Using Practical Exhibits to Present Selected Measurement Topics

This Exhibit: “Test13”

Larry R Nelson[1]

Dated: 12 September 2022 (check for update)

● About Test13

Test13 is a 13-item multiple-choice cognitive test having arithmetic items written decades ago, well before electronic calculators were developed. The items in Test13 were intended to assess how well primary school students had mastered selected objectives and presumably were based on, and drawn from, a maths syllabus.

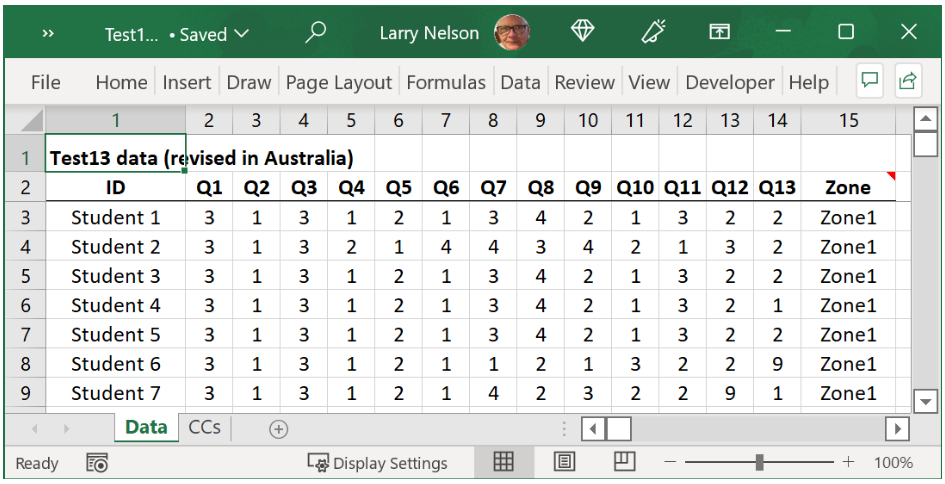
The actual items are seen in Appendix A. Each item presents a problem to solve and offers four possible answers. Students were instructed to circle the answer they considered to be correct.

● Terminology

The question, or problem to solve, is said to be the item’s “stem”. As seen in Appendix A, the stem for the first item is “Solve the following, 7097 + 1903”.

There is just one right answer to each item, referred to as the “keyed-correct” response. The incorrect answers are known as “distractors”. In Test13, possible item responses are 1,2,3,4 for each item. We will be using the Lertap5 app to analyse Test13 results and, in Lertap5, res=(1,2,3,4) is the way this test’s items will have their valid response codes specified (seen in the “CCs” lines on p.11).

● Student responses recorded in an Excel workbook



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**Test13 data (revised in Australia)**														
2	**ID**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**	**Q13**	**Zone**
3	Student 1	3	1	3	1	2	1	3	4	2	1	3	2	2	Zone1
4	Student 2	3	1	3	2	1	4	4	3	4	2	1	3	2	Zone1
5	Student 3	3	1	3	1	2	1	3	4	2	1	3	2	2	Zone1
6	Student 4	3	1	3	1	2	1	3	4	2	1	3	2	1	Zone1
7	Student 5	3	1	3	1	2	1	3	4	2	1	3	2	2	Zone1
8	Student 6	3	1	3	1	2	1	1	2	1	3	2	2	9	Zone1
9	Student 7	3	1	3	1	2	1	4	2	3	2	2	9	1	Zone1

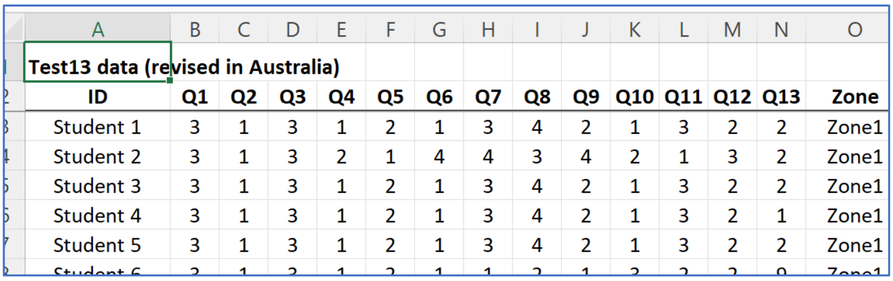
Figure 1

Figure 1 displays the initial nine records in an Excel workbook with the item responses given by seven students. Column 15 has a code indicating the geographic zone a student resided in at the time of the test – there were four zones in the country in which Test13 was administered.

[1] l.nelson@curtin.edu.au

Of the seven students whose item responses are seen in Figure 1, note that two students, the 6th and the 7th, have an item response of “9”. In this dataset, a 9 was used to indicate a missing response. Student 6 did not answer Q13. Student 7 did not answer Q12.

