Iteman 4 and Lertap 5

Larry R Nelson
Curtin University (Australia)
Burapha University (Thailand)
www.lertap5.com
Last updated: 12 January 2022

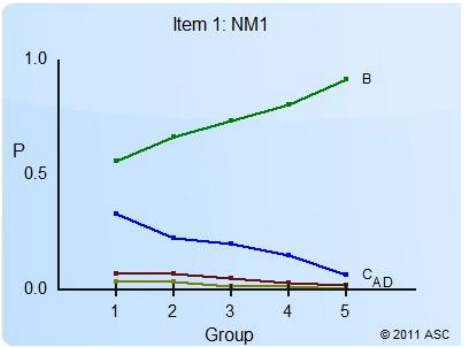
Version 4.5 of the venerable Iteman program was announced at the beginning of 2022. An email message announcing this new release said "*While the overall functionality has not been changed, the look and feel has been upgraded and the MS Word report given a modern new design.*"

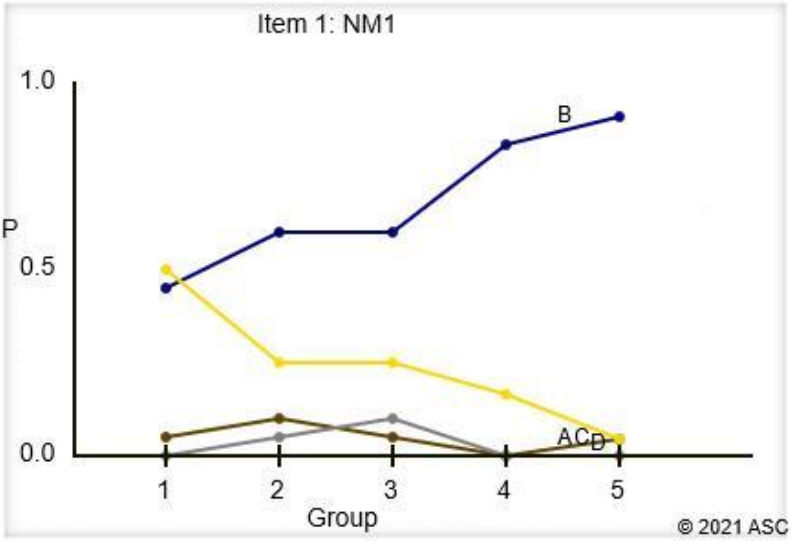
I downloaded a copy of the new version from here. It is available as a stand-alone "exe" app for Windows users (as was the previous version), and also offered as a "cloud" app. Of note was the version number in the downloaded demo version: 4.4, not the "4.5" mentioned in the promotional email message.

My original review of Iteman, as found below, involved version 4.3. My review of the new release, Version 4.4, easily confirmed that "*the overall functionality has not been changed*".

Now the report created when using the "exe" app for Windows is a standard Microsoft Word "docx" file, not the "rtf" file formerly created. This change results in a file type many users are likely to be more familiar with.

Report content is the same. One change I do like for sure concerns the graphics – whereas before they were small and had a blue background, now the background is gone and the graphics are larger, a welcome change.





While acknowledging that the overall functionality of Iteman has not changed, ASC has indicated that there are plans for new functionality. Readers may want to keep track of developments, or perhaps even provide input on what they might want to see in the next version (visit the website).

Meanwhile, Lertap5's functionality has not been standing still. Check out recent developments here.

The Original Paper (2019)

The purpose of this document is to point out some of the similarities and differences found in two classical test analysis software systems, Iteman 4, and Lertap 5.

Iteman 4 is available from ASC, Assessment Systems Corporation. Lertap 5, formerly distributed by ASC, is now available from Lertap5.com.

Lertap grew out of work initiated in 1972 when the Venezuelan Ministry of Education undertook a national assessment of educational progress in mathematics and the Spanish language. Program instructions were written in Spanish, and all of the program's output was also in Spanish.

An English-language version of Lertap emerged the following year in the United States. Versions for personal computers first appeared in 1980; in 2000 a version was created for use with Microsoft Excel – as this was the fifth "genus", this version was called "Lertap 5". Click here for a more extensive summary of Lertap's development.

Iteman's pedigree does not go back quite as far as Lertap's. It was first developed for use with personal computers in the late 1980s as part of a system called "MicroCAT". A Windows version was developed in the mid 1990s. The present version, known as "Iteman 4", has been in use since 2011.

Lertap 5 requires a version of Microsoft Excel in order to run. Under Windows, present versions of Lertap 5 work with Excel 2007, 2010, 2013, 2016, and 365. On Macintosh computers, the present version works with Excel 2016. Use this link to visit the downloads page.

Iteman 4 is distributed as a stand-alone Windows executable file. It requires that users have prepared a file with item-response data beforehand[1], as well as a file called the "item control file" with scoring information for each item.

A sample dataset

I selected "M.Nursing", a freely-available dataset from the internet, and ran it through both programs. Below I provide samples of the output created by each program, and discuss some of the differences between Iteman 4 and Lertap 5[2].

One of the greatest differences in the two programs relates to how they output information. Iteman 4's main output is in an RTF file, a rich-text file for viewing in a word processor, ready to print. Lertap 5's output is in an Excel workbook.

Iteman4's RTF file is generally many pages in length. In this case, with the "M.Nursing" data, Iteman 4 produced 68 pages of output, with the RTF file's size being just over 5MB (five megabytes). Here's a link to the file with all of Iteman4's output.

Lertap5's output, in this case, consisted of eight worksheets. They were titled "Freqs", "Scores", "Stats1f", "Stats1b", "Stats1ul", "csem1", "Stats1ulChta", and "Histo1". These were nested within the Excel workbook which contained the original item response data. The workbook's size was just under half a mega-byte. Here's a link to the complete Excel workbook after M.Nursing was processed with Lertap5.

Test scores

As would be expected, both programs produce test scores. Iteman4's output for M.Nursing was as follows:

Table 2: Summary statistics

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
Scored Items	60	40.608	7.710	2	60	0.677	0.246

[1] Most users of Lertap 5 will also have prepared their data beforehand; this is often done by using a mark-sense scanner, or an online testing system. But it's possible to enter response data directly into Lertap; teachers with small classes often enter their quiz results directly.

[2] I used Iteman version 4.3.0.3 and Lertap version 5.10.7.2

Figure 1: Total score for the scored items

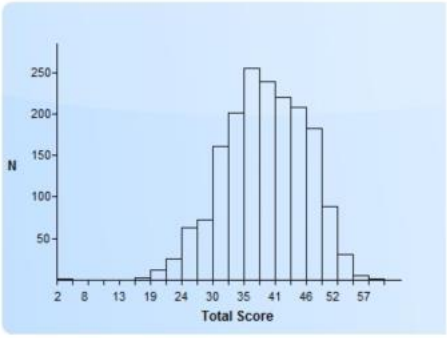


Table 5: Frequency Distribution for Total Score

Range	Frequency
1 to 4	1
5 to 7	0
8 to 10	0
11 to 13	0
14 to 16	0
17 to 19	3
20 to 22	12
23 to 25	26
26 to 28	63
29 to 30	72
31 to 33	161
34 to 36	201
37 to 39	255

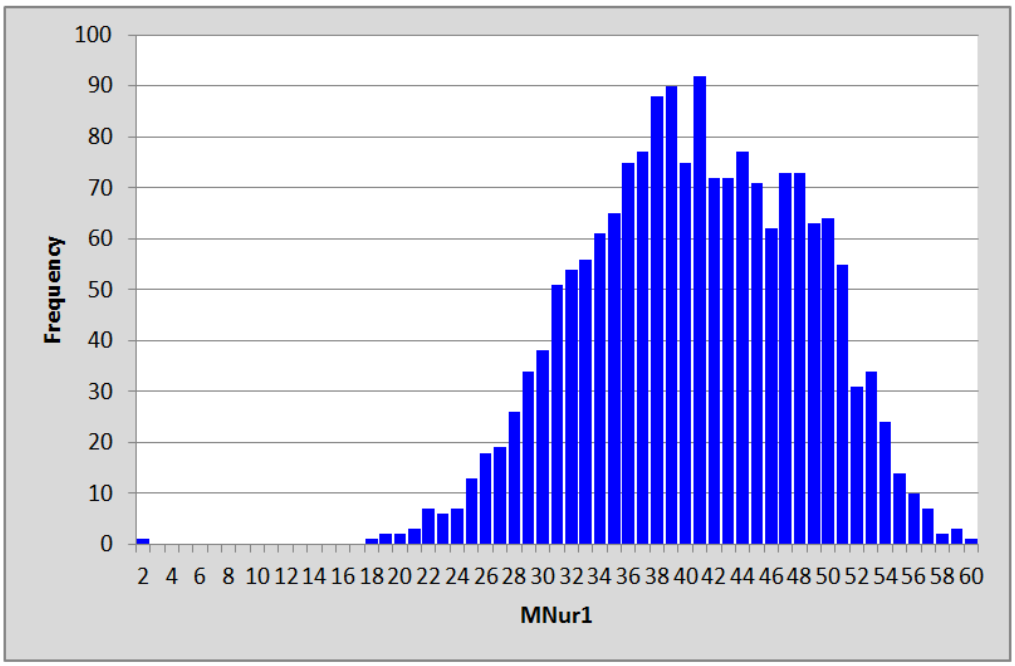
I have not copied all of Iteman4's "Table 5" in order to save space.

