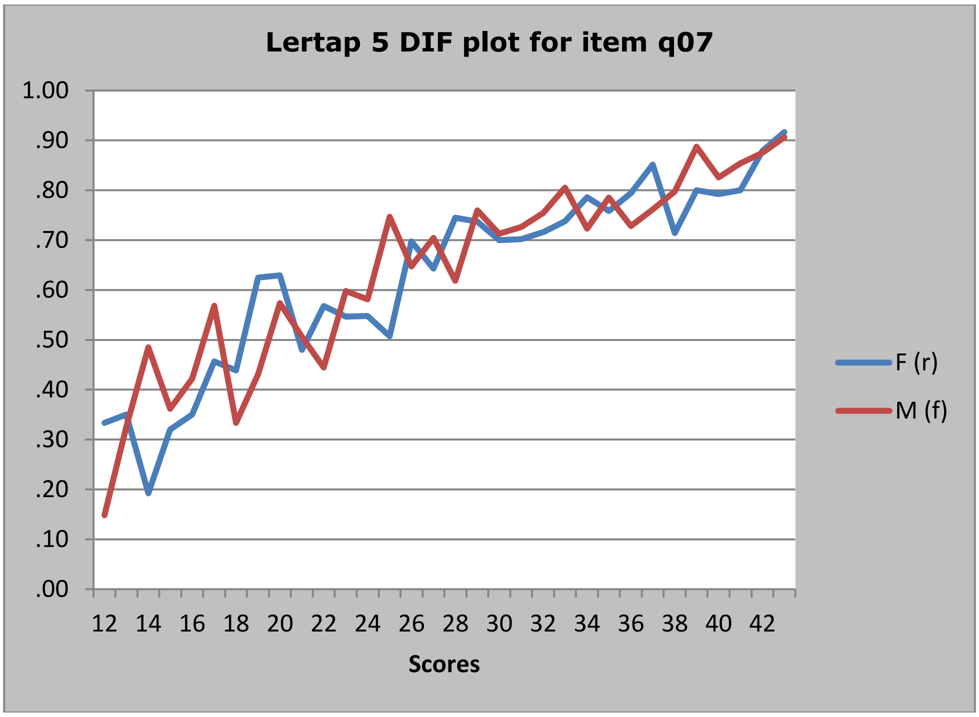
The "Make M-H charts" option will create a "DIF plot" for each and every test item, such as that for q17 seen here:



Now it's fairly easy to see that there's quite a range of score levels where the girls (F) outscored the boys (M) on q17. The difference isn't uniform in the score level range I've used here –there are two or three levels where the advantage went to the boys; these are at both ends of the plot. However, the number of cases in the extreme score levels is not great; the M-H statistic is a weighted average of an item's odds ratios, so those few levels where the boys were stronger, coming in relatively sparse regions of the score distribution, will be washed out when statistics are calculated, and they, the M-H statistics, will suggest what the plot confirms: something favoured the girls on q17.

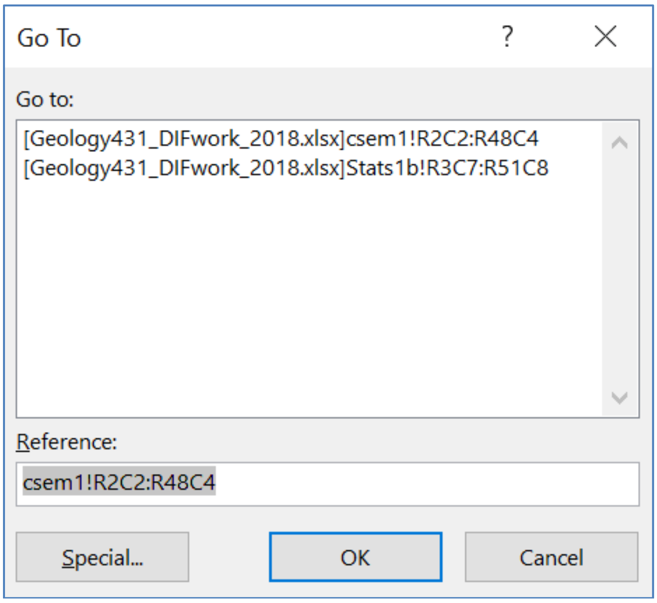
We don't know what it may have been, but test developers and test users who concern themselves with fairness may have something to ponder – q17 *might* be said to be "unfair" – there may be something to q17 which, for some reason, results in girls have a greater likelihood of identifying the item's correct answer.

The vast majority of items on this test had DIF plots indicating no difference, such as this one for q07:



It is possible to copy all of these "DIF plots" and paste them in a word processor, such as Word – follow the following steps.

