Visual Item Analysis

Larry R Nelson
Curtin University & Burapha University
Document date: 28 August 2016

This document discusses graphics-based methods for assessing the quality of cognitive test items. It suggests that, with some experience, Lertap's quintile plots may serve as an effective visual alternative to the many tables commonly found in Lertap output, permitting item effectiveness to be determined by eye. The document also suggests that "non-standard" quintile plots may be used to visually detect group differences in item response patterns[1].

Updates: Lertap 5's capabilites for plotting item responses and creating response trace lines have been extensively enhanced in Version 5.10.5, released in January 2015. Admire the changes here.

Quintile plots

Quintile plots are Excel charts created from the information commonly found in the "Statsul" reports created by Lertap; such plots will play a big role in this paper. Here's an example of the output found in a typical Statsul report:

Options->	1	2	3	4	5	other	U-L diff.	U-L disc.
A27mc Grp1	0.00	0.00	0.00	0.95	0.05	0.00	**0.56**	**0.77**
A27mc Grp2	0.04	0.02	0.05	0.70	0.19	0.00		
A27mc Grp3	0.02	0.12	0.10	0.47	0.27	0.03		
A27mc Grp4	0.09	0.12	0.07	0.26	0.46	0.00		
A27mc Grp5	0.19	0.12	0.09	0.18	0.37	0.05		

The table above provides data for one item, item "A27mc". The item used five options, with corresponding response codes of Res=(1,2,3,4,5). The correct answer to this item was 4, signified by the underlining seen in the column under the 4.

The results in the table have been broken out by five groups, one row per group: "Grp1", to "Grp5". These groups are based on dividing the distribution of test scores into five levels: the upper 20% (Grp1), the 2nd-highest 20% (Grp2), the 3rd-highest 20% (Grp3), the 4th-highest 20% (Grp4), and the lower 20% (Grp5). The "other" column indicates the proportion in each group who did not answer the item.

The Statsul report above reveals that the proportion of students in the top group who were able to identify the correct answer, 4, was 0.95, or 95%. Note how this proportion drops as we go down the correct answer's column: it goes from 0.95, to 0.70, to 0.47, to 0.26, and then, in the "lower" group, to 0.18. Fewer than 20% of the students in the weakest group were able to pick out the correct answer to item A27mc. The report also indicates that option 5, an incorrect answer, a "distractor", was popular with the two lowest groups, Grp4 and Grp5.

---

[1] My thanks to Dr. Leon Gross, Director of Psychometrics & Research, National Board of Examiners in Optometry, U.S.A., for his support and encouragement in the development of this paper. Any errors are due to my short-sightedness, not his.

It was also selected by 27% of the middle group, by 19% of the 2nd highest group, and even by 5% of the top students, those in Grp1.

The end of a Statsul report has a descriptive summary of the five groups, as seen here:

Summary group statistics

	n	n(%)	avg.	avg%	s.d.	max.
Grp1	57	19.79%	32.8	66%	2.9	40
Grp2	57	19.79%	26.4	53%	1.4	29
Grp3	60	20.83%	21.5	43%	1.3	24
Grp4	57	19.79%	17.0	34%	1.1	19
Grp5	57	19.79%	11.9	24%	2.8	15
everyone	288		21.9	44%	7.5	40

In this case, 288 students were tested. Lertap has made an effort to form five groups, each with an equal number of students. This it couldn't do as 288 divided by 5 results in 57.6. In this case we've ended up with four groups having 57 students, with Grp3, the middle group, having 60.

The "Summary group statistics" indicate: "n", the number of students in a group; "n%", the percentage in each group; "avg.", the average test score for a group; "avg.%", avg. expressed as a percentage of the maximum possible score; "s.d.", the standard deviation of group scores; "min.", the lowest score found in a group; "mdn.", the median of the test scores found in each group; and "max.", the highest score found in each group.

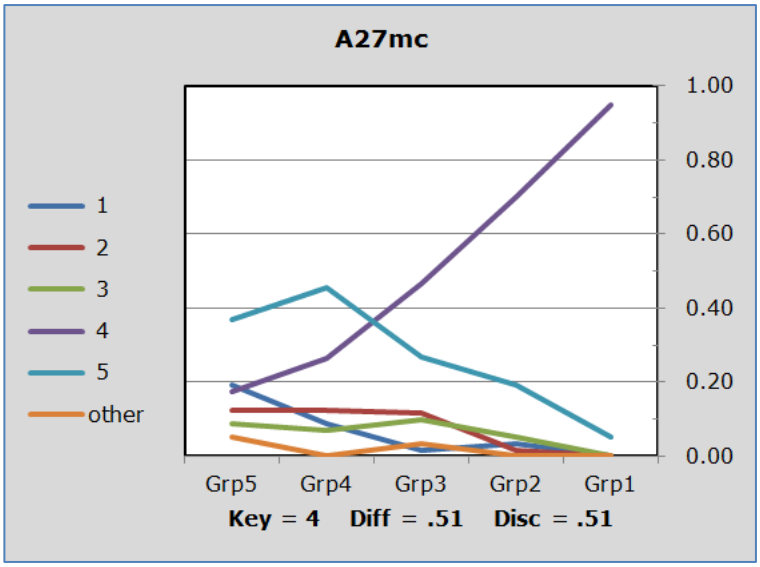
Item analysis involves a process of determining how well test items meet the objectives they have been designed for. Cognitive tests are often meant to help us identify the "best" students, those whose knowledge of relevant subject matter is strong, above average. Items which assist in this process are said to be ones which can "discriminate" the good students from the weak ones.

Traditional item analysis has been based on a study of numeric summaries of results. If an item is discriminating, only the best students will get it right; the other students will be drawn off by the distractors. Our item A27mc would seem to be a good item: the proportion of students who correctly answered the item was higher the stronger the group. Conversely, the proportion of students who selected the distractors was highest in the lowest groups.

Lertap provides several numeric summaries of results: they're found in the Statsul report exemplified above, and in two other report formats which will be mentioned later, Statsf and Statsb.

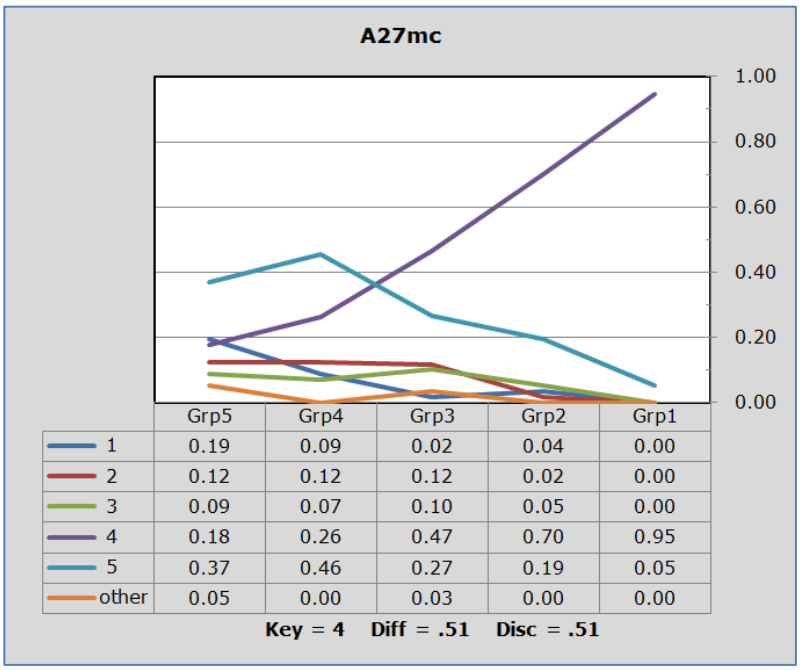
In this paper we'll look at an alternative way to go about determining whether or not an item has discriminated – we'll use graphs referred to as "quintile plots". Such plots can very effectively portray how well items have discriminated.