The columns in Figure1 are numbered. In Excel terminology, the “R1C1” reference style has been activated in this case. Some Excel users will find Excel to be using letters instead of numbers to label columns, with A being the first column, a reference style referred to as “A1”. See Figure 2 for an example. The Excel-based data analysis app used in this document, Lertap5, has an inbuilt shortcut for switching from R1C1 to A1 – see “Ref. style”. The R1C1 reference style is preferred when working with Lertap5. This page from Microsoft shows another way to change the reference style.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Test13 data (revised in Australia)														
ID	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**	**Q13**	**Zone**
Student 1	3	1	3	1	2	1	3	4	2	1	3	2	2	Zone1
Student 2	3	1	3	2	1	4	4	3	4	2	1	3	2	Zone1
Student 3	3	1	3	1	2	1	3	4	2	1	3	2	2	Zone1
Student 4	3	1	3	1	2	1	3	4	2	1	3	2	1	Zone1
Student 5	3	1	3	1	2	1	3	4	2	1	3	2	2	Zone1
Student 6	3	1	3	1	2	1	1	2	1	3	2	2	9	Zone1

Figure 2

● Lertap5

Lertap5 is an Excel-based app used to analyse tests and surveys. An introduction to Lertap5 is found at this website. The use of Lertap5 is addressed in a number of places. A recommended start is this PowerPoint presentation (also available as a PDF). Another suggested guide is the “Cook’s Tour”.

To follow the steps presented in this document readers will need to (1) have Excel running on their computer (Windows or Macintosh); (2) install Lertap5; and (3) get a copy of the Test13 dataset (an Excel workbook).

This web page has a link to download the Lertap511.zip file. Within this zip file are three Excel files which, taken together, constitute the Lertap5 data analysis system (or “app”). The zip also contains a few PDF documents. All of these files are extracted from Lertap511.zip by unzipping it using the instructions found on the web page.

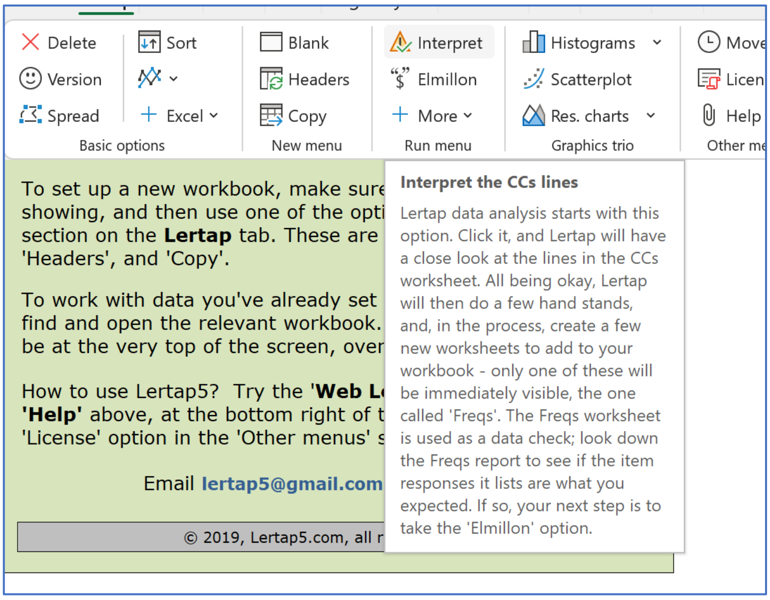
A copy of the Test13 Excel workbook used in this discussion may be downloaded from this link.

● Get “Freqs”

We’ll begin by getting Lertap5 to summarise item response frequencies for Test13.

To do so we need to have Lertap5 open and running in Excel, and we also need to have opened that copy of Test13 results.

After opening Lertap5 and the workbook with Test13 results, click on the “Interpret” option seen in the Lertap5 “tab” in Excel; refer to Figure 3.



To set up a new workbook, make sure showing, and then use one of the opti section on the **Lertap** tab. These are 'Headers', and 'Copy'.