Lertap5 has information about test scores in four of its reports. Here are excerpts:

Summary statistics

number of scores (n):	1,769	
lowest score found:	2.00	(3.3%)
highest score found:	60.00	(100.0%)
median:	41.00	(68.3%)
mean (or average):	40.61	(67.7%)
standard deviation:	7.71	(12.8%)
standard deviation (as a sample):	7.71	(12.8%)
variance (sample):	59.44	

number of subtest items:	60	
minimum possible score:	0.00	
maximum possible score:	60.00	
reliability (coefficient alpha):	0.82	
index of reliability:	0.91	
standard error of measurement:	3.23	(5.4%)



z	score	f	%	cf	c%	h
-5.01	2.00	1	0.1%	1	0.1%	|
-4.88	3.00	0	0.0%	1	0.1%	
-4.75	4.00	0	0.0%	1	0.1%	
-4.62	5.00	0	0.0%	1	0.1%	
-4.49	6.00	0	0.0%	1	0.1%	
-4.36	7.00	0	0.0%	1	0.1%	
-4.23	8.00	0	0.0%	1	0.1%	
-4.10	9.00	0	0.0%	1	0.1%	
-3.97	10.00	0	0.0%	1	0.1%	
-3.84	11.00	0	0.0%	1	0.1%	
-3.71	12.00	0	0.0%	1	0.1%	
-3.58	13.00	0	0.0%	1	0.1%	
-3.45	14.00	0	0.0%	1	0.1%	
-3.32	15.00	0	0.0%	1	0.1%	
-3.19	16.00	0	0.0%	1	0.1%	
-3.06	17.00	0	0.0%	1	0.1%	
-2.93	18.00	1	0.1%	2	0.1%	|
-2.80	19.00	2	0.1%	4	0.2%	||
-2.67	20.00	2	0.1%	6	0.3%	||
-2.54	21.00	3	0.2%	9	0.5%	|||
-2.41	22.00	7	0.4%	16	0.9%	|||||||
-2.28	23.00	6	0.3%	22	1.2%	||||||

Median	41.00
Mean	40.61
Max	60.00
s.d.	7.71
var.	59.41
Range	58.00
IQRange	12.00
Skewness	-0.19
Kurtosis	-0.25
MinPos	0.00
MaxPos	60.00

In taking these screenshots, I have not re-sized those from Iteman4. In general, the graphs found in Iteman4 tend to be on the smaller side, often considerably more condensed than the counterparts found in Lertap5[3]. The graphics "engine" used in Iteman4 results in relatively weakly-formatted displays, of lower quality than the graphics capabilities of Excel used by Lertap5.

A close study of the test scores produced by both programs will reveal that there was an outlier, a very low score of just 2. One of the Lertap5 tables indicates that this outlier was about five standard deviations below the mean (z=-5.01).

In Lertap5, finding the data record corresponding to this score is a reasonably straightforward process. Not so in Iteman4, not at all – Iteman4 lacks a data editor. In my opinion, this is a significant limitation; I say this as basic data analysis requires that data be subject to careful screening beforehand in order to weed out data preparation and collection errors. Lertap5, an Excel "app", makes it easy to do this. In Iteman4 it's nigh impossible. (The M.Nursing webpage has more about this outlier; it turned out to be from a student who completed only a few of the test items and then was excused for medical reasons.)

Test reliability

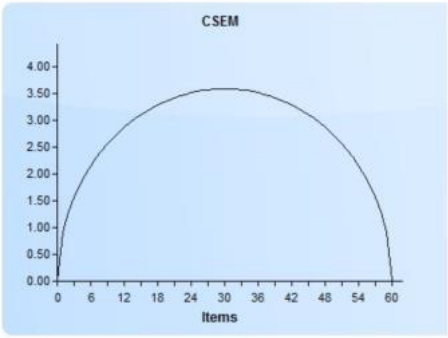
Both Iteman4 and Lertap5 produce statistics and graphs related to test reliability and estimates of measurement error.

The table and graph below are from Iteman4:

Table 3: Reliability

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0.825	3.228	0.692	0.601	0.701	0.818	0.751	0.824

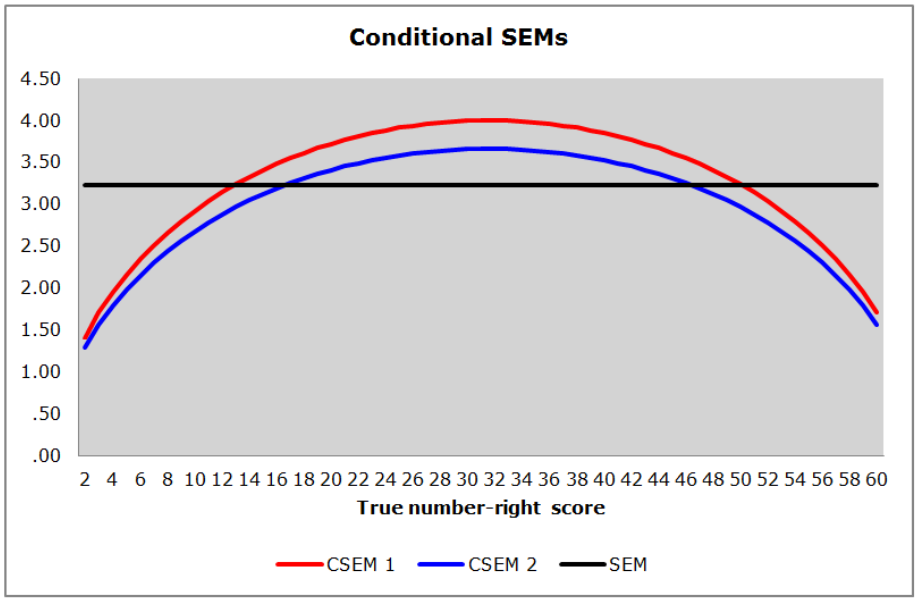
Figure 5: CSEM



[3] In comparing the two score histograms above, Iteman4 has used collapsed score intervals, producing a small graph. Lertap5 will not collapse intervals unless specifically directed to do so.

Lertap5 output was as follows:

number of subtest items:	60	
minimum possible score:	0.00	
maximum possible score:	60.00	
reliability (coefficient alpha):	0.82	
index of reliability:	0.91	
standard error of measurement:	3.23	(5.4%)



Conditional standard errors of measurement for "M.Nu

Score	CSEM 1	CSEM 2	SEM
2	1.40	1.28	3.23
3	1.70	1.56	3.23
4	1.95	1.79	3.23
5	2.16	1.98	3.23
6	2.35	2.15	3.23
7	2.51	2.30	3.23
8	2.66	2.44	3.23
9	2.80	2.56	3.23
10	2.92	2.68	3.23
11	3.04	2.78	3.23
12	3.14	2.88	3.23
13	3.24	2.97	3.23

Iteman4's standard reliability output, Table 3 above, includes more estimates of reliability than found in the standard output from Lertap5. However, Lertap5 is capable of producing the split-half figures output by Iteman4 (see this example), and one of the Lertap5 websites incudes an "S-B" calculator (click here).

Iteman4 outputs one of the two common CSEM estimates used in classical test theory; Lertap5 outputs both of them, and its graph incudes SEM as a reference line.

Overall item statistics

In classical test theory, item difficulty and item discrimination are the key measures of item quality. Iteman4 refers to item difficulty as the "P value", and to item discrimination as "Rpbis". When Lertap5 needs to use an abbreviation to refer to item difficulty, it uses "p" in its Stats1f report, and "U-L diff." in the Stats1ul report.