Step (1) hold down the Control key (Ctrl) and simultaneously press the G key. This will result in a dialog box like this (it may very well contain other lines, even no lines, in the Go to section, and the Reference box may be empty):

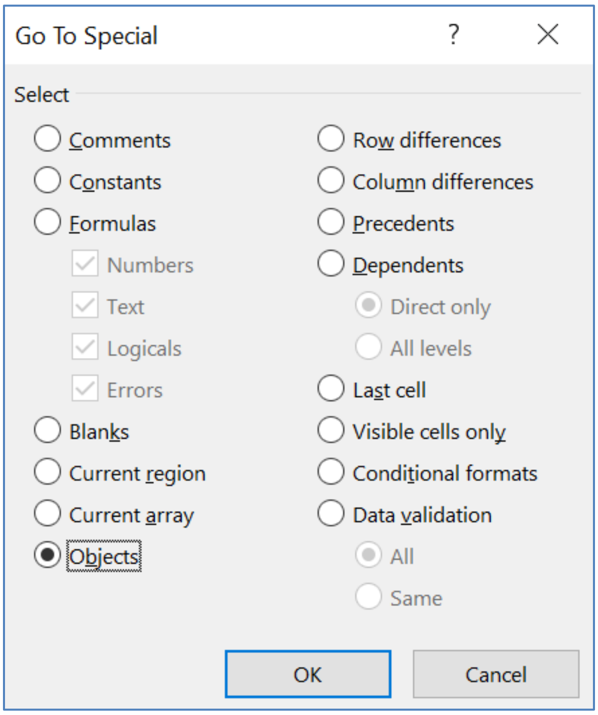


[Geology431_DIFwork_2018.xlsx]csem1!R2C2:R48C4
[Geology431_DIFwork_2018.xlsx]Stats1b!R3C7:R51C8

Reference:

csem1!R2C2:R48C4

Step (2): click the rectangular Special button in the lower left (above).



○ Comments
○ Constants
○ Formulas
☑ Numbers
☑ Text
☑ Logicals
☑ Errors
○ Blanks
○ Current region
○ Current array
● Objects
○ Row differences
○ Column differences
○ Precedents
○ Dependents
● Direct only
○ All levels
○ Last cell
○ Visible cells only
○ Conditional formats
○ Data validation
● All
○ Same

Step (3): select the round Objects button in the lower left (above).

Step (4): all of the charts will now be selected; press the copy button (or option) in Excel, or hold down the Ctrl key and press C.

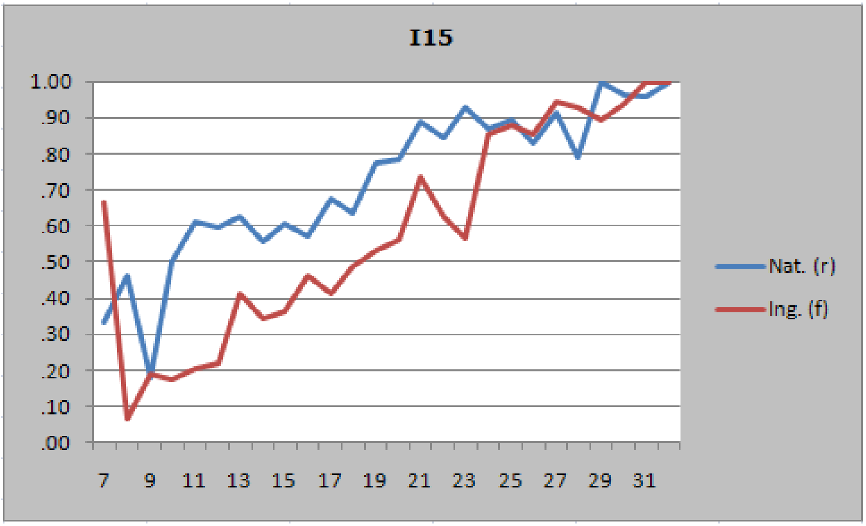
Step (5): go to the word processor and use the paste button (or option), or hold down the Ctrl key and press V.

Remember that Lertap made a copy of the IbreaksMH worksheet before making all of these DIF plots. The copy may be used to get new DIF plots based on another range of score levels.

Non-uniform DIF

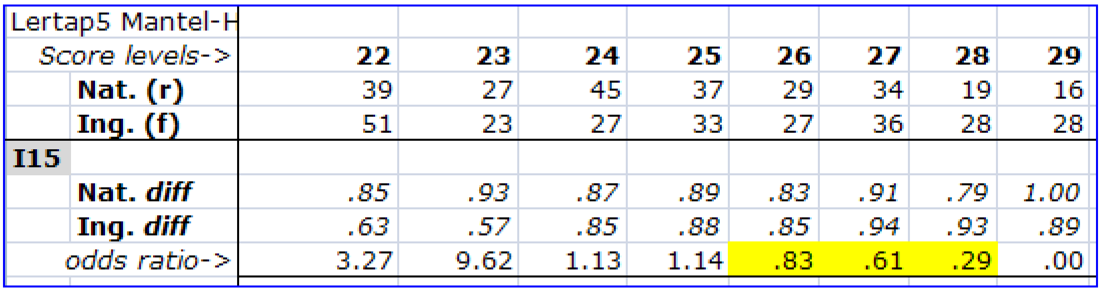
Lertap uses Mantel-Haenszel methods for its DIF analysis. While these methods generally receive good support in the literature, it is recognised that M-H procedures are not capable of uncovering non-uniform DIF, that is, the case where an item may favour one group in one area of the scores range, but, in another part of the scores range, favour the other group.

As an example, consider item "I15" from the test mentioned in Lertap help:



MH alpha for I15 was 2.39, with MH D-DIF at -2.04, an ETS level of C (large), indicating DIF in favour of the reference group, Nat.