To work with data you've already set find and open the relevant workbook. be at the very top of the screen, over

How to use Lertap5? Try the '**Web L** '**Help**' above, at the bottom right of t 'License' option in the 'Other menus' s

Email **lertap5@gmail.com**

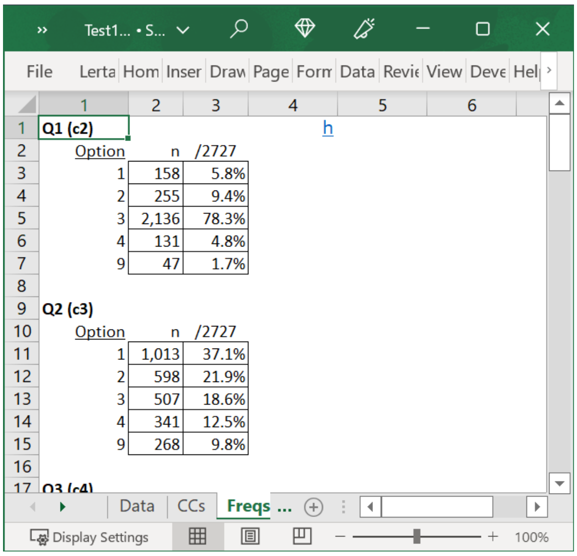
© 2019, Lertap5.com, all r

Interpret the CCs lines

Lertap data analysis starts with this option. Click it, and Lertap will have a close look at the lines in the CCs worksheet. All being okay, Lertap will then do a few hand stands, and, in the process, create a few new worksheets to add to your workbook - only one of these will be immediately visible, the one called 'Freqs'. The Freqs worksheet is used as a data check; look down the Freqs report to see if the item responses it lists are what you expected. If so, your next step is to take the 'Elmillon' option.

Figure 3

The Interpret option adds a worksheet called “Freqs” to the Test13 workbook. Figure 4 displays response frequencies for the first two test items.



	1	2	3	4	5	6
1	Q1 (c2)				h	
2	Option	n	/2727			
3	1	158	5.8%			
4	2	255	9.4%			
5	3	2,136	78.3%			
6	4	131	4.8%			
7	9	47	1.7%			
8						
9	Q2 (c3)					
10	Option	n	/2727			
11	1	1,013	37.1%			
12	2	598	21.9%			
13	3	507	18.6%			
14	4	341	12.5%			
15	9	268	9.8%			
16						
17	Q3 (c4)					

Figure 4

We see that 78.3% of the students chose 3 as their answer to Q1. This was the keyed-correct response for that item. Of interest is the considerable number of students who did not answer the second item, Q2 – Figure 4 shows that almost 10% of the students did not respond to this item. The keyed-correct response for Q2 was 1.

We’d be inclined to say that Q1 was easy as so many students (78%) got it right, but Q2 was <u>not</u> easy; less than half of the students (37%) were able to correctly answer Q2.

Note the small underlined blue h towards the top of Figure 4 – if clicked on it will open a help topic with an explanation of the contents of a Freqs “report”. (See that help topic here.)

At this stage we would want to check on the response frequencies for items Q3 through Q13. In most cases, we will probably want to make sure that the number of students failing to answer each item is well below 10% or so. If this is not the case, we might think that the test items may not have been appropriate for the students who were assessed.

● Get item statistics and student scores with “Elmillon”

Elmillon is the name of the main item analysis routine in Lertap5. The meaning of the name is given here. In Figure 3 the Elmillon option is seen immediately below the Interpret option.

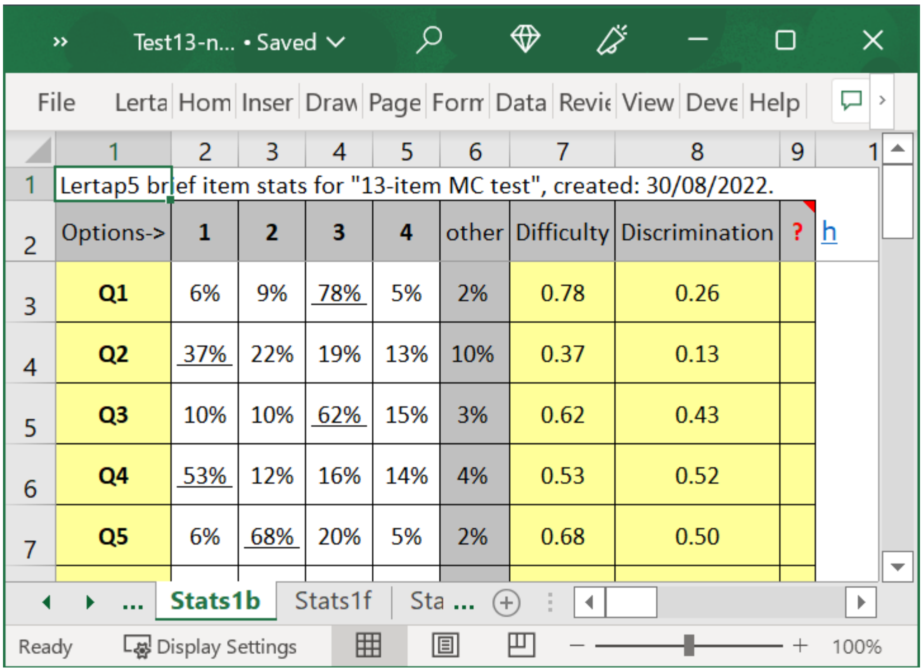
A classical item analysis for a cognitive test, such as Test13, will tell us how difficult each item was, how the responses students gave to an item correlated with their responses on the other items (called the “item discrimination”), and provide an estimate of the reliability of the test (how much measurement error might there be in the test scores?).