Both Iteman4 and Lertap5 allow multiple-choice items to have more than one correct answer. Iteman4 assumes that every correct answer will add one point to a student's score, Lertap5 does not – in Lertap5, it's possible to award partial points, such as half a point, to correct answers, in fact, in Lertap5, any item option may have any number of points, they can even be negative.

A slight error in Iteman4's output appears (in version 4.3.0.3) when a multiple-choice item has more than one correct answer. A point-biserial correlation, "Rpbis" in Iteman4, is not the correct estimate of item discrimination for items having more than one right answer. The correlation index normally used in this case is "Pearson's r", the Pearson product-moment correlation.

But, a little checking reveals that, in fact, Iteman4 (like Lertap5) does use Pearson's r as the discrimination figure for multiple-choice items with multiple scored options – it's just that Iteman4's output is incorrectly labelled – where it says "Rpbis" it should use the correct label, such as "r" or "R", in its little "Item statistics" table. And, another minor point with regard to Iteman4's output: the "Rbis" figure reported for items with more than one right answer would not be expected (it doesn't make sense).

The snapshots below show Iteman4 output for the M.Nursing test:

Figure 2: P values for the scored items

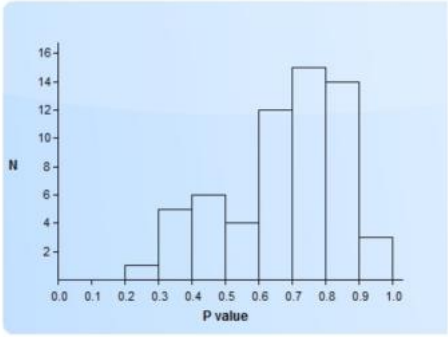


Table 6: Frequency Distribution for the P values

Score	Frequency
0.0 to 0.1	0
0.1 to 0.2	0
0.2 to 0.3	1
0.3 to 0.4	5
0.4 to 0.5	6
0.5 to 0.6	4
0.6 to 0.7	12
0.7 to 0.8	15
0.8 to 0.9	14
0.9 to 1.0	3

Figure 3: Rpbis for the scored items

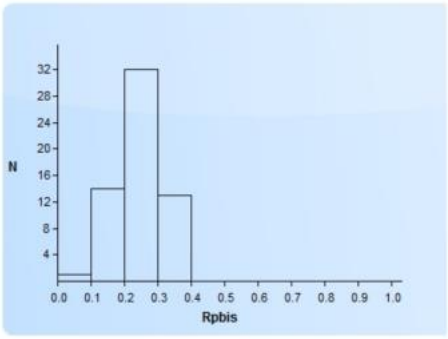
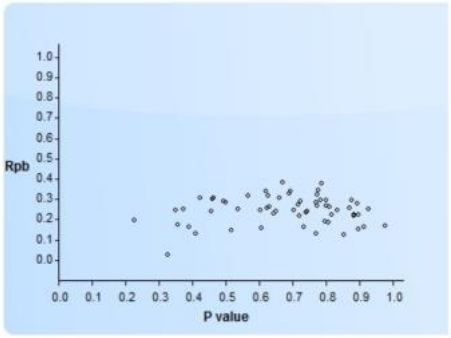


Table 7: Frequency Distribution for the Rpbis

Score	Frequency
0.0 to 0.1	1
0.1 to 0.2	14
0.2 to 0.3	32
0.3 to 0.4	13
0.4 to 0.5	0
0.5 to 0.6	0
0.6 to 0.7	0
0.7 to 0.8	0
0.8 to 0.9	0
0.9 to 1.0	0

Figure 4: P by Rpbis



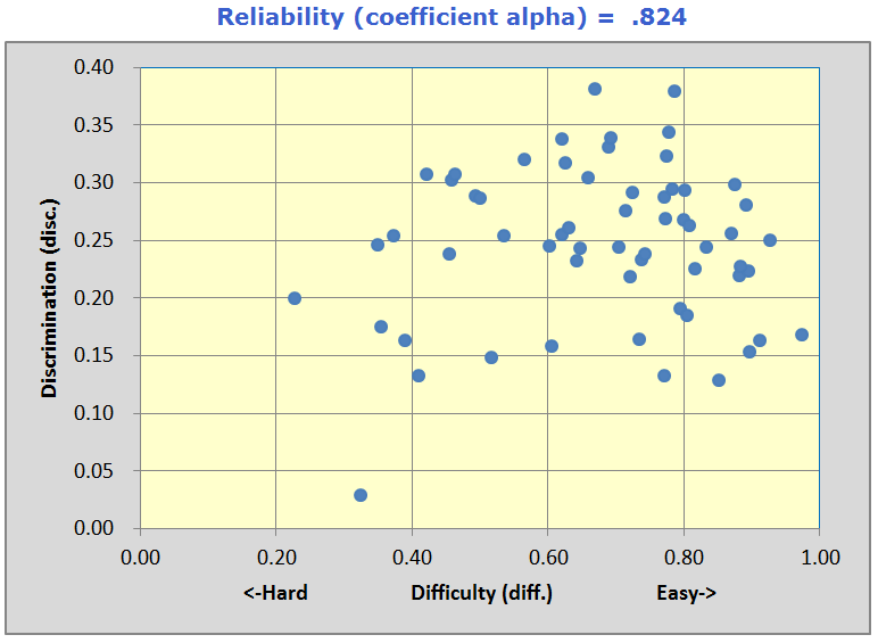
Lertap5 output for the M.Nursing test:

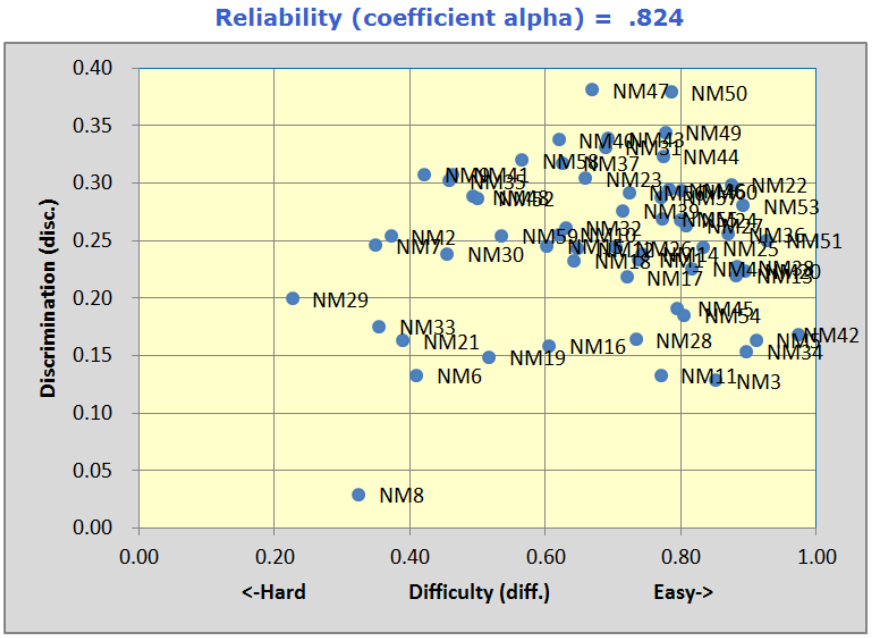
item difficulty bands

.00:
.10:
.20: NM29
.30: NM2 NM7 NM8 NM33
.40: NM6 NM9 NM21 NM30 NM35 NM41
.50: NM19 NM48 NM52 NM58 NM59
.60: NM10 NM12 NM15 NM16 NM18 NM23 NM31 NM32 NM37 NM40 NM47
.70: NM1 NM11 NM14 NM17 NM26 NM28 NM39 NM43 NM44 NM46 NM49 NM50 NM55 NM56 NM57
.80: NM3 NM4 NM13 NM22 NM24 NM25 NM27 NM36 NM38 NM45 NM54 NM60
.90: NM5 NM20 NM34 NM42 NM51 NM53

item discrimination bands

.00: NM8
.10: NM3 NM5 NM6 NM11 NM16 NM19 NM21 NM28 NM33 NM34 NM42 NM54
.20: NM1 NM2 NM4 NM7 NM10 NM12 NM13 NM14 NM15 NM17 NM18 NM20 NM24 NM25 NM26 NM2
.30: NM9 NM22 NM23 NM31 NM35 NM37 NM40 NM41 NM43 NM44 NM46 NM47 NM49 NM50 NM56
.40:
.50:
.60:
.70:
.80:
.90:





I think it would be safe to suggest that Lertap5's output makes it much easier to identify problematic items. This is because Lertap5 uses item labels in its reports. Above, for example, item "NM8" clearly stands out as an item which

performed poorly; the little "item difficulty bands" and "item discrimination bands" above also quickly point to potential problem areas.