For most of the scores range (seen along the x-axis), there seems to be something about I15 which works to favour the Nats. However, as shown in the small table below, there is an area of the range, corresponding to test scores of 25, 26, and 27, where the odds ratio is in favour of the focal group, Ing. And, there are quite a few students at these three score levels: 82 Nats and 91 Ings (respective total group sizes were 876 and 844, so we're talking about approximately 10% of each group; the odds ratio at a score level of 7 was .25, but there were few students with this score, only 3 in each group).



Lertap5 Mantel-H								
Score levels->	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**
Nat. (r)	39	27	45	37	29	34	19	16
Ing. (f)	51	23	27	33	27	36	28	28
I15								
Nat. diff	*.85*	*.93*	*.87*	*.89*	*.83*	*.91*	*.79*	*1.00*
Ing. diff	*.63*	*.57*	*.85*	*.88*	*.85*	*.94*	*.93*	*.89*
odds ratio->	3.27	9.62	1.13	1.14	.83	.61	.29	.00

I myself would not have singled out I15 as displaying non-uniform DIF, but a logistic regression program did, finding "moderate" DIF for I15, saying that it "exhibited non-uniform DIF too".

Purification of the matching score (example 2)[4]

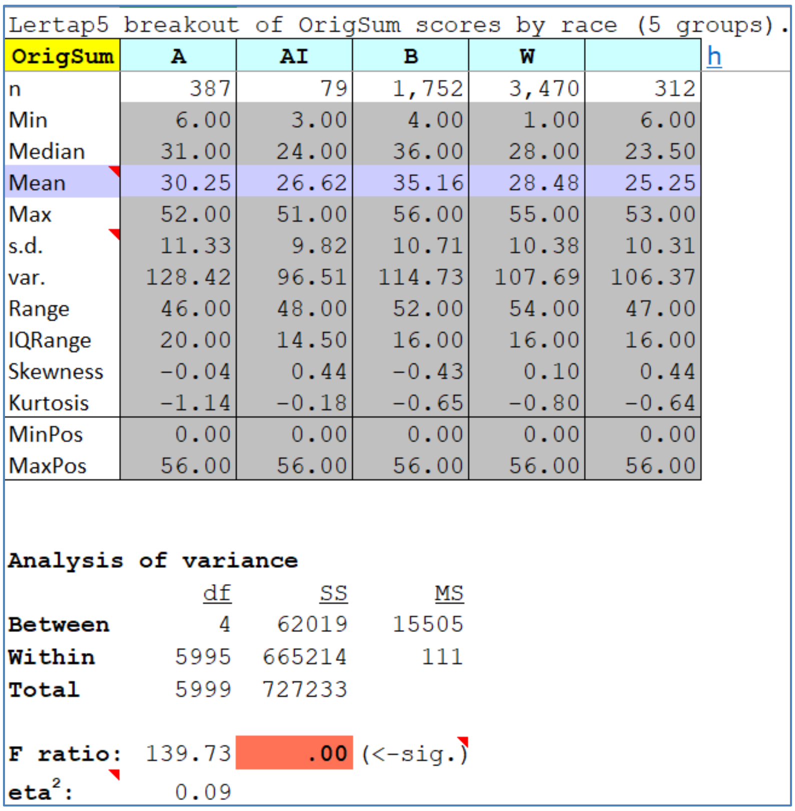
Here’s another example of using the Mantel-Haenszel approach in DIF analysis, based on the “exam1” test discussed in “Applied Measurement with jMetrik” (Meyer 2014). (Copies of the original exam1 data are available here. A link to my Excel workbook, created by importing the original data, is here.)

Numerous authors, Meyer among them, suggest that, when the matching score in a DIF study is a “sum score”, a total test score formed by adding up the number of correct answers on a test (such as the G431-49 score used above), M-H results might be adversely impacted if the score is based on the use of test items known to have DIF.

A recommended procedure in such cases is to form a new “sum score” which excludes those items previously found to exhibit DIF, and to then undertake a new M-H analysis using the new score.

Chapter 6 of Meyer’s 2014 text has a discussion of using the “race” variable in the exam1 results to look for possible DIF among the 56 test items.

Before asking Lertap to get DIF results for race, I’ll look at group differences using the “Breakout scores by groups” option, exactly as I did above when looking at gender differences on the G431-49 test.



Lertap5 breakout of OrigSum scores by race (5 groups).

OrigSum	A	AI	B	W		h
n	387	79	1,752	3,470	312	
Min	6.00	3.00	4.00	1.00	6.00	
Median	31.00	24.00	36.00	28.00	23.50	
Mean	30.25	26.62	35.16	28.48	25.25	
Max	52.00	51.00	56.00	55.00	53.00	
s.d.	11.33	9.82	10.71	10.38	10.31	
var.	128.42	96.51	114.73	107.69	106.37	
Range	46.00	48.00	52.00	54.00	47.00	
IQRange	20.00	14.50	16.00	16.00	16.00	
Skewness	-0.04	0.44	-0.43	0.10	0.44	
Kurtosis	-1.14	-0.18	-0.65	-0.80	-0.64	
MinPos	0.00	0.00	0.00	0.00	0.00	
MaxPos	56.00	56.00	56.00	56.00	56.00	