Quintile plots are made from the proportions shown in the first table above, from the Statsul report. The "standard" Lertap quintile plot for item A27mc is shown below:



How many lines are included in this graph? Six: see the legend to the left of the plot. There's one line for each item option, 1 to 5, and a sixth line, "other", corresponding to people who did not answer the item.

It's possible to have Lertap add a table at the bottom of the plot, as shown below:



	Grp5	Grp4	Grp3	Grp2	Grp1
1	0.19	0.09	0.02	0.04	0.00
2	0.12	0.12	0.12	0.02	0.00
3	0.09	0.07	0.10	0.05	0.00
4	0.18	0.26	0.47	0.70	0.95
5	0.37	0.46	0.27	0.19	0.05
other	0.05	0.00	0.03	0.00	0.00

The purpose of Lertap's standard quintile plot is to graphically trace the performance of each item option and, in so doing, allow us to visually assess the quality of the item in terms of its ability to help us discriminate strong students from weaker ones.

A cognitive item meant to identify the best students, those whose mastery of test content is strong, will have a plot similar to the one above.

There will be one line, that corresponding to the right answer, which rises from left to right.

For item A27mc, the correct answer was 4. Look at the proportion of students in each group who were able to pick out the item's correct answer: we start with 18% in Grp5, the weakest group, the "lower" group, increase to 26% in the next-lowest group, the "4th" group; go up to 47% in the middle group; to 70% in the 2nd-best group; and then, in Grp1, the top group, hit 95%.

A discriminating item will have trace lines for the wrong answers, the distractors, which drop, which dip down, as we go from left to right. The purpose of the distractors is to pose item answers which will seem quite plausible to weaker students, but not to the top students. Effective distractors are an absolute must for discriminating items; if they do not appear as reasonably possible answers to some, almost everyone will get the item right, thwarting our effort to identify the most proficient students.

Item A27mc has one distractor which worked particularly well: 5. This distractor was selected by 37% of the lower group, and by almost half, 46%, of the next-lowest group, the 4th group. As we go from left to right across the columns of the table below the plot, we find a pattern: the popularity of distractor 5 falls off, dropping to 5% for the top group. This is what we expect of well-performing distractors: they will seem plausible only to those whose content mastery is weak – the top students will see through them.

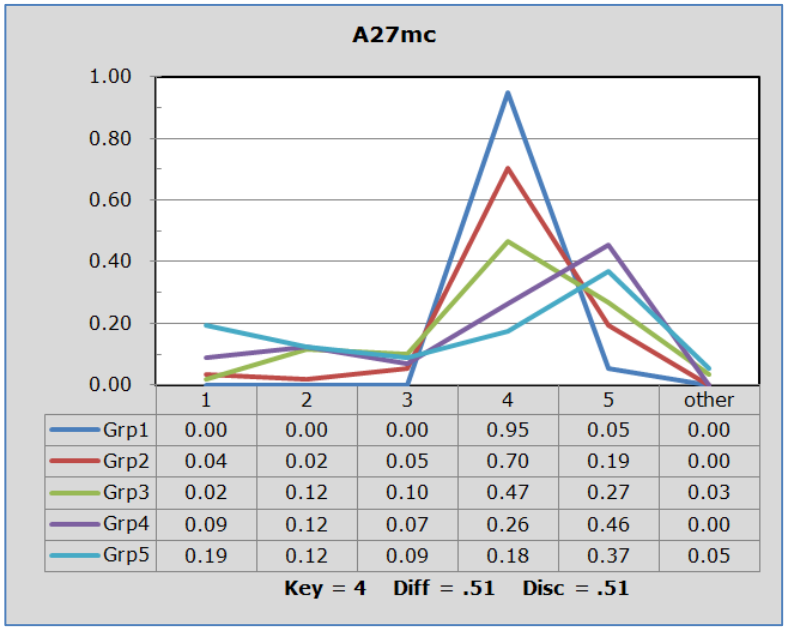
What's found in a Statsul table is pictured in the quintile plot: we can easily see the graph of the right answer increasing as we go from left to right, from weaker to stronger students. The distractors do the opposite: their popularity falls away from left to right, generally steadily dipping down.

Quintile plots can take the place of Lertap's various reports, at least to a certain extent. Once you've got the hang of them, it's possible to pick out poorly-functioning items by scrolling down the graphs, looking for items which don't display the desired pattern. The easiest give-away is simple: look for plots which do not have one line rising rapidly from left to right; these are likely to be the items most in need of attention.

The non-standard quintile plot in Lertap

The quintile plots pictured above have been referred to as the "standard" one. This is because Lertap has a second quintile plot, one where the x-axis corresponds to item options instead of quintile groups.

Here's our item A27mc in a "non-standard" plot:



	1	2	3	4	5	other
Grp1	0.00	0.00	0.00	0.95	0.05	0.00
Grp2	0.04	0.02	0.05	0.70	0.19	0.00
Grp3	0.02	0.12	0.10	0.47	0.27	0.03
Grp4	0.09	0.12	0.07	0.26	0.46	0.00
Grp5	0.19	0.12	0.09	0.18	0.37	0.05

Key = 4 Diff = .51 Disc = .51

The lines in this plot now trace the five groups over each of the item options. In the so-called "standard" plot, the lines traced the options over each of the five groups – here we have simply swapped things around.

Let's review what we expect of a "good" cognitive item: its correct answer will be identified by the strongest students, those in Grp1 and Grp2; the distractors will be selected by the weakest students. Thus, we would expect the lines for the top two groups to be below the other lines for each distractor, and above them for the right answer.

Is this what happens in the graph? Almost; there's a small problem with distractor 1: the 2nd-highest group comes in above the 3rd group. Otherwise we're okay.

Standard versus non-standard

Given that Lertap will make two types of plot, which is best? That would be up to you to answer, of course. The first type, referred to here as the "standard", is probably the most popular. The format of the "standard" quintile plot bears some similarity to the "ICC", the item characteristic curve found in item response theory; here's an example of an ICC taken from Crocker & Algina (1986):

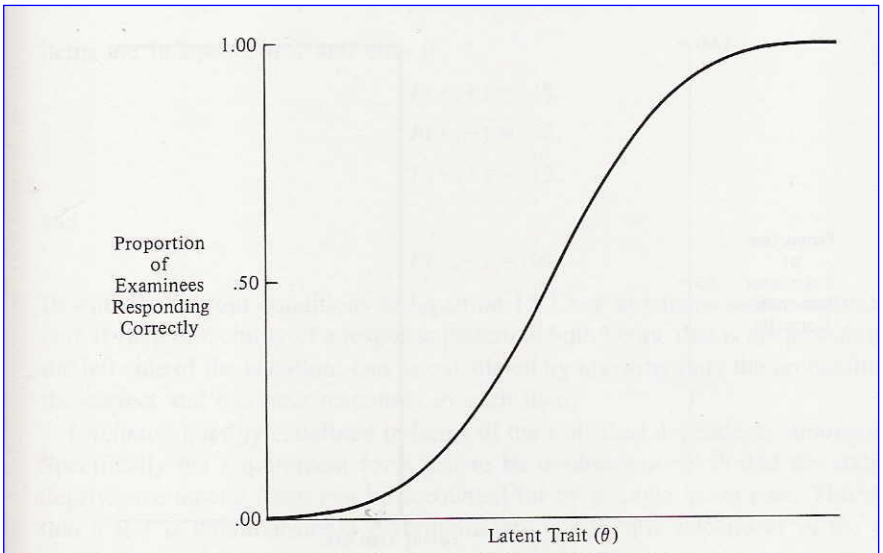


FIGURE 15.1 The Relationship Between Ability and Item Response

The y-axis of the ICC, the vertical axis, is the same as that found in Lertap's quintile plots. The abscissa, on the other hand, the x-axis, is similar but not identical. The ICC abscissa refers to a "latent trait" of the people tested; du Toit (2003, p.832) writes "In much of the IRT literature, this latent variable is referred to as 'ability', but in an educational context a more apposite term is 'proficiency'". (Note that the ICC does not concern itself with distractors, only with correct answers; Figure 15.1 indicates that the proportion of test takers getting the item correct steadily increases in accordance with their latent ability.)