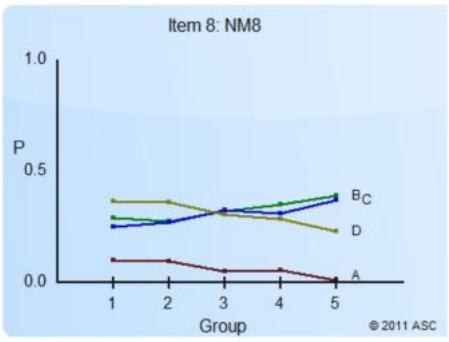
Note that there are two displays above for the M.Nursing difficulty by discrimination graphs. The second one has the item labels; the first does not. Lertap5 usually outputs just one of these two, not both – a toggle is used to switch the labels on and off, as wanted.

Item-level results

Both Iteman4 and Lertap5 will output a graph with response trace lines for each of an item's options, and tables of statistics. Iteman4 will always dedicate a minimum of one printed page for each item; Lertap5 will try to fit more than one item per page if it can.

I have based the sample output below on the M.Nursing item labelled "NM8". This was the worst performer of all items as it had the lowest discrimination. Generally, poorly-performing items have issues with their incorrect answers, that is, with their "distractors". Both programs use special flags to draw attention to distractor problems.

Iteman4's output for item NM8 includes a graph followed by four tables:



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
8	NM8	B	Yes	4	1	K

Item statistics

N	P	Total Rpbis	Total Rbis	Alpha w/o
1769	0.325	0.029	0.037	0.827

Option statistics

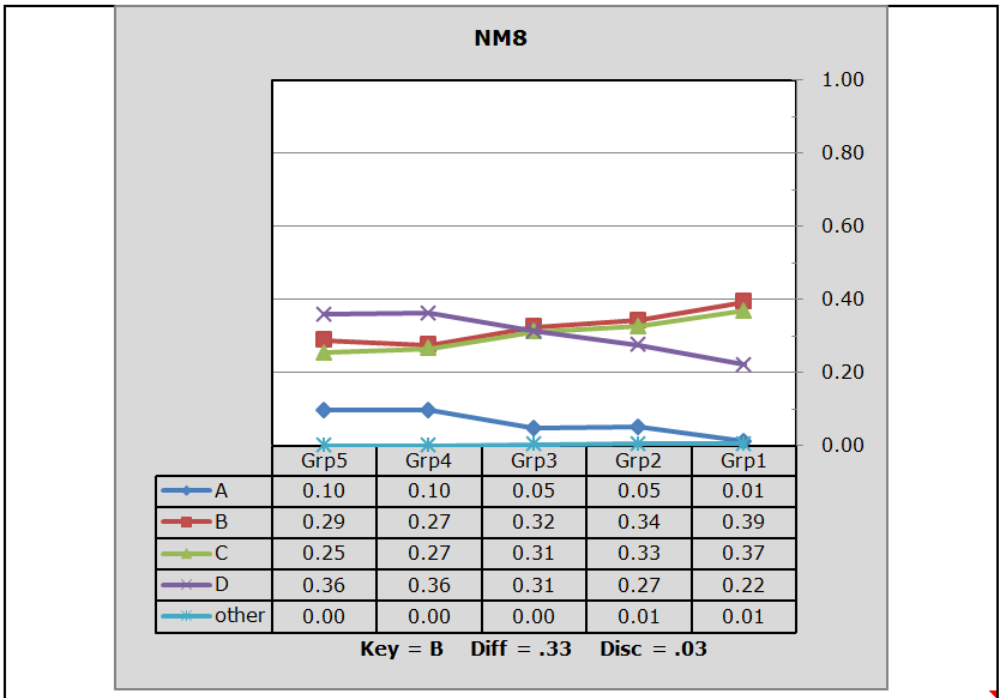
Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	107	0.060	-0.131	-0.261	36.308	6.676	Maroon	
B	575	0.325	0.029	0.037	41.602	7.855	Green	**KEY**
C	540	0.305	0.103	0.136	41.481	7.613	Blue	
D	542	0.306	-0.071	-0.093	39.463	7.479	Olive	
Omit	5	0.003	0.031	0.192	48.200	4.147		
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
A	107	0.098	0.096	0.049	0.054	0.011	Maroon	
B	575	0.290	0.276	0.319	0.351	0.389	Green	**KEY**
C	540	0.249	0.270	0.325	0.312	0.370	Blue	
D	542	0.364	0.358	0.307	0.283	0.230	Olive	

Note the "Scored" column in Iteman4's "Item information" table. Iteman4 allows unscored items to be easily included in an analysis – this is often done when a test includes pilot items. Lertap5 can be made to process pilot items too, but the way it does so involves more steps (see this example).

Lertap5's output for item NM8 is shown below:



NM8 (c10) *alpha with & without:* *0.825* *0.827*

option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	0.00	107	0.06	-0.14	-0.28	36.31	-0.56	
B	1.00	575	0.33	0.03	0.04	41.60	0.13	
C	0.00	540	0.31	0.08	0.10	41.48	0.11	<-aa
D	0.00	542	0.31	-0.10	-0.13	39.46	-0.15	
other	0.00	5	0.00	0.05	0.32	48.20	0.98	<-aa

Both Iteman4 and Lertap5 have essentially output the same information for item NM8, but the format of their graphs and tables has very obvious differences.

The equivalent to Iteman4's "Quantile plot data" table is found in the shaded area immediately below Lertap5's graph. The equivalent to the "Item statistics" table in Iteman4 is the Key = B line in the shaded area of Lertap5's output. The "Disc = .03" figure in Lertap5 is a Pearson product-moment correlation coefficient; this will equal Iteman4's "Total Rpbis" when the item is dichotomously scored, as was the case for item NM8.

The most "meaty" tables are the "Options statistics" table in Iteman4 and the equivalent results table in the clear area at the bottom of Lertap5's output. Lertap5's pb(r) is Iteman4's Rpbis, while b(r) in Lertap5 is Rbis is Iteman4. Here, "pb" means point-biserial correlation; "b", or "bis", means biserial correlation.

Iteman4 has an "SD" (standard deviation) column in the "Option statistics" table whereas Lertap5 replaces this with "z", the z-score corresponding to the "avg." test score for the students selecting each option.

Lertap5 has a "wt." column which indicates the number of points a student will get on selecting one of the item options. Iteman4 doesn't include this very basic information as it always gives one point for scored options, it does not have any flexibility for assigning other scores (or points).

When it comes to suggesting what may be wrong with item NM8, Iteman4 has placed a "K" flag in the small "Item information" table. The Iteman4 manual says that K signifies a *"Key error (Rpbis for a distractor is higher than Rpbis for the key)"*. Both Iteman4 and Lertap5 show that this was so: Rpbis for the keyed-correct option, B, was 0.03, less than 0.08 for option C, and also less than 0.05 for the Omit / Other rows.

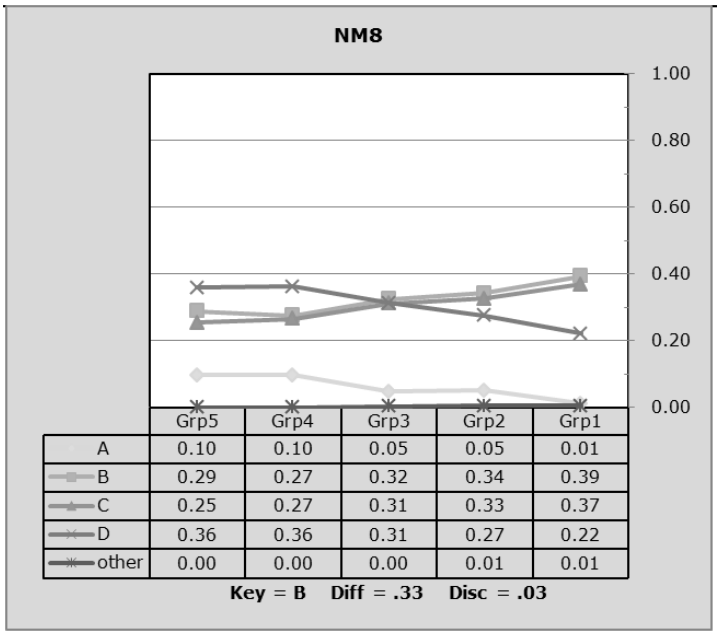
Lertap5 places its flags in the margin to the right of each row of results for each item option. For item NM8, Lertap5 has placed "aa" flags for option C and also for the "Other" row. Lertap5 is saying that the students who selected these options were above-average students – a glance down the "Avg." column will confirm this; it's even easier to see in the "z" column. In this case, the five students in the "Other" row, with an average test score of 48.20, were, on average, about one standard deviation above the mean (z=0.98). These students may have left item NM8 unanswered ("omitted"), or may have shaded in more than one "bubble" on their answer sheets, causing the scanner to enter an error code (which could be an asterisk, a Z, or a 9, depending on how the scanner was set up – Z was used in the case of the M.Nursing test).