Analysis of variance

	df	SS	MS
Between	4	62019	15505
Within	5995	665214	111
Total	5999	727233	

F ratio:	139.73	.00	(<-sig.)
eta^2:	0.09		

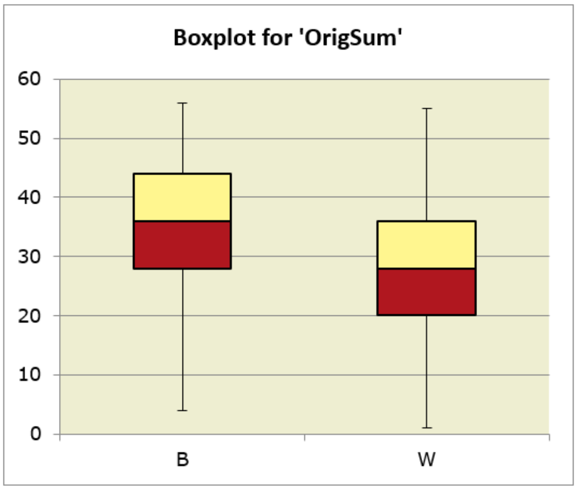
The exam1 dataset involves four race “levels”: Asian, American Indian, Black, and White. (312 of the Data records did not have an entry for race.)

Meyer (2014) discusses the application of the jMetrik program’s M-H option to look for possible black/white DIF, and does so without first commenting on the difference between mean test scores – they’re rather substantial: 35.16 for

[4] All results in this section have been corroborated by running jMetrik.

blacks, 28.48 for whites – as percentage-correct scores (out of 56 items): 63% correct for black students, 51% for white students.

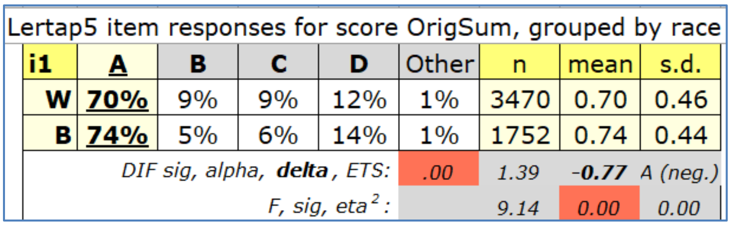
A boxplot from Lertap highlights the difference:



DIF investigations generally assume that the groups involved in a DIF study are samples from populations where group means are equal, or nearly so. We want to believe that the groups have near-equal proficiency levels in the subject area in which they're being tested. When this is not the case, looking for differences in response patterns for each item will be impacted to start with.

As we then go on to look at Lertap's DIF results, it will bode well to keep in mind that the B group in this example seems to have an edge to begin with, something that could well affect our interpretation of the results.

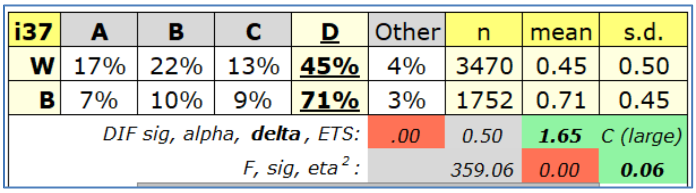
On all 56 items, black students had a higher percent-correct score in the "Ibreaks1" report made by taking Lertap's "Item responses by groups" option. An example is this one for item i1 where the B group's 74% bettered the 70% from the W group:



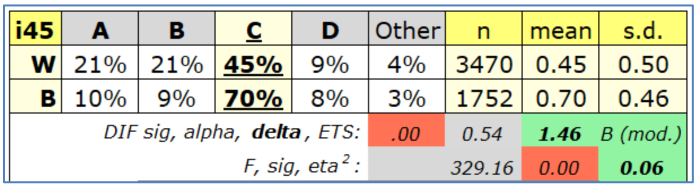
Lertap5 item responses for score OrigSum, grouped by race

i1	A	B	C	D	Other	n	mean	s.d.
W	**70%**	9%	9%	12%	1%	3470	0.70	0.46
B	**74%**	5%	6%	14%	1%	1752	0.74	0.44
				DIF sig, alpha, ***delta****, ETS:*	*.00*	*1.39*	***-0.77***	*A (neg.)*
				F, sig, eta^2:		*9.14*	*0.00*	*0.00*

The "A *(neg.)*" seen under the s.d. column indicates that i1 had negligible DIF as indexed by the ETS scale. A scan down these little tables reveals just two items whose ETS flag is other than A:

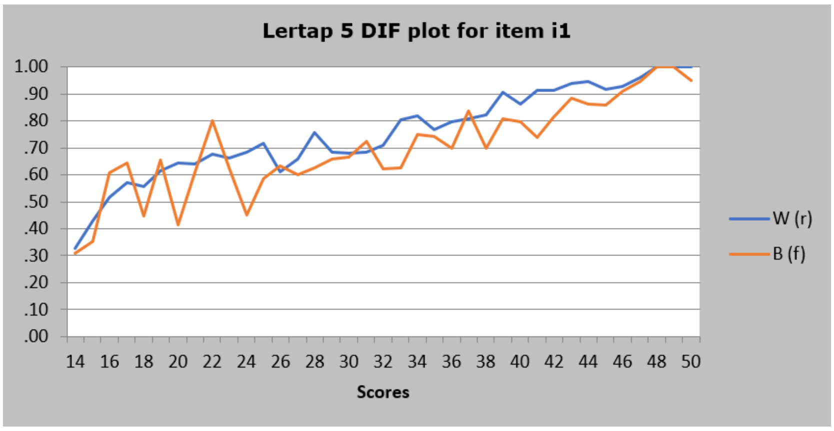


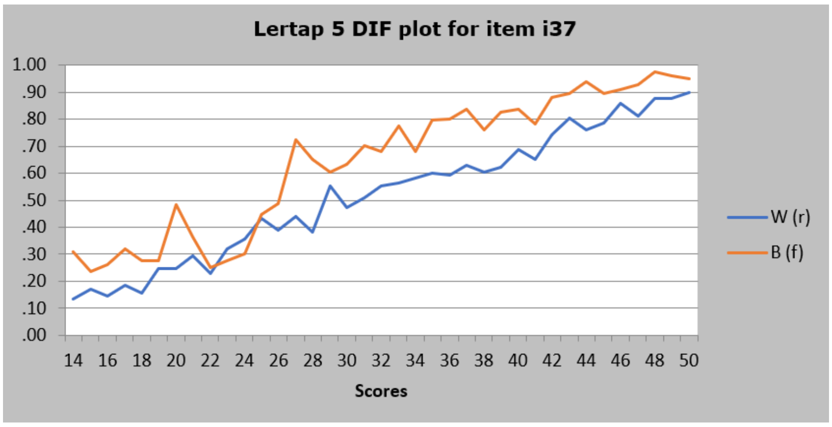
i37	A	B	C	D	Other	n	mean	s.d.
W	17%	22%	13%	**45%**	4%	3470	0.45	0.50
B	7%	10%	9%	**71%**	3%	1752	0.71	0.45
				DIF sig, alpha, ***delta****, ETS:*	*.00*	*0.50*	***1.65***	*C (large)*
				F, sig, eta^2:		*359.06*	*0.00*	***0.06***

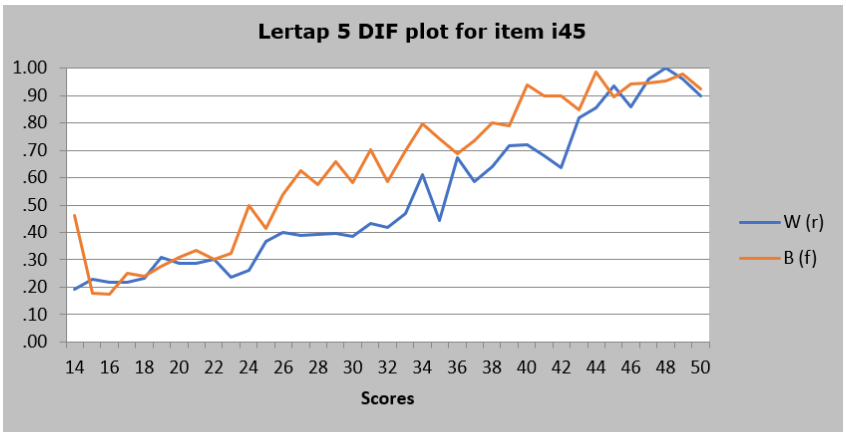


i45	A	B	C	D	Other	n	mean	s.d.
W	21%	21%	**45%**	9%	4%	3470	0.45	0.50
B	10%	9%	**70%**	8%	3%	1752	0.70	0.46
*DIF sig, alpha, **delta**, ETS:*					*.00*	*0.54*	***1.46***	*B (mod.)*
F, sig, eta^2:						*329.16*	*0.00*	***0.06***

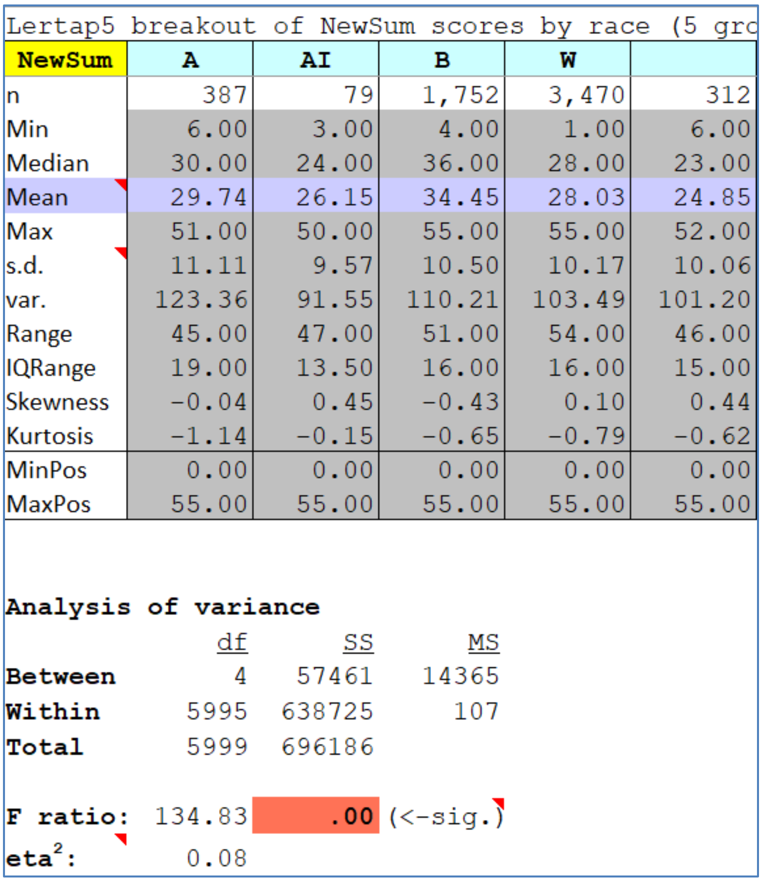
Corresponding DIF plots from the IbreaksMH1 report are seen below; in this the "Scores" along the x-axis are those formed by summing the number of correct answers over all 56 items:







Let's now "purify" the matching score ("OrigSum") by computing a new score ("NewSum") which leaves out i37, the item with DIF at the ETS C level. (We get Lertap to do this behind the scenes, and run "Breakout" score by groups", followed by "Item responses by groups".)



Lertap5 breakout of NewSum scores by race (5 grc

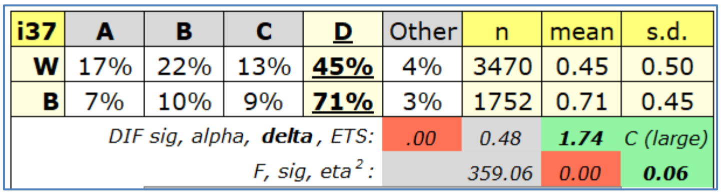
NewSum	A	AI	B	W	
n	387	79	1,752	3,470	312
Min	6.00	3.00	4.00	1.00	6.00
Median	30.00	24.00	36.00	28.00	23.00
Mean	29.74	26.15	34.45	28.03	24.85
Max	51.00	50.00	55.00	55.00	52.00
s.d.	11.11	9.57	10.50	10.17	10.06
var.	123.36	91.55	110.21	103.49	101.20
Range	45.00	47.00	51.00	54.00	46.00
IQRange	19.00	13.50	16.00	16.00	15.00
Skewness	-0.04	0.45	-0.43	0.10	0.44
Kurtosis	-1.14	-0.15	-0.65	-0.79	-0.62
MinPos	0.00	0.00	0.00	0.00	0.00
MaxPos	55.00	55.00	55.00	55.00	55.00

Analysis of variance

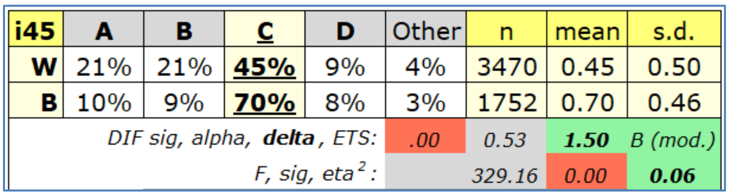
	df	SS	MS
Between	4	57461	14365
Within	5995	638725	107
Total	5999	696186	

F ratio: 134.83 .00 (<-sig.)