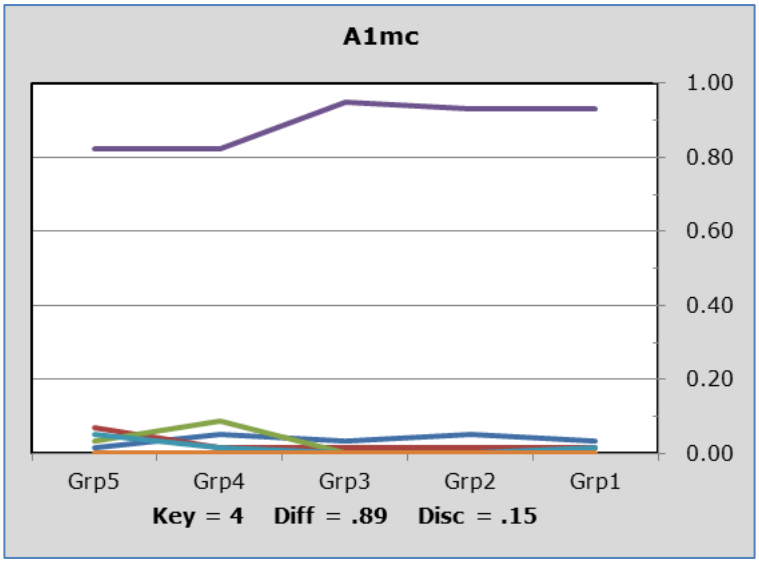
The x-axis seen in Lertap's "standard" quintile plot is usually based on the scores of those tested. If we take these scores as being related to "ability", then the standard quintile is plotting over something similar to the ICC. Of course, the ICC uses a continuous x-axis whereas Lertap's x-axis is based on just five grouped levels of test scores.

Lertap's standard quintile plot is essentially identical to the item trace lines seen in Wainer (1989), although Wainer's x-axis, like that of the ICC, is continuous, i.e., not broken into the five grouped levels of test scores seen in Lertap's plots.

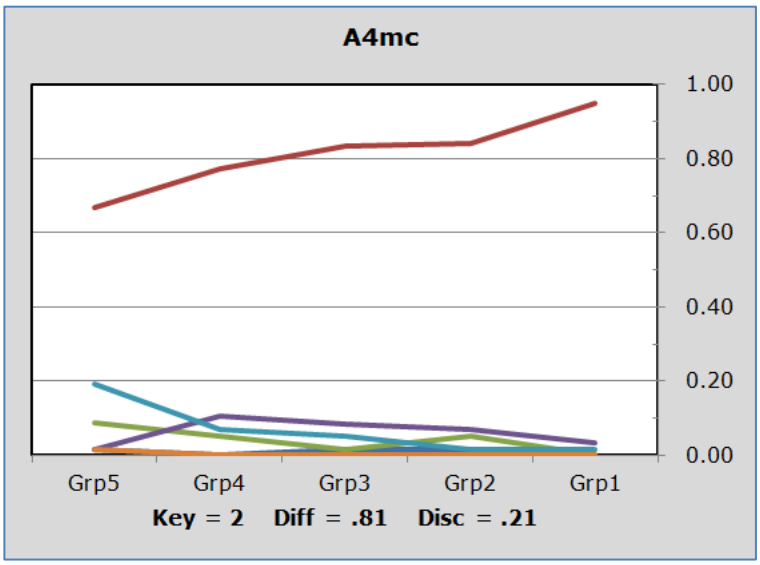
We might conclude, then, that the "standard" quintile plot is more aligned to the types of item response information displays found in the literature.

Sample quintile plots

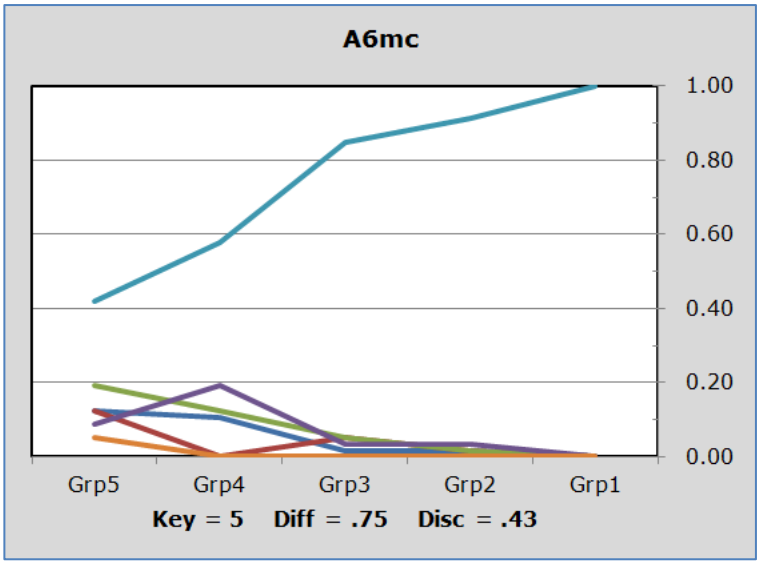
A look at sample standard quintile plots will be useful. Most of the plots below are taken from items belonging to a professionally-developed aptitude test for high school students. The test had 70 items. Fifty (50) of the items were five-option multiple-choice questions, using response codes of {1,2,3,4,5}. The remaining twenty items were short-answer questions.



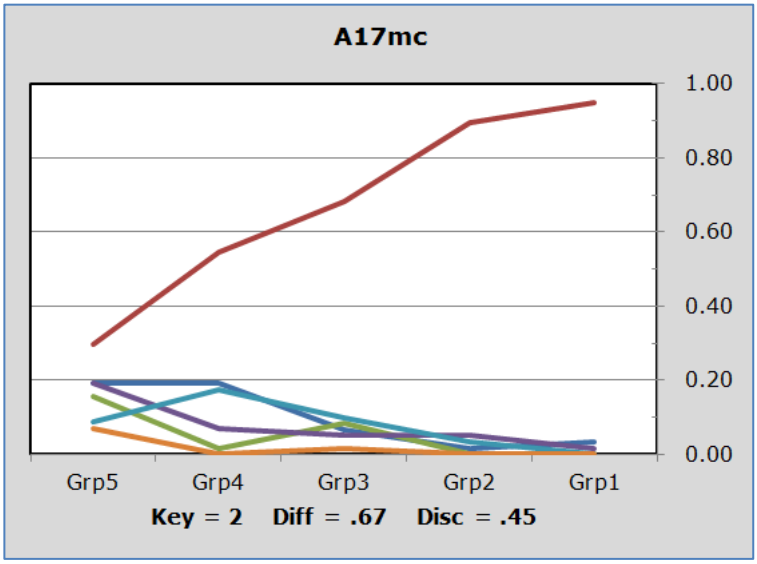
This item, A1mc, reflects the pattern found for easy test items. The item is not discriminating well at all; the proportion of people able to find the correct answer is high even in the lowest groups, Grp5 and Grp4; the four distractors are apparently not seen as plausible answers by more than 10% of the test takers in any of the five groups. This was the first item in the test, and it was intentionally designed to be easy.



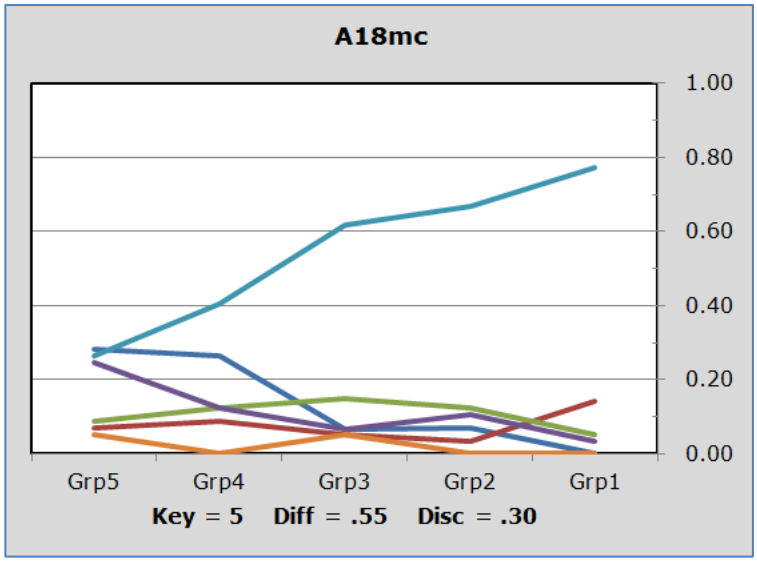
Item A4mc is also easy, but not quite as easy as A1mc. Here we're starting to see the pattern desired of a discriminating item: the trace for the right answer dips on the left, and then steadily rises. The trace lines for the distractors are highest on the left, generally dipping down to the right.