Elsewhere I have argued that Lertap5's flags have some advantages over those used in Iteman4.

A unique feature of Iteman4 is that the criteria for some of its flags may be set by users. For example, it is possible to get Iteman4 to use an "LR" (low R) flag whenever an item's discrimination is less than, say, 0.05. This somewhat compensates for Iteman4's inability to use item labels in its tables, and for the lack of something like Lertap5's "Stats1b" report mentioned below.

Note the different markers used in Lertap5's plots for each line. These make the use of grayscale shading more feasible, something of issue when the plots are

printed on a black and white printer[4]. In Excel, grayscale shading is activated by using the "Colors" option on the Page Layout tab.



Lertap5 has two item response summary reports with no equivalent in Iteman4. They're called "Stats1b" and "Stats1ul".

The **Stats1b** report in Lertap5 looks like this:

Lertap5 brief item stats for "M.Nur Licensing E390v6.3", created: 4/04/2016.

Options->	A	B	C	D	other	Difficulty	Discrimination	?	h
NM1	5%	74%	19%	2%	0%	0.74	0.23		
NM2	36%	18%	37%	9%	0%	0.37	0.25		
NM3	12%	1%	1%	85%		0.85	0.13		
NM4	4%	14%	82%	0%		0.82	0.23		
NM5	91%	2%	6%	1%		0.91	0.16		
NM6	21%	41%	34%	4%	0%	0.41	0.13		
NM7	11%	35%	40%	14%	0%	0.35	0.25		
NM8	6%	33%	31%	31%	0%	0.33	0.03	C	
NM9	36%	15%	42%	7%		0.42	0.31		

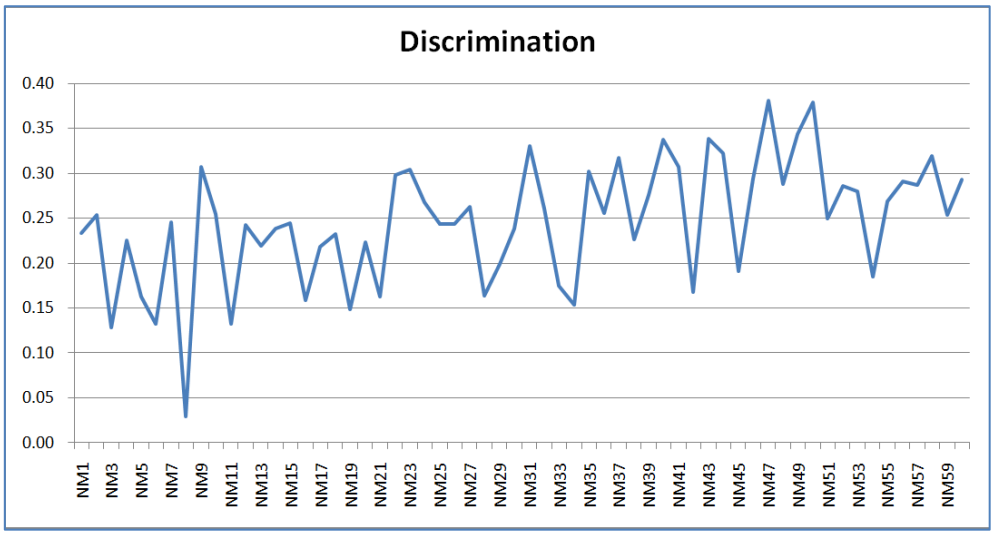
This report (above) summarizes item responses using just a single line for each item. Those item options which have been scored are underlined; for example, the correct answer to item NM4 was C, with 82% of students getting the item right.

[4] Iteman's various graphs look best with a color printer. The item response graphs from Iteman can at times be rather difficult to interpret when printed in black and white, in part because Iteman does not use unique line markers along the trace lines.

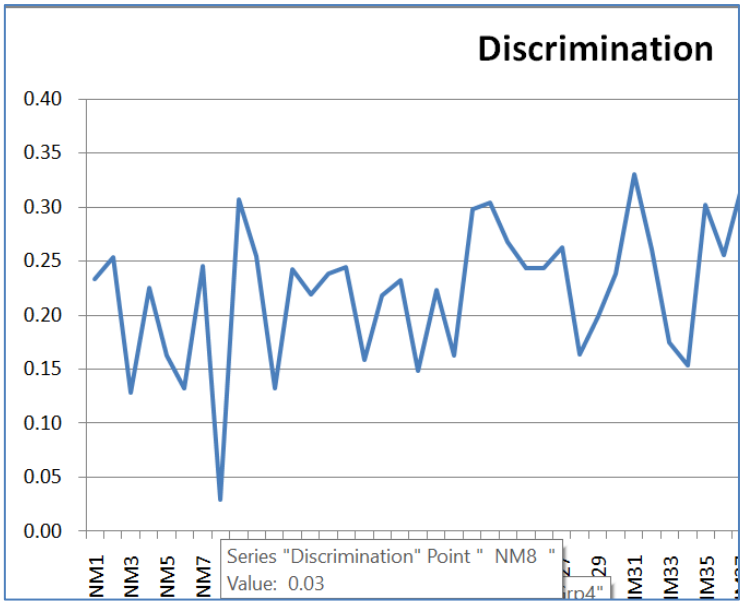
The **?** column is for flags; NM8's option C has been flagged in the snapshot above. A simple scan of the response percentages for NM8 hints at the item's problem: students are indicating that there wasn't a clear best answer, at least not in their opinion (about a third of the students took options B, C, and D).

The purpose of the Stats1b report is to give an easy-to-read, concise summary of how test items have behaved.

It's possible to sort a Stats1b report on the difficulty and discrimination columns; this makes it even easier to discern how the items have performed. And, a graph of item difficulty and discrimination results is easy to get using the "Line" option on the Lertap5 ribbon tab. An example is shown below, plotting the item discrimination figures for the M.Nursing test:



When Excel's graphs are looked at on screen, it's possible to move the mouse cursor along a line and get point readouts. For example, when I hovered over the lowest point in the graph above, Excel gave a readout: NM8 with an item discrimination of 0.03 (pictured below).



Many of Lertap5's reports have a small h like the one seen above at the top right of the Stats1b summary. This is a link to context sensitive online help. Were I able to click on the h above, I'd be taken to this webpage.

Lertap5's **Stats1ul** report has two sections; a snippet from the top section is shown below, followed by a snapshot of the bottom section:

Lertap5 U-L stats for "M.Nur Licensing E390v6.3", created: 4/04/2016.

Options->	A	B	C	D	other	U-L diff.	U-L disc.	h
NM1 Grp1	0.02	0.91	0.07	0.00	0.00	**0.74**	**0.35**	
NM1 Grp2	0.03	0.84	0.13	0.01	0.00			
NM1 Grp3	0.05	0.72	0.21	0.02	0.00			
NM1 Grp4	0.07	0.67	0.22	0.04	0.00			
NM1 Grp5	0.07	0.56	0.33	0.03	0.00			
NM2 Grp1	0.21	0.12	0.64	0.03	0.00	**0.44**	**0.41**	
NM2 Grp2	0.33	0.18	0.46	0.04	0.00			
NM2 Grp3	0.41	0.19	0.32	0.08	0.01			
NM2 Grp4	0.43	0.21	0.22	0.12	0.01			
NM2 Grp5	0.42	0.20	0.23	0.15	0.00			
NM3 Grp1	0.06	0.00	0.00	0.94	0.00	**0.85**	**0.17**	
NM3 Grp2	0.09	0.00	0.01	0.90	0.00			