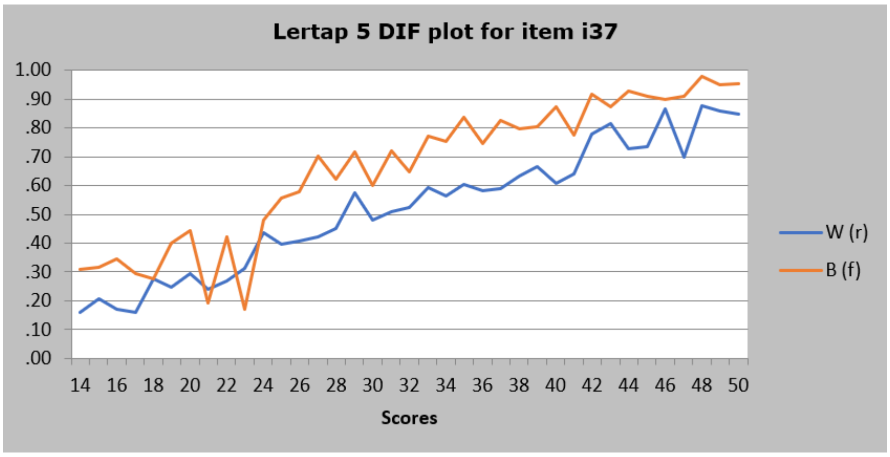
eta^2: 0.08

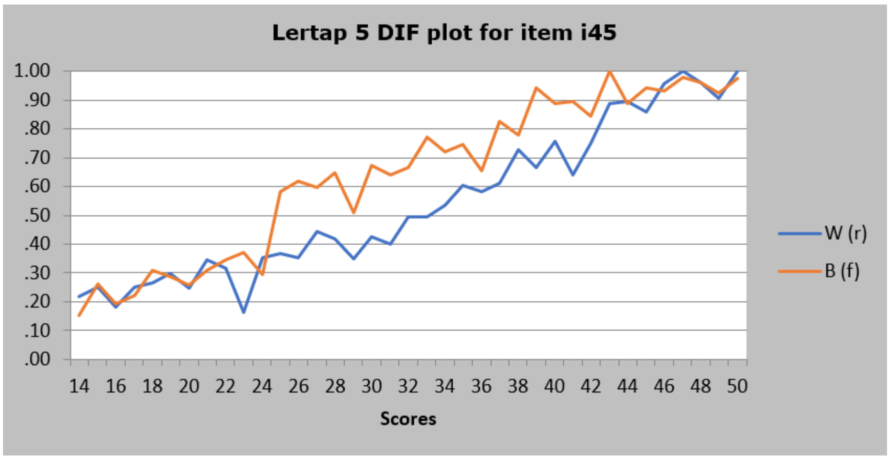


i37	A	B	C	D	Other	n	mean	s.d.
W	17%	22%	13%	**45%**	4%	3470	0.45	0.50
B	7%	10%	9%	**71%**	3%	1752	0.71	0.45
		DIF sig, alpha, ***delta****, ETS:*			*.00*	*0.48*	***1.74***	*C (large)*
			F, sig, eta^2:			*359.06*	*0.00*	***0.06***



i45	A	B	C	D	Other	n	mean	s.d.
W	21%	21%	**45%**	9%	4%	3470	0.45	0.50
B	10%	9%	**70%**	8%	3%	1752	0.70	0.46
		DIF sig, alpha, ***delta****, ETS:*			*.00*	*0.53*	***1.50***	*B (mod.)*
			F, sig, eta^2:			*329.16*	*0.00*	***0.06***





Not much changed. There were still just the two items with an ETS level above A. If you look closely, the alpha and delta figures in the tables are different to what they were before (but not by much), and the DIF plots haven't changed very much either.

This is in some contrast to what Meyer (2014, Chapter 6) found – his analyses with the jMetrik program found that the ETS classification for i37 remained at the C level (same as what we just found here), but i45's ETS classification went from B to C (different to what we've just found).

Closer examination of Lertap's figures shows that i45's delta value, seen immediately above as 1.50, has been rounded up from 1.49716 – had Lertap used the rounded value, 1.50, instead of the true internal value, 1.49716, then i45 would indeed been classified as ETS level C, in agreement with Meyer.

In this example, purifying the matching score had little effect. It may be due to having a relatively large number of test items (56) and a very small number of items showing DIF (2). Others have suggested that purification is at times of benefit, but, generally, such studies have had a higher number of DIF items. (Try searching for "purifying the matching variable in dif", or "purifying the matching score in dif".)

SPSS and DIF

It's possible to get **SPSS** to calculate Mantel-Haenszel statistics[5]. Lertap can be used to prepare data for SPSS (which of course is just as well as all my data are in Lertap to start with).

Here's how to go from Lertap to SPSS.

Starting with Lertap:

> Make sure the categorical variable in the Lertap Data sheet has a numeric code. If it was in a column which used {M, F} for codes, for example, use Lertap's recoder to map these to {1, 2}. The "Recode a Data column" is an option under the Move+ option under the 'Others menu'.
>
> Copy the categorical variable over to the Lertap Scores sheet. Use the "Copy a Data column to the Scores worksheet" option, also found under the Move+ menu.

[5] SAS will also produce MH statistics, as will jMetrik, as will numerous other programs. Lertap5, SAS, difR, and jMetrik have the advantage of being free for students.

Create a Lertap IStats worksheet by using the "Item scores and correlations" option found under the More+ option on the 'Run menu'. Then, from IStats, select the columns and rows with the 1.00 / 0.00 item scores (they always start in row 3), and copy them to columns on the Scores worksheet, to the right of the column with the categorical variable. Yes, this can be a big copy and paste when there are thousands of test takers.

Make sure you have the Excel referencing style set to "A1": this is simple to do – just use the "Ref. style" option found under "Excel" in the 'Basic options' menu, to the right of the "Spread" option. Click it, and the columns in the Scores worksheet should change from having numbers as labels to letters, starting, naturally, with the letter A.

Note the cells where the scores begin; most likely this will be A3. Note where they end; on a sample job I ran, with 40 items and 1,720 test takers, the scores ended in cell AQ1722. And note: this cell will **not** be at the bottom of the Scores worksheet, just in the row where the scores end. And another note: getting stuff from Scores into SPSS is not difficult, but it can be made just a bit easier if you delete the first Scores row so that "variable labels" come up to reside there, in the first row. If you don't want to fiddle with the Scores worksheet in this manner, use Excel to make a copy of it, and then delete the first row in the copy. Then, once again note the cell where the scores begin, and that where they end; A2 to AQ1721, for example.

Now, close the workbook. No need to exit Excel, but you do need to close the workbook to avoid a potential file sharing violation when getting into SPSS.

Going over to SPSS:

Use the File menu to / Open / Data

Tell SPSS you're looking for Files of type Excel (*.xls, *.xlsx, *.xlsm)

Find and Open the Lertap Excel workbook you just closed.

If you did delete the first Scores row, then leave a tick in the little box which says "Read variable names from the first row of data".

Pick out the worksheet in the workbook which has your scores. Put in the range, including the first row if you have the variable names in it (in my example, this would be A1:AQ1721).

That should do it. SPSS should now have a dataset with a tab for "Data View" and another for "Variable View".