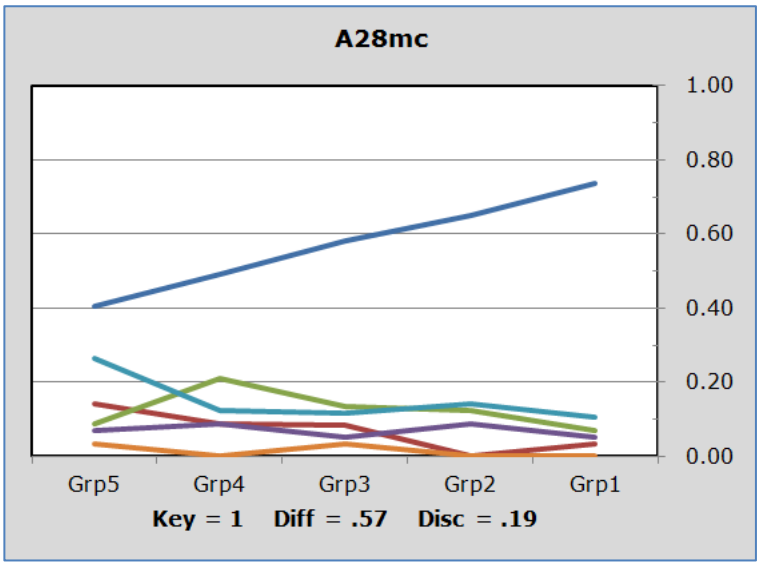
The desired pattern continues to develop here. Item A6mc would still be considered easy, but it is a better discriminator: the left-end dip for the correct answer is more pronounced, the distractors are appealing to a higher proportion of people in the lowest groups, but the top students, those in Grp2 and Grp1, are able to rule out each of them (the distractors). We could say that the distractors "zero-out" as their trace lines dip completely to touch the x-axis on the right.



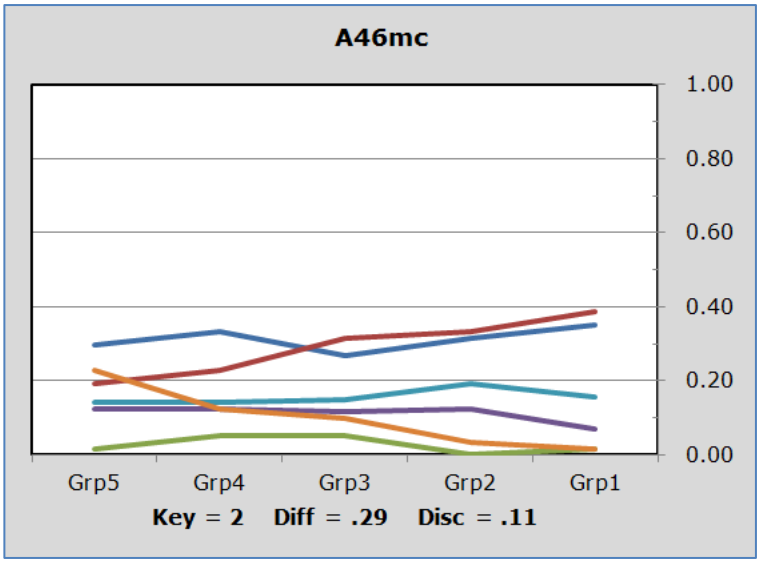
Item A17mc's pattern is strong on the left, where the trace of the correct response falls more than that seen in item A6mc. This is a good pattern, but it's not quite as strong on the right as is A6mc. A17mc's distractors dip down at the right end, as we want – however, they don't dip to zero; one or two of the distractors still appear plausible to a small proportion of students in Grp1, the best group.



Item A18mc shows a big left-end dip for the right answer, but the distractors don't fall off as much as we'd like. Still, this would be considered to be a good item if we use the classical item quality measures found in the literature: a difficulty in the range 0.40 to 0.60, and a discrimination of at least 0.30 (see Chapter 7 of the Lertap manual, Nelson (2000), for more discussion).

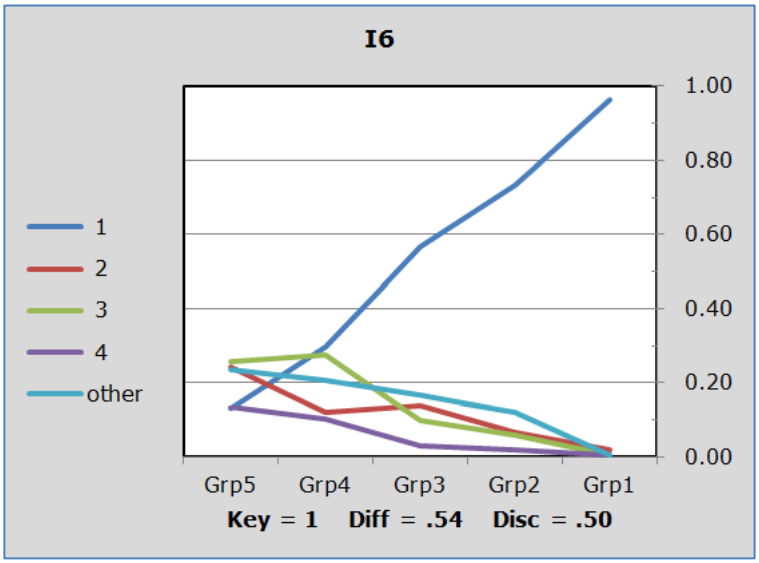


The steady rise shown by A28mc's correct answer looks good, but this item's distractors do not fall off sharply; about a quarter of the students in the upper group, Grp1, have been attracted to what are supposed to be distractors, options intended to seem plausible only to the less proficient students.



Item A46mc shows a poor pattern. The right answer rises from left to right to be the highest point for the upper group, but the rise is very slow, hardly compelling. Just over 60% of the students in the upper group were distracted by one of the incorrect answers, and the distractors do not fall off as we move from left to right. In fact, one of the distractors actually climbs a wee bit.

Let's end these sample shots with a nicely-functioning item from another test, one having four-option items, with {1,2,3,4} used as response codes:



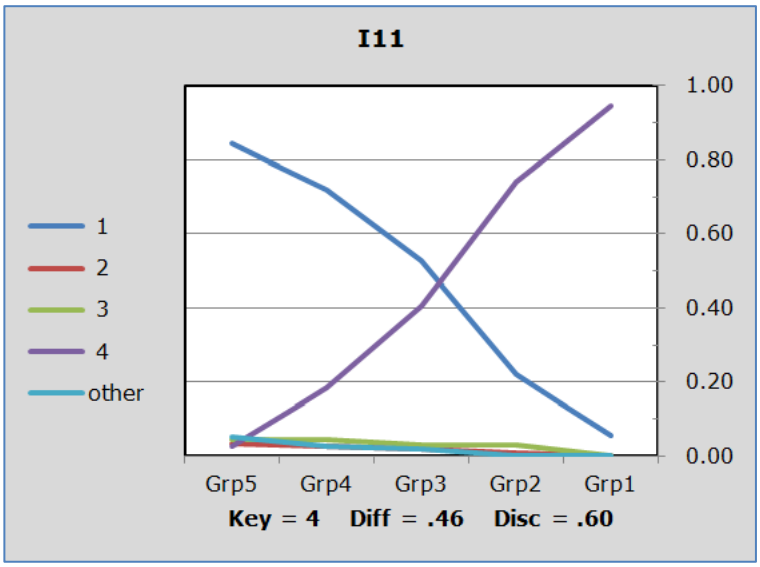
I6 exhibits a near ideal pattern for a discriminating question. The trace for the correct answer rises sharply. In Grp1, the top group, very close to 100% of the students selected the correct answer. In the weakest group, Grp5, there was indecision, with the most popular answers being two distractors, options 2 and 3.