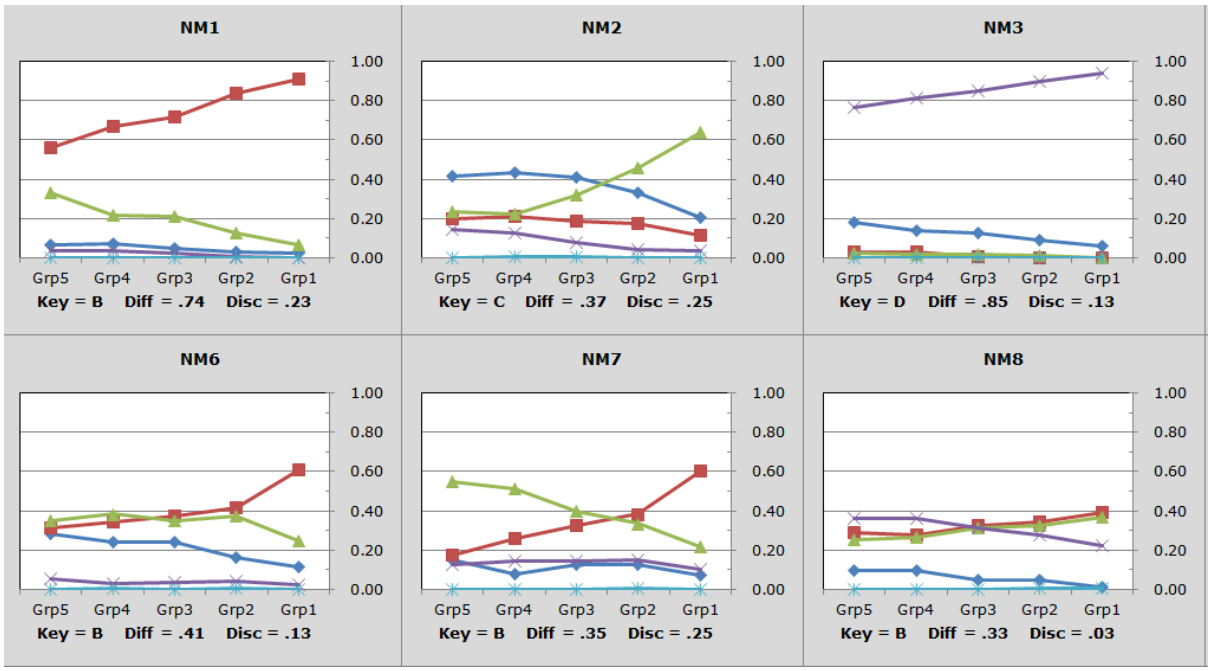
Summary group statistics

	n	n(%)	avg.	avg%	s.d.	min.	mdn.	max.
Grp1	353	19.95%	51.1	85%	2.5	48	51	60
Grp2	353	19.95%	45.4	76%	1.5	43	45	48
Grp3	357	20.18%	40.7	68%	1.3	39	41	43
Grp4	353	19.95%	36.3	60%	1.4	34	36	39
Grp5	353	19.95%	29.6	49%	3.6	2	30	34
everyone	1,769		40.6	68%	7.7	2	41	60

This was an upper-lower analysis with more than two groups.

Clicking on the little h for the Stats1ul report leads to this webpage.

It is possible to pack the item response plots made in Lertap5. With a large computer monitor, "packed plots" are another *very* effective way to spot poorly performing items – it's possible to display as many as 50 item plots at once on a 24" monitor. I present an example below, using six of the M.Nursing items. The problem with NM8 is readily spotted on the right side of the plot; the line tracing response proportions for the correct option, the red line, with little squares used as markers, does not clearly rise above the trace lines for the distractors.



Here's a webpage from another test, providing a better example of packed plots in action.

PDF or xlsx; Iteman 4 or Lertap 5?

To this point, my comparison of the two programs is far from complete. Each program has more options and capabilities. Readers might ask, for example, how their output compares when survey items are processed, such as those found in rating scales.

Each program has an option to check for possible cheating; how do they compare? (Please see Appendix A.)

What about mastery, licensing, and certification tests? Both Iteman 4 and Lertap 5 have options to analyse such tests. Lertap 5's methods are exemplified here; what about Iteman 4, what's in its output for these tests? (Please refer to Appendix C.)

Both programs support DIF analysis, differential item functioning. What does their output look like, how do they compare? (Please see Appendix B.)

As time permits I will add answers to these questions (and others) in this paper[5]. In the meantime, this link will provide more details about Lertap 5's options and features for those interested.

A fundamental difference, in my opinion, is that Iteman 4 is a "passive" system. It's flat. You prepare all data offline as it cannot be done within Iteman 4. There's no data editor, no way to look at the quality of the incoming data, no way to check for errors, no way to weed out bad data records. I could never have undertaken this study of a country's national assessment project with Iteman 4.

The main output in Iteman 4 is a word processor file. You can't change the color used in its graphs – the output is designed for printing; it's not live. You can't swipe down a column of figures, such as discrimination coefficients, and get a graph. The graphs themselves often tend to be a bit small.

Yes, Iteman 4's RTF file prints well; that can be quite handy. But Iteman's output is verbose; the RTF file from the M.Nursing test comprised 68 pages. It doesn't look so good in black and white. Printing in color can be expensive, especially when copies are wanted.

Enter Lertap 5. It's a self-contained system – the actual item response data are contained within Lertap 5. Most of Lertap's reports are relatively short; users will seldom print them all, they'll select the one most suited to their needs, and print that one – if this were, for example, the Stats1b report, the M.Nursing results would print in four pages or less.

Lertap's CCs worksheet, the counterpart to Iteman's item control file (ICF), is more often than not very easy to set up. The ICF for M.Nursing required that 60 lines be created in a word processor and saved as a text file (Excel could also be used); just three lines were required in Lertap's CCs worksheet, and, again, being a self-contained system, these control lines were prepared within Lertap 5 itself.

Excel's graphics are clearly superior to those found in Iteman 4. It's easy to change their colors; even simple to change their style (here's an example).

Lertap's documentation is more extensive than Iteman's. There are four supporting websites, sample data sets, context-sensitive help, technical support papers, videos, materials for use in measurement classes. (See this website.)

In recent years, Lertap 5 has experienced more growth, with new features and enhancements frequently added, often at user request. See this link.

[5] Feedback from readers is welcome; write support@lertap.com and mention what you'd most like to see.

Try 'em

It doesn't cost even a penny to try these two programs. Both Iteman 4 and Lertap 5 have trial systems available for test drives. Iteman's is called the "demonstration version" – it allows processing up to 50 items for up to 100 people. Lertap's is called the "mini version"; it is also limited to processing results from 100 people, but has no limit on the number of items. Both versions are absolutely free. And both of these "trial systems" may be converted to unrestricted versions with full power without downloading and installing them again, something that's done with the purchase of a license at one of the respective stores.

The store for Iteman 4 is here, while the gateway to purchasing Lertap licenses is here. Lertap has a free one-year license for students; Iteman offers a similar deal but for just six months.

The many screen snapshots in this paper have been taken from the Iteman 4 RTF output file, and the Lertap 5 Excel workbook for the M.Nursing test. You can see the real stuff, the whole picture, all of the output, with a couple of mouse clicks: for Iteman click here, and for Lertap click here.

Appendix A: Response Similarity Analysis

I selected a 44-item multiple-choice test completed by 2,508 university students, and processed their item responses with Wesolowsky's SCheck program, Lertap 5, and Iteman 4.

SCheck produces quite a bit of output; from it I have extracted the following summary statistics and "QQPLOT". SCheck found 157 suspicious pairs of students.

SUMMARY

==========

mean of Z's = 0.0452 stdev= 0.6995

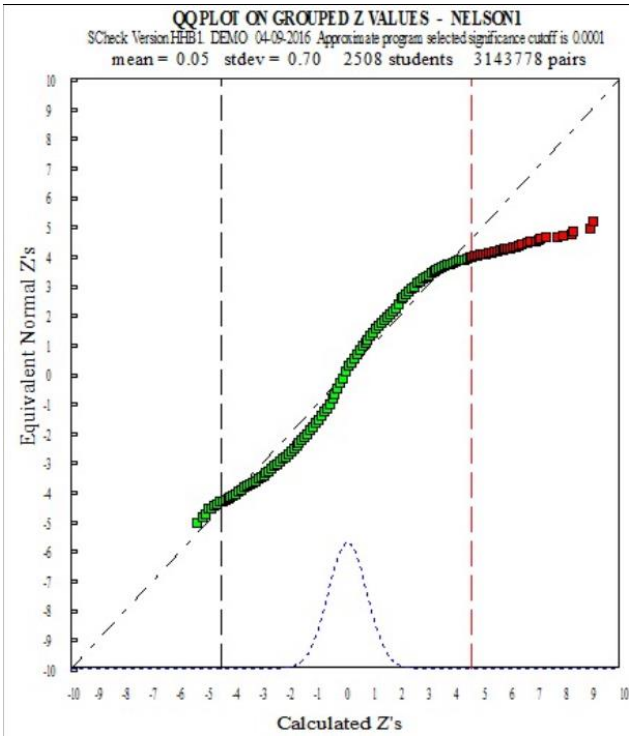
The number of pairs checked is 3143778

The specified significance for program-selected pairs is 0.000100

The approx. cutoff Z based on the specified significance is 4.571000

The cutoff Z based on the Bonferroni bound is 6.53

Number of observations below -4.571 or above 4.571 is 157



Lertap 5 also creates a lot of output when its RSA routine runs. It found 172 suspect pairs, with 107 of them judged to be "very suspect". The little table below is from Lertap 5's "RSAsig" report. Of the 157 suspects found by SCheck, 141 were included in Lertap 5's list of suspects.