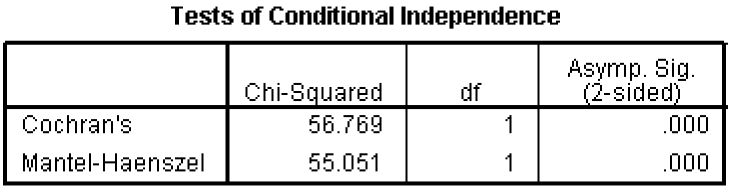
Save this SPSS dataset.

Still in SPSS, go to Analyze / Descriptive Statistics / Crosstabs

Put the categorical variable into the Row(s) box, all the items into the column(s) box, and the actual test score into the Layer 1 of 1 box. Under the Statistics option, tick "Cochran's and Mantel-Haenszel statistics".

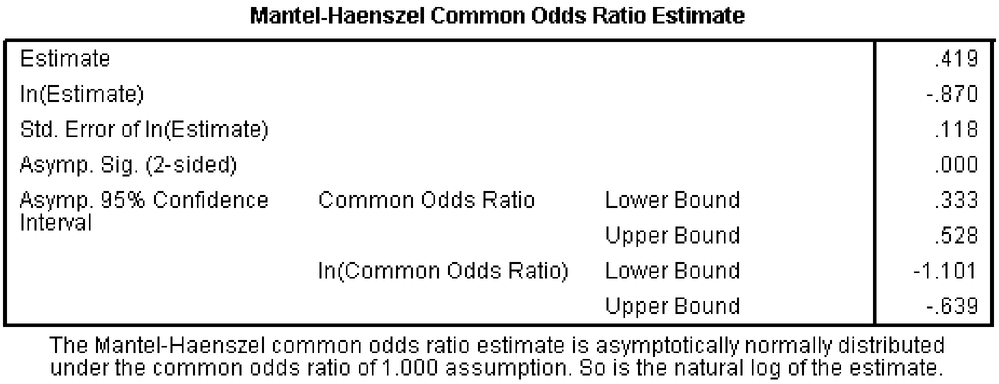
Click Continue / OK

SPSS grinds out lots of results. Here are its calculations for I15:



Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	56.769	1	.000
Mantel-Haenszel	55.051	1	.000

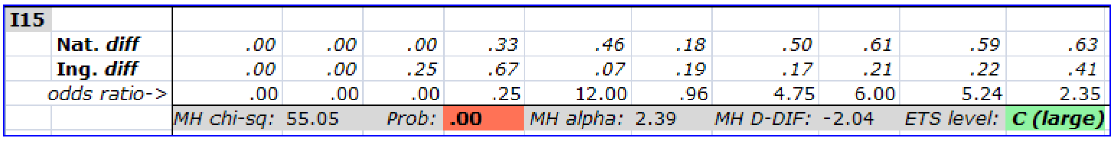


Mantel-Haenszel Common Odds Ratio Estimate

Estimate			.419
ln(Estimate)			-.870
Std. Error of ln(Estimate)			.118
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	.333
		Upper Bound	.528
	ln(Common Odds Ratio)	Lower Bound	-1.101
		Upper Bound	-.639

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

Compare these results with I15 data given in an IbreaksMH report from Lertap:



I15											
	Nat. ***diff***	*.00*	*.00*	*.00*	*.33*	*.46*	*.18*	*.50*	*.61*	*.59*	*.63*
	Ing. ***diff***	*.00*	*.00*	*.25*	*.67*	*.07*	*.19*	*.17*	*.21*	*.22*	*.41*
	odds ratio->	.00	.00	.00	.25	12.00	.96	4.75	6.00	5.24	2.35
		MH chi-sq: 55.05		*Prob:*	**.00**	*MH alpha:* 2.39		*MH D-DIF:* -2.04		*ETS level:*	***C (large)***

The Mantel-Haenszel chi-square values agree exactly (55.05), but the SPSS estimate of the Mantel-Haenszel common odds ratio is .419, compared to an MH alpha of 2.39 from IbreaksMH. What appears to be a major discrepancy has come about because I took little care when using Lertap to recode the categorical variable from Nat and Ing to numeric codes: I let Ing become 1, while Nat became 2. SPSS effectively takes the lowest code as the reference group, so now I need to divide SPSS' .419 into 1 (that is, 1.00 divided by .419) to reverse this coding error. What do I get? Agreement: 2.39[6].

difR package in CRAN

An R package called "difR" was available as at 21 January 2019. It was used as another cross-check on Lertap's dif statistics. difR allows users to look for possible differential item functioning using a variety of procedures, including Mantel-Haenszel, SIBTEST, and logistic regression. difR's M-H statistics were found to be 100% in agreement with Lertap's when the G431-49 test data were analyzed. Read about using difR with Lertap here.

This paper has to do with using Lertap in conjunction with R packages.

[6] Note that SPSS may not use the continuity correction for chi-square. Lertap 5 has a setting in row 53 of its System worksheet which turns the correction on or off.

Selected URLs for DIF

This paper by Stout (et al., 2003) has an excellent discussion of what constitutes DIF, and when. It's highly recommended, a fine read.

Mention of "SIBTEST" (the Simultaneous Item Bias Test) by NAEP in this website is also good, with mention of Mantel-Haenszel.

Larry R Nelson, PhD

School of Education
Curtin University, Western Australia

Questions welcomed. Please send to: l.nelson@curtin.edu.au