So, then, let's see. We've looked at how many plots? Eight. What do you think? Is it easier to assess item quality by using graphs, or by using tables with numbers? The plots are kind of catching, are they not?

The Statsb and Statsf reports

The quintile plots are based on Lertap's Statsul report, as mentioned above. Lertap produces two other reports for cognitive items, Statsb and Statsf. The letters at the end of the report names mean "upper-lower", "brief", and "full", respectively.

Something which the quintiles may be somewhat poor at is pointing to distractors which have completely failed in their task, that is, distractors which have not appeared as plausible to anyone. Have a look at the following plot:



The item plotted above, I11, had three distractors. Two of them were selected by very few students, and their trace lines lie almost flat along the x-axis, rather hard to detect.

Such distractors are unwanted. In a test meant to discriminate among students, picking out the strongest from the weakest, items with non-functioning distractors need fixing. A "dead" distractor, one obviously incorrect even to weak students, increases the chances that the less knowledgeable students will get an item correct without knowing the underlying concept or fact.

Lertap's Statsb report is made to help detect poorly-functioning distractors. Here's a sample from a Statsb report:

Lertap5 brief item stats for "Form A MC", created: 16/05/2006.

Res =	1	2	3	4	5	other	diff.	disc.	?
A18mc	14%	8%	11%	11%	55%	2%	0.55	0.30	2
A21mc	26%	56%	8%	2%	8%	1%	0.56	0.26	
A22mc	55%	8%	6%	24%	6%	2%	0.55	0.39	
A23mc	19%	11%	8%	16%	44%	2%	0.44	0.46	
A24mc	12%	50%	8%	9%	17%	5%	0.50	0.43	
A26mc	16%	4%	15%	49%	11%	5%	0.49	0.27	1

The last column of the Statsb report, the one with the ? mark heading, will point out distractors which have not been seen as plausible by any of the test takers, and/or have appeared as plausible to strong students.

In this example, item A18mc has distractor 2 parked in the ? column. Looking to the left, we see that 8% of the students selected this distractor, so it's not really a "dead" distractor (one not selected by anyone). What has happened in this case is that the students who selected option 2 were above-average students.

If you look above, way above, item A18mc's quintile plot is on display. You can see distractor 2 sticking up on the right, attracting about 18% of the students in the strongest group. As we trace this distractor's performance across the graph, from left to right (cleaning our glasses, as required), we see that it was most popular with the upper group.

This is a curious result. It has served to bring down the item's discrimination index, "Disc". Item A18mc could very well be a better item if we could find out what it was that prompted almost one in five of the strongest students to select it, something we might do by talking to the students, or perhaps by extrapolating from the response patterns.

Lertap has another report with item statistics; here's a snippet:

Lertap5 full item stats for "Form A MC", created: 16/05/2006.

A18mc

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	0.00	39	0.14	-0.34	-0.53	15.51	-0.86
2	0.00	22	0.08	0.05	0.10	23.32	0.19
3	0.00	31	0.11	-0.03	-0.05	21.32	-0.08
4	0.00	33	0.11	-0.16	-0.26	18.61	-0.44
5	1.00	157	0.55	0.30	0.37	24.34	0.32
other	0.00	6	0.02	-0.11	-0.32	16.17	-0.77

The table above is from a Statsf report. Not everyone likes to look at this report as it involves so many statistics; a fair proportion of Lertap users often cannot recall what pb(r) and b(r) are, even though they fondly re-read their favourite chapters in the Lertap manual several times a week.

Not to worry; the matter worthy of note is the "avg." value of 23.32 for the 22 students who selected option 2. This is their average test score, and it's high, almost as high as the average of 24.34 for the 157 students who selected the right answer.

For this test, the overall average score, for all 288 students, was 21.92, a value found in the Scores report:

Lertap5 Scores worksh

Record No.	FA-MC
n	288
Min	5.00
Median	21.50
Mean	21.92
Max	40.00
s.d.	7.48
var.	55.90
Range	35.00
IQRange	11.00
Skewness	0.14
Kurtosis	-0.61
MinPos	0.00
MaxPos	50.00

Scores / Stats1f / Sta

As a z-score, the average of 23.32 for the 22 students selecting item A18mc's option 2 was (23.32 – 21.92)/7.48, or 0.19. This is just another reflection of what we can see in the item's quintile plot: some of the strongest students thought this was the correct answer to the item. This is an unwanted outcome; when it happens we usually suspect there may be some ambiguity in the item – there may be something that has to be fixed.

Recapping

To review, we started this discussion by looking at a Statsul report, and the quintile plots which graphically portray the information contained in the report.

Quintile plots are undoubtedly neat. With some practice they can be used to quickly get an idea of how test items are functioning. However, they're not perfect – if a distractor is "dead", not distracting anyone, its plot line will lie flat along the x-axis, and may be very difficult to detect.

The Statsb report provides a more concise indication of how distractors have functioned. A scan down the last Statsb column will immediately reveal items whose distractors may require some attention.

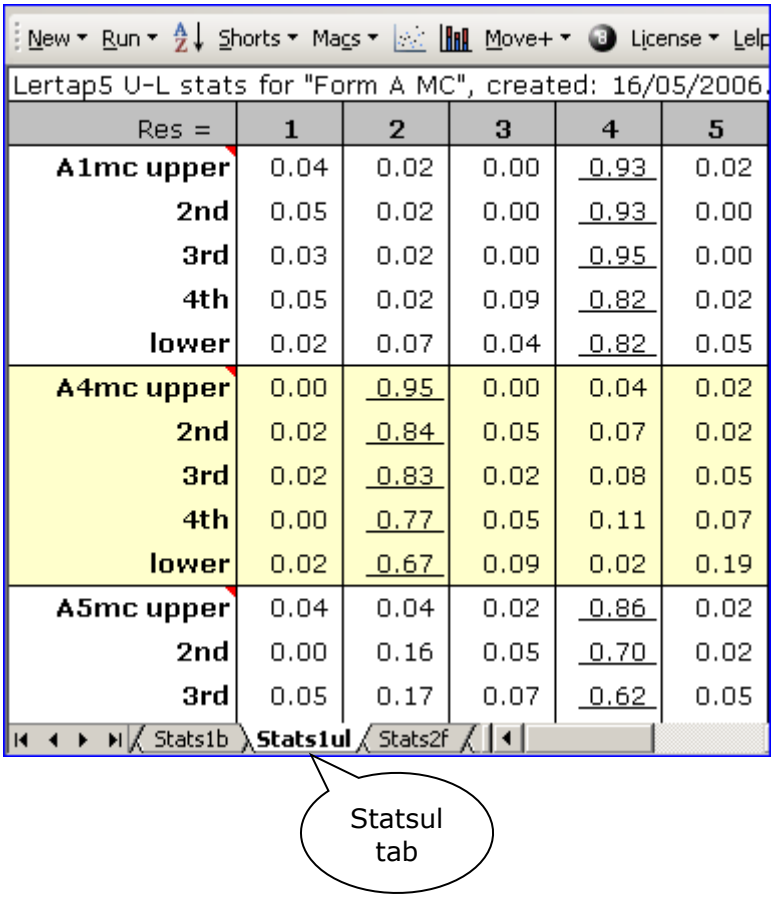
There's much more about how to use and interpret the Statsul, Statsb, and Statsf reports in the manual (Nelson, 2000).