Pairings	
Suspect:	172
Not suspect:	1,459,314
Total:	1,459,486
Inclusions	
Number of items:	43
Number of students:	1709
Run control	
EEIC minimum:	8
H-H index minimum:	1.5
H-H sigma minimum:	5
Items excluded:	1
Minimum score setting:	8
Maximum score setting:	37

In comparison, Iteman 4 produces very little output, just a "csv" file with a "BBO matrix" readily viewed in Excel. A sample snapshot is shown below; the whole spreadsheet may be downloaded from [this link](#).

IM4_Cheat_Check1 BBO-matrix.csv

	1	2	3	4	5	6	7	8	9	10
1		Flag	S1	S2	S3	S4	S5	S6	S7	S8
2	S1	0								
3	S2	0	1							
4	S3	0	1	0.53						
5	S4	0	1	0.28	0.53					
6	S5	0	1	1	1	1				
7	S6	0	1	1	1	1	1			
8	S7	0	1	1	1	1	1	1		
9	S8	0	1	0.53	0.53	0.53	1	1	1	
10	S9	0	1	0.53	0.53	0.53	1	1	1	0.5
11	S10	0	1	1	1	1	1	1	1	
12	S11	0	1	1	1	1	1	1	1	
13	S12	0	1	0.53	0.53	0.53	1	1	1	0.5
14	S13	0	1	0.78	0.53	0.78	1	0.53	1	0.5
15	S14	0	1	0.53	0.53	0.53	1	1	1	0.5
16	S15	0	1	0.28	0.53	0.28	1	1	1	0.5
17	S16	0	1	0.28	0.53	0.28	1	1	1	0.5

IM4_Cheat_Check1 BBO-matrix

That's it. All we get from Iteman 4 is this "BBO matrix". There is no supporting documentation; the manual makes reference to "Bellezza and Bellezza (1989)[6]" without providing a complete citation. We are left with something of a mystery. Perhaps this is some sort of work in progress; at the moment we might only conclude that Iteman 4 does not truly have any support for what I have termed response similarity analysis.

[6] Here it what the manual should have included: Bellezza, F.S., & Bellezza, S.F. (1989). Detection of cheating on multiple-choice tests by using error-similarity-analysis. *Teaching of Psychology*, 16(3), 151-155.

Update August 2019: it was indeed a work in progress. ASC ended up recommending that users might want to use their "forensics" package called "SIFT". This came to be a free, unsupported package. It has a manual, quite useful for its review of response similarity analysis statistics and methods, but the software itself is in a very primitive state, hardly useful at all.

{This webpage has two sample datasets to use with cheat-checking software.}

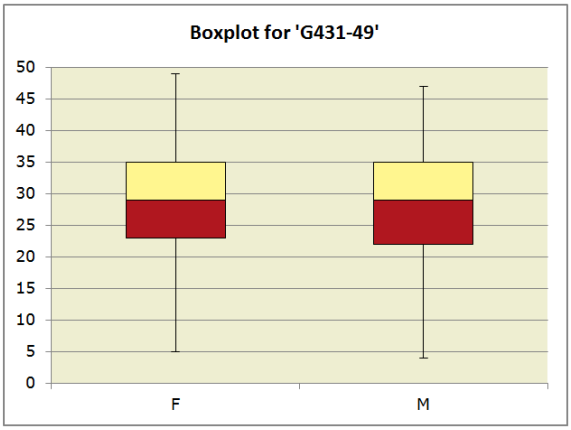
Appendix B: Differential Item Functioning, "DIF"

I selected a 49-item multiple-choice test completed by 4,712 middle-school geology students, and processed their item responses with SPSS[7], Lertap 5, and Iteman 4. All three use Mantel-Haenszel procedures to look for any possible DIF effects which may underlie responses to each of the items.

Before applying DIF methods, I needed to make sure that these two groups did not have different overall test scores, something I would be willing to accept as an indication of similar subject proficiency.

Two of the three programs, SPSS and Lertap, have routines that may be used to compare group scores – the results below were obtained by using Lertap 5's "Breakout scores by groups" option and indicate that the overall test scores were similarly distributed in the two groups.

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



My work was based on the original item response data as found in this Excel workbook. I prepared data for use with SPSS by following the steps described in the "SPSS and DIF" section of this document[8]. Iteman 4 requires two input files, one with data, one with item control information; I used Lertap 5 to prepare them – see the data file here, and the item control file here.

As to comparing results, I will start with those produced by Iteman 4 as found in the program's main output file, an RTF file made to be read with a word processor – see it here; in this case, the RTF file was 57 pages in length.

Iteman 4 found evidence of differential item functioning for the four items mentioned (and "flagged") in the following table:

[7] I used SPSS 17.0

[8] I set the chi-sq. "continuity correction" in Lertap's System sheet to "yes" as SPSS uses the correction.

Table 4: Summary Statistics for the Flagged Items

Item ID	P / Item Mean	R	Flag(s)
q11	0.583	0.391	DIF
q17	0.779	0.338	DIF
q28	0.350	0.384	DIF
q43	0.678	0.428	DIF

Iteman's item-level results for q11 are seen below:

Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
11	q11	3	Yes	4	1	DIF

Item statistics

N	P	Total Rpbis	Total Rbis	Alpha w/o	M-H	p	Bias Against
4712	0.583	0.391	0.494	0.868	0.610	0.002	Females

Item q11 has been given a DIF flag by Iteman as the M-H statistic of 0.610 is statistically significant with p=0.002. Being less than 1.00, the M-H statistic is suggesting that this item favours the focal group. The reference and focal groups for DIF are defined just before Iteman runs; in this case I defined females as the reference group, males as the focal group. As will be seen later in the Lertap output, the males, the focal group, did better on q11; this is why Iteman says this item appears to be biased against females.

SPSS produced 60 pages of tables. The two tables below pertain to item q11:

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	52.468	1	.000
Mantel-Haenszel	51.600	1	.000

Mantel-Haenszel Common Odds Ratio Estimate

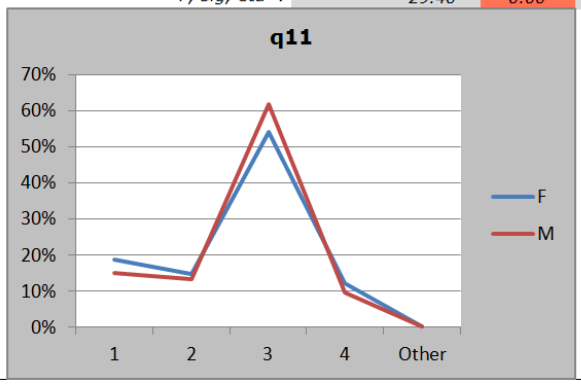
		Estimate	.617
		ln(Estimate)	-.482
		Std. Error of ln(Estimate)	.067
		Asymp. Sig. (2-sided)	.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	.541
		Upper Bound	.704
	ln(Common Odds Ratio)	Lower Bound	-.614
		Upper Bound	-.351

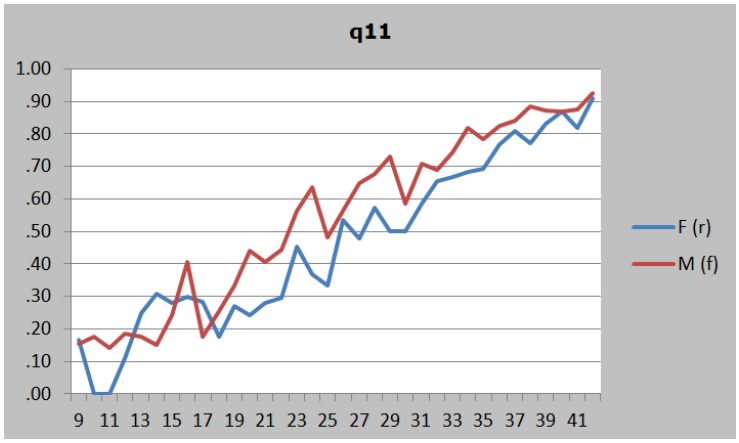
Lertap 5 output for q11 is shown below:

q11										
F *diff*	.00	.00	.00	.00	.00	.17	.00	.00	.11	.25
M *diff*	.00	.50	.00	.29	.23	.15	.18	.14	.19	.18
odds ratio->	.00	.00	.00	.00	.00	1.10	.00	.00	.55	1.57