How to get quintile plots

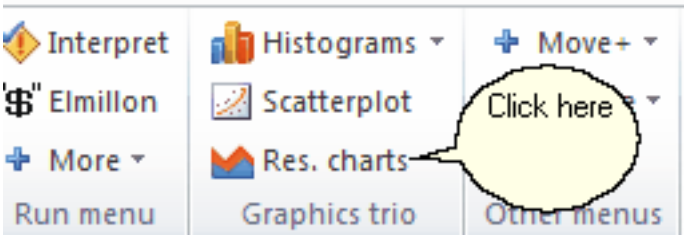
For those readers who have not used quintile plots before, it will be a good idea to mention the steps required to make them.

Once data have been prepared, and CCs control lines made, a click on Lertap's Run menu's "Interpret CCs lines" option, followed by a click on the Run menu's "Elmillon item analysis", will result in the Statsf, Statsb, and Statsul reports.

To get quintile plots, open the Statsul report by clicking on its tab:



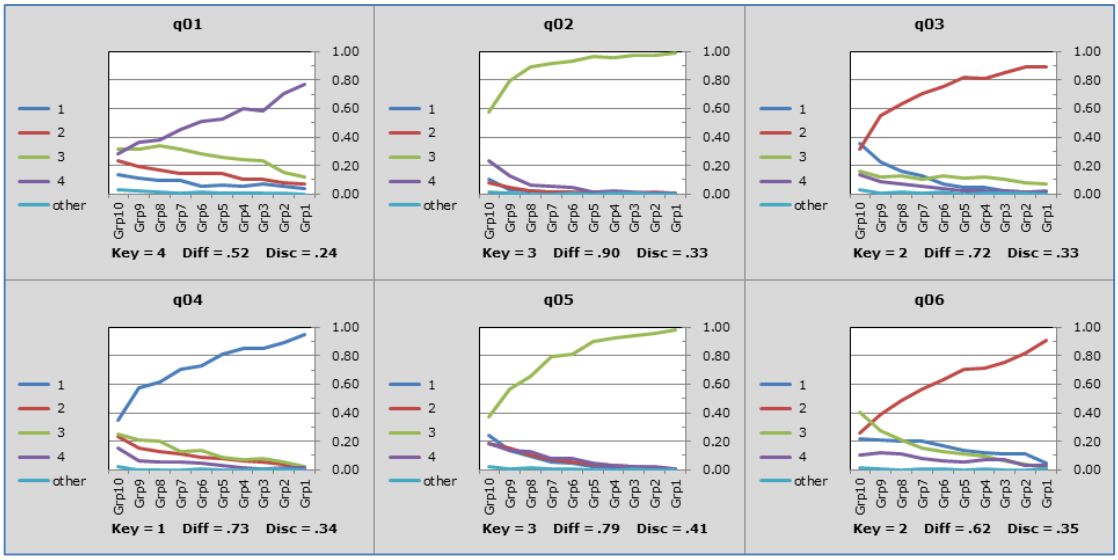
Then, with Excel 2007, Excel 2010, Excel 2013, or Excel 2016 click as shown:



Quintile plot options

There are a few options which pertain to how and when Lertap makes its quintile plots. For example, you don't have to have five score groups: you may have as few as two, or as many as 10. You can have Lertap output non-standard quintiles instead of standard ones. Having a table appended to each plot is another option.

A particularly useful option is to get Lertap to make "packed plots", a collage which has all the plots collected and presented together. Here's a small example:



Read all about the quintile plot options in "Lelp", Lertap help.

Quintile plots and item statistics

Have you been looking at the Diff. and Disc. values which appear below the x-axis in Lertap's quintile plots? These statistics have intentionally been excluded from the discussion above – we've wanted to stand back and look at the gestalt, at the whole picture, without having to get down to the nitty-gritty of statistics.

But there are some interesting observations to make.

"Diff." is item difficulty; it's the proportion of people who correctly answered an item. The higher item Diff. is, the higher the proportion of people getting the item right, the easier the item. (Diff. is actually a great misnomer; the untrained would expect high Diff. to mean high difficulty, but the opposite is true.)

Study the graphs again, looking at the pattern reflected by the trace lines, and at "Diff.". Higher Diff. values might seem to be associated with two characteristics of the plots: a relatively flat trace for the correct answer, with no sharp left-end dip.

Is this not the general case?

No, it's not. Look at item A46mc. The correct answer's trace starts at 0.20, rising to just under 0.40 at the right. Not much of a rise, a fairly flat trace. There's no left-end dip to speak of . But this is a low-Diff. item.

There's a third characteristic of high-Diff. items: their correct-answer trace starts high, and continues so across the plot. This being the situation, let's say this: an item will have high Diff. if the trace line for the correct answer is high on the left, and stays high, maybe even rising a bit as we move to the right.

Such a line will not have much slope. Slope? The slope of a line is an index of how unflat it is, to put it in unglorious terms. A flat line has no slope. A line with a slope of 1 (one) will rise at a 45-degree angle from left to right.

A high-Diff. item, an easy item, is one whose correct-answer trace starts high on the left, and is fairly flat, having little slope.

With regard to the item discrimination index displayed at the bottom of the plots, "Disc.", what pattern is seen in the plots?

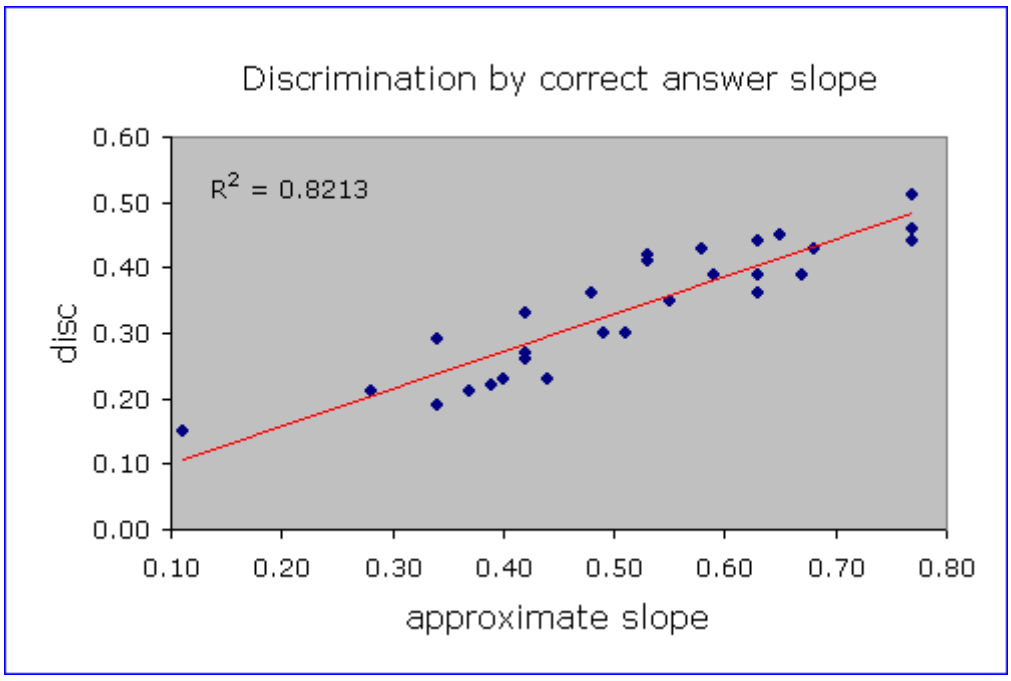
The greater the slope for the correct option's trace line, the greater Disc.

Is this right? Let's see. Let's approximate the slope of the correct answer's trace by subtracting the proportion for the lower group from the proportion for the upper group. Here's a little list, including respective Disc. values.

```
For  A1mc we'd have 0.96 - 0.81, or 0.15.  Disc. = 0.15.

For A46mc we'd have 0.38 - 0.20, or 0.18.  Disc. = 0.11.

For  A4mc we'd have 0.98 - 0.67, or 0.31.  Disc. = 0.21.

For A28mc we'd have 0.73 - 0.40, or 0.33.  Disc. = 0.19.

For A18mc we'd have 0.78 - 0.27, or 0.51.  Disc. = 0.30.

For  A6mc we'd have 1.00 - 0.41, or 0.59.  Disc. = 0.43.

For A17mc we'd have 0.97 - 0.30, or 0.67.  Disc. = 0.45.

For   Q51 we'd have 0.89 - 0.19, or 0.70.  Disc. = 0.45.
```

Is there not a pattern? As the approximation of slope increases, the Disc. values tend to increase.

A scatterplot of the approximate slope index by Disc. value for the first thirty multiple-choice items in the aptitude test is shown below[2]; the line superimposed on the plot is a "regression line", a "trend line", which lets us see the general drift of the relationship between Disc. and approximate slope:



[2] Only the first 30 items were studied as this was a "speeded test"; there was insufficient time for many of the students to work beyond the 30th item.

There is a relationship, at least for the selected items: higher Disc. values are associated with higher slope estimates.

The greater the slope for the correct answer's trace line, the higher the Disc. value.

This is only to be expected. The Disc. statistic seen in Lertap is a point-biserial correlation coefficient formed by correlating student test scores with a select / not-select index for the correct answer. If a student selected the correct answer, we could give him/her a select / not-select "score" of 1 (one); if s/he did not select the correct answer, we could give a select / not-select score of 0 (zero). We'd then correlate the select / not-select scores with test scores.

Higher correlations will result when those selecting the correct answer have the highest test scores. In terms of a standard Lertap quintile plot, those with highest test scores are in the "upper" group. The number in this group with a 1 on the select / not-select index can be found by multiplying group n by group proportion; call this n(high). At the other end, the left end of the plot, we have the students with the lowest test scores. The number in this group with a 1 on the select / not-select index is again group n times group proportion; call this n(low). If n(low) equals n(high), the trace for the correct option is going to have little if any slope, and the correlation between the select / not-select index and test score is going to be low.

So, what might be concluded? *Items having a correct-answer trace which has a sharp dip at the left end, and rises steadily to approach the top of the plot at the right, will have a high discrimination index.*

Quintile plots with a demographic variable

Note: the information below remains useful, but Lertap now has additional, vastly extended capabilities for comparing group responses to items, including support for "DIF", differential item functioning. Please refer to this site.

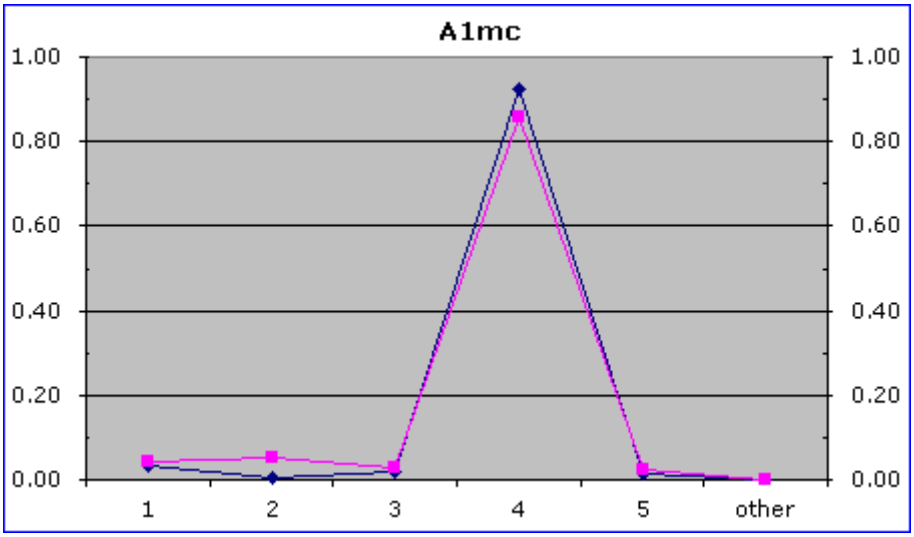
When conditions are right[3], Lertap's quintile plots may be used to compare the way two, three, four, or five groups have responded to an item.

The high school aptitude test featured in most of the plots seen above involved 288 students, 140 girls and 148 boys. A gender code was entered in the third column of the Data worksheet, with girls=1 and boys=2. We can use this column as a basis for the plots.

To do so, we go to the Run menu, and request an "External criterion analysis", asking it to use the gender codes found in the third column as the stratifying variable for quintile plots.

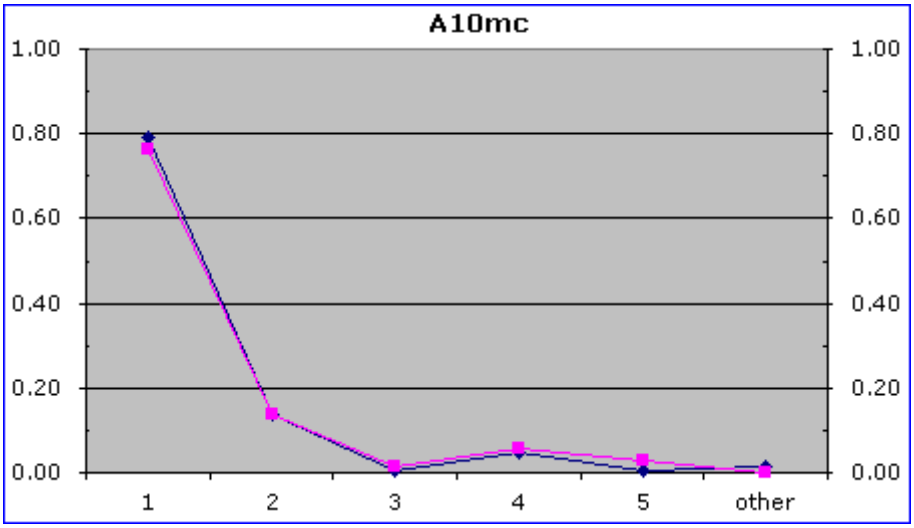
We got Lertap to make non-standard quintile plots. Look at some of the results:

[3] There are two main conditions: the external criterion must be numeric, and should have no more than five discrete values. Lertap and Excel permit demographic variables to be recoded so that these conditions are met.

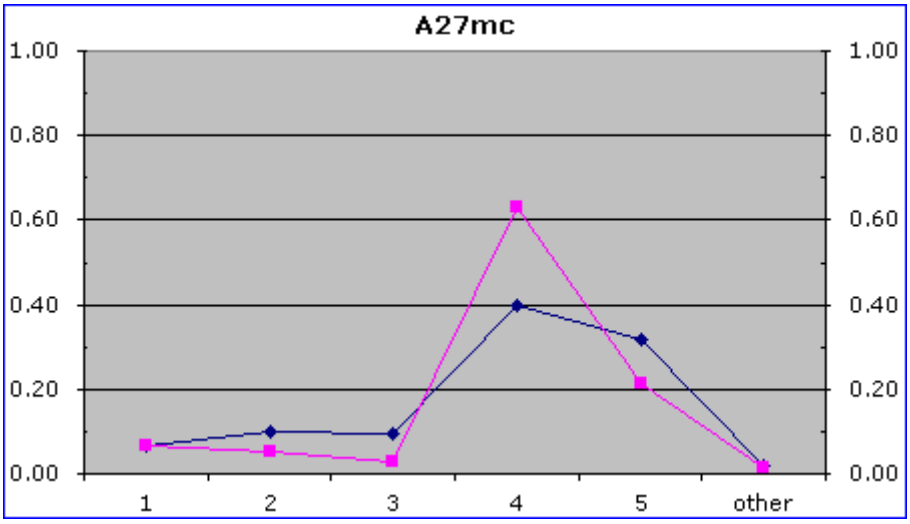


There are only two lines in the plot, one for girls (pink), one for boys (blue).

The correct answer to item A1mc is 4, selected by 86% of the girls, and 93% of the boys. The response pattern is similar – there do not seem to be any substantial differences by gender.



There are even smaller gender differences in the responses to item A10mc; the lines are almost overlapping across the graph, with about 79% of the boys getting this item correct, and 76% of the girls.