MH chi-sq: 51.60 *Prob:* **.00** *MH alpha:* .62 *MH D-DIF:* **1.13** *ETS level:* **B (mod.)**

q11	1	2	3	4	Other	n	mean	s.d.
F	19%	15%	**54%**	12%	0%	2153	0.54	0.50
M	15%	13%	**62%**	10%	0%	2559	0.62	0.49
				*DIF sig, alpha, **delta**, ETS:*	.00	0.62	***1.13***	*B (mod.)*
				F, sig, eta²:		29.40	0.00	0.01





The interpretation of Lertap's DIF output is the subject of this document. The small table above, on top of the two graphs, shows a chi-sq. value of 51.60, p of .00, agreeing with the SPSS output. Iteman does not output the chi-sq. value; we can assume that the p of 0.002 reported by Iteman is its test of the corresponding (but unreported) chi-sq.

Iteman 4 allows for a maximum of seven score intervals; when there are more than seven test scores, Iteman collapses the scores so that they fit into seven intervals; SPSS and Lertap do not have this restriction -- they will use the entire range of scores when possible, resulting in 46 possible score intervals for this example.

Despite the differences in the number of score intervals, for q11 all three programs have similar estimates of M-H alpha, the common odds ratio:

Iteman=0.610, SPSS=.617, and Lertap=.617 (rounded to 0.62 in the tables above).

Over all 49 test items, Iteman found just four with a significant chi-sq. value: items q11, q17, q28, and q43[9]. Lertap and SPSS found *thirteen*, all with p<.05: q11, q13, q15, q17, q22, q23, q24, q28, q30, q32, q43, q44, and q47. I would suggest that the reason for these differences has to do with the number of score intervals used to estimate MH-alpha and to calculate the chi-sq. value – both Lertap and SPSS use many more intervals.

Lertap goes beyond Iteman and SPSS in its DIF output, creating more statistics, and providing graphs to aid in the interpretation of how the item responses of the two groups differed[10]. Iteman 4's DIF output is, in comparison, limited, and, as just mentioned in the paragraph above, in disagreement with the results found by Lertap 5 and SPSS. Iteman 4 does not have an option for comparing overall test scores for the two DIF groups beforehand – it must be assumed that this has been done prior to applying the program[11]. Iteman 4's manual does not provide information about the continuity correction sometimes used in DIF work; consequently it is impossible to determine if this correction is used.

Update August 2019: there is a free R package, "**difR**", with more DIF options than those found in Lertap5. See this paper.

[9] The Iteman manual does not give the cutoff for p; from looking at the output, it appears that p<.05 is the cutoff used to determine statistical significance in Iteman 4.

[10] Note: of the two graphs shown above, the first is standard Lertap output; the second requires a bit of extra work from users as described in this document.

[11] The two groups in a DIF analysis are assumed to have equal or near-equal subject proficiency. This should be tested before the analysis is initiated.

Appendix C: Mastery Testing

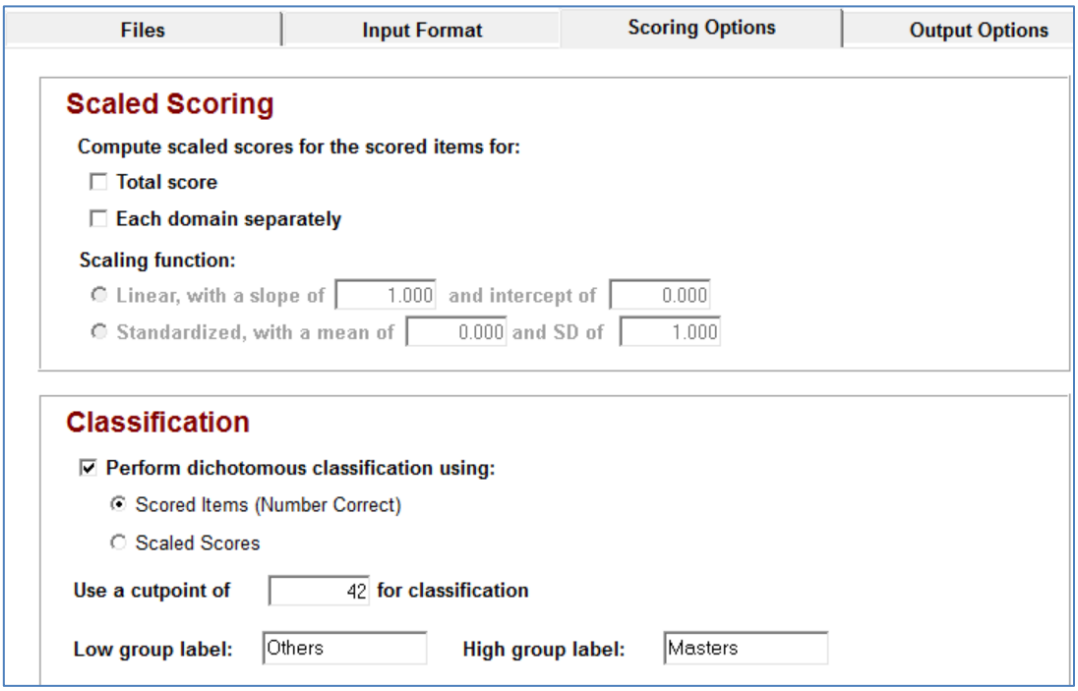
Lertap 5 has had support for criterion-referenced, mastery, licensing, and certification testing since the year 2001; a reference is Chapter 7 of the manual. Iteman 4 now also has some support in these areas, and I thought I'd compare what the two programs do.

I began by making a copy of the M.Nursing dataset mentioned here. The actual copy and Lertap results relevant to this section may be perused by downloading this Excel workbook. I used Lertap to prepare the two input files required by Iteman.

Page 1 of the Iteman 4.3 manual says: "Scores can be classified into two groups at a specified cut score, and the two groups can use your labels."

Later on in the manual, on page 13, instructions are provided "to perform a dichotomous classification" using a cutpoint.

I set these options in Iteman's "Scoring Options" tab:



What happened when I ran Iteman 4 with the scoring options set above? I got the usual RTF output file with test and item results as summarised above, in the main part of this paper. A link to the Iteman RTF report is here. The following is found after Table 2: Summary Statistics:

> The cutscore on this exam was 42.000, producing a pass rate of 45.7%. The Livingston index of classification consistency at the cut-score was 0.828.

Still in the same report, after Figure 4, just before Figure 5, the Iteman 4 report provided the following data:

> The CSEM at the cutscore of 42.000 equaled 3.288.

The output below was created by Lertap 5 and found in its Stats1ul report:

Summary group statistics

	n	n(%)	avg.	avg%	s.d.	min.	mdn.	max.
masters	808	45.70%	47.5	79%	3.9	42	47	60
others	960	54.30%	34.9	58%	4.7	18	36	41
everyone	1,768		40.6	68%	7.7	18	41	60

This was an upper-lower analysis based on a mastery cutoff percentage of 70% (cut score = 42) .

Variance components

	df	SS	MS
Persons	1767	1726.70	0.98
Items	59	3343.03	56.66
Residual	104253	18120.57	0.17

Hoyt's reliability coefficient:	0.822	
CSEM at the cut score:	3.260	
Livingston's coefficient:	0.828	
Index of dependability:	0.796	
Estimated error variance:	0.003	
For 68% conf. intrvl. use:	0.059	
Prop. consistent placings:	0.808	(Estimated number of incorrect classifications: 338)
Prop. beyond chance:	0.614	

Both of the programs output the core information necessary to complete an NCCA report (see this little paper).

Lertap 5 has provided particularly extensive information, with a variety of relevant statistics to consider. The "**Prop. consistent placings**" statistic is a most compelling consistency index, easy to interpret, with strong support in the literature (see this paper for a technical discussion). It is what I recommend for the NCCA report.

This webpage has more coverage of Lertap 5's output in the area of mastery testing and classification consistency.