Here, at item A27mc, gender differences seem evident: the lines for the two groups diverge, especially at response 4, the right answer. Of the boys, 40% got the item right, compared to 63% of the girls.

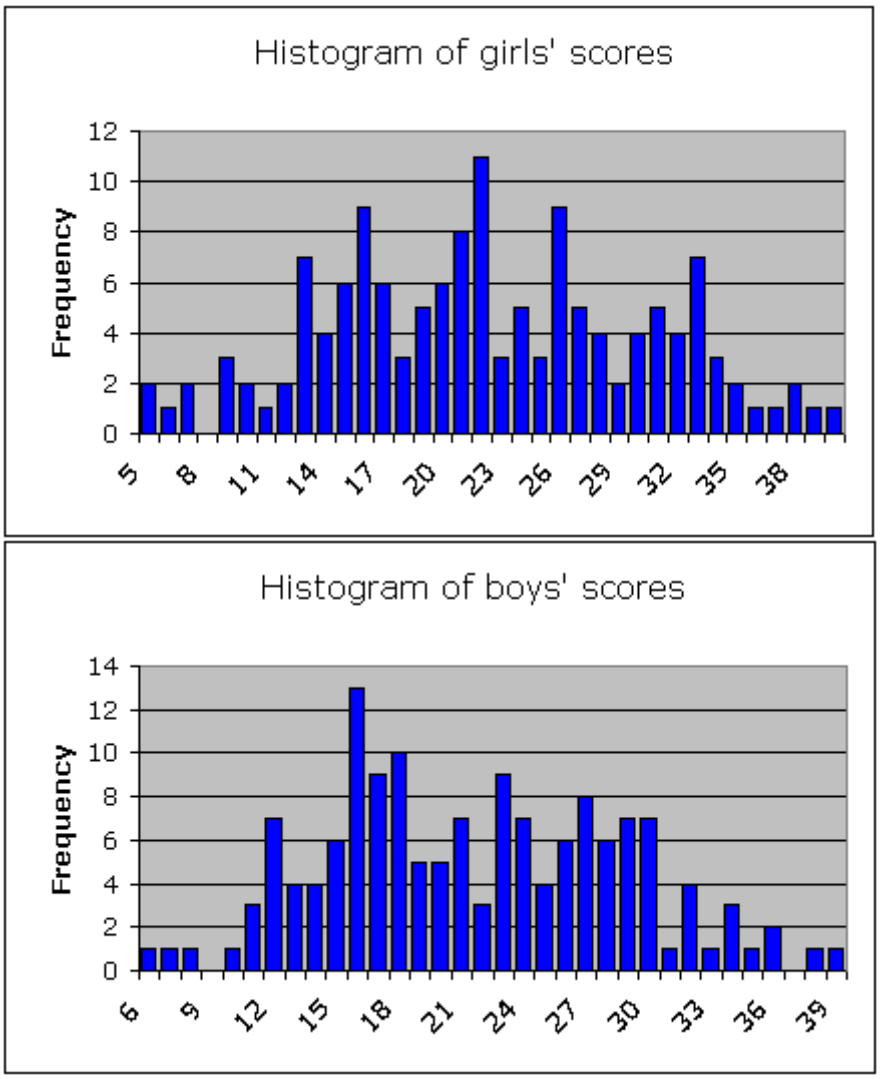
Distractor 5 fooled a higher percentage of boys than girls, 32% compared to 21%.

Plots such as these make it very easy to spot items which appear to be gender sensitive.

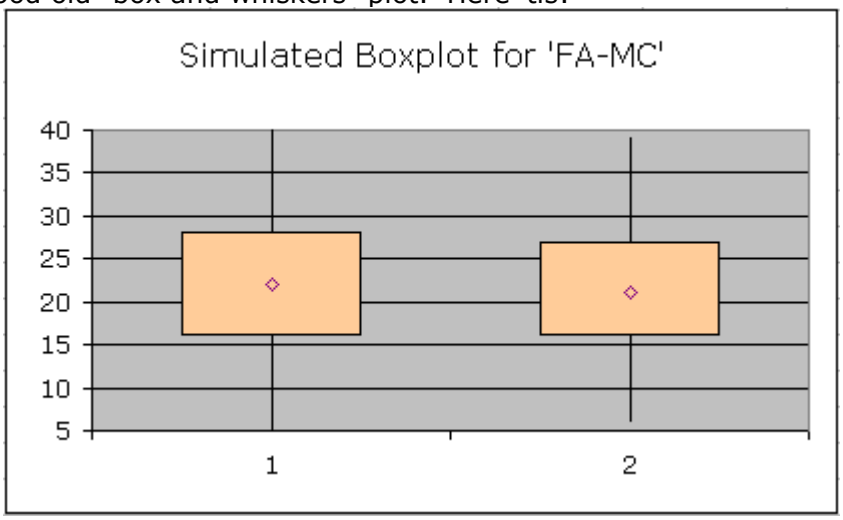
We probably would not usually want to have an aptitude test which is gender sensitive. To see if there were gender difference on the overall test score, we could use Lertap's Run menu to request "Breakout scores by groups":

FA-MC	1	2
n	140	148
Min	5.00	6.00
Median	22.00	21.00
Mean	22.22	21.63
Max	40.00	39.00
s.d.	7.97	6.96
var.	63.53	48.50
Range	35.00	33.00
IQRange	12.00	11.00
Skewness	0.06	0.20
Kurtosis	-0.65	-0.63
MinPos	0.00	0.00
MaxPos	50.00	50.00

The overall mean for the girls (code 1) is slightly higher than that for the boys (code 2), but there's not much in this little table. We might get a better picture by having Lertap make score histograms:



Another way to look at potential group differences would be to get Lertap to make a good old "box and whiskers" plot. Here 'tis:



Box and whiskers plots, also know as "boxplots", are a creation of John Tukey (1977). Like quintile plots, they present a great deal of information without leaning heavily on numbers.

The plot above indicates that test scores were slightly more spread out for girls (code 1). This is so as the box for the girls' scores is slightly taller than the one for boys, and because the whiskers for the girls, the lines stemming from the top and bottom of the box, extend further than the boys' whiskers (obviously the girls had not been using an epilator).

But, basically, the boxplot indicates that the groups do not differ much. Even though there seemed to be meaningful gender differences on item A27mc, over the whole test there appeared to be gender balance[4].

For more about using an external criterion for quintile plots, making boxplots, and getting histograms, please refer to "Lelp", Lertap help.

Visually yours?

In conclusion, Lertap 5 provides test developers and users with two fundamental methods for item analysis: a series of Stats reports with item statistics presented in different ways (Statsul, Statsb, Statsf), and a series of pictures, the quintile plots.

Which of these methods is better? The reports, as a group, give more information than the pictures do. The Statsf report, for example, has summary test statistics not found in the pictures, such as reliability, and the standard error of measurement estimate. The Statsb report has a single column devoted to highlighting weak distractors.

However, for assessing the performance of an individual item, the pictures have much to say for themselves. They obviate the need to understand statistical measures, such as "Diff." and "Disc.". They quickly and colourfully reveal how all item options worked.

Is this visual approach to item analysis a serious research method? Yes. Without a doubt. Is a quintile picture worth a thousand of Lertap's retinue of Stats reports? Hmmm ….

[4] A one-way analysis of variance table, not shown here, is also standard output (with effect-size estimator, in addition to the usual F-ratio).

References

Crocker, L.M. & **Algina**, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Du Toit, M. (Ed.) (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood (IL): Scientific Software International.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items.* New Jersey: Lawrence Erlbaum Associates. (See pp.221-223.)

Nelson, L.R. (2000). *Item analysis for tests and surveys using Lertap 5.* Perth, Western Australia: Curtin University of Technology (www.lertap.curtin.edu.au).

Tukey, J.W. (1977). *Exploratory data analysis.* Reading, Massachusetts: Addison-Wesley.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement,* 26, 191-208.

Larry R Nelson, PhD

Curtin University
Perth, Western Australia

Burapha University
Chonburi, Thailand

www.lertap.com