I’ll use Elmillon now. When it finishes it will have added four new worksheets: Stats1b, Stats1f, Stats1ul, and Scores.

These three Stats worksheets all have to do with item statistics, looking at item performance from different perspectives. I’ll pose a few questions to guide a discussion of the steps commonly found in classical item and test analyses for cognitive items.

Q1: Which items were the easiest / which the most difficult?

We could answer this question just by using the Freqs report (Figure 4). But Stats1b makes it much easier, have a look at Figure 5:



Lertap5 brief item stats for "13-item MC test", created: 30/08/2022.

Options->	1	2	3	4	other	Difficulty	Discrimination	?	h
Q1	6%	9%	78%	5%	2%	0.78	0.26		
Q2	37%	22%	19%	13%	10%	0.37	0.13		
Q3	10%	10%	62%	15%	3%	0.62	0.43		
Q4	53%	12%	16%	14%	4%	0.53	0.52		
Q5	6%	68%	20%	5%	2%	0.68	0.50		

Figure 5

The “Difficulty” column in a Stats1b report (or worksheet) gives the proportion of students getting each item correct. Of the five Test13 items seen in Figure 5, the easiest item was Q1; Q2 was the most difficult. Columns 2 through 5 repeat the information found in the Freqs report but now it’s more obvious which options were the distractors, and which was the item’s keyed-correct response:

the keyed-correct option is underlined. The “Other” column corresponds to the percentage of students who did not answer an item.

To be noted is that difficult items are those having the lowest proportion correct. Q2 was more difficult than Q1, 0.37 vs 0.78. A high item difficulty figure will actually be an easy item. This is obviously the reverse of what would be expected, but this terminology dates way back to the beginning of CTT, classical test theory, and is now embedded.

Over the years some have proposed solutions to this semantic anomaly by, for example, changing the label to “item facility”, but, except in rare cases (such as “MOODLE”) , this suggestion has not been adopted.

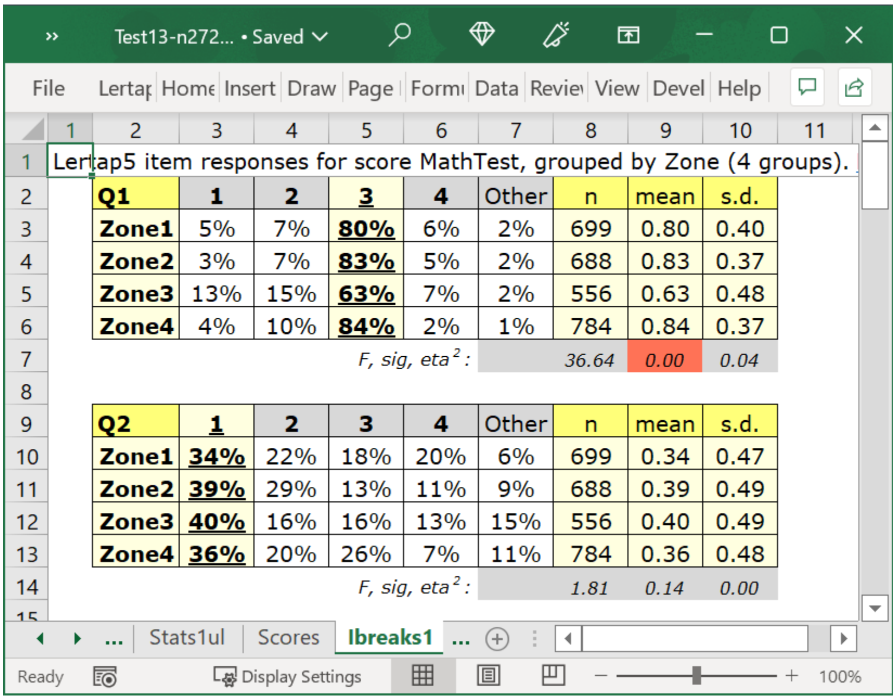
Answering Q1: Which items were the easiest / which the most difficult? Figure 5 displays results just for the first five of Test13’s items. Of those five, the easiest was Q1 (difficulty 0.78), the most difficult, the hardest, was Q2 (difficulty 0.37). It turns out that none of the other eight items were easier or harder than these two.

There are times when our analysis of test results may end with Stats1b. This may happen, for example, when a teacher simply wants to identify problem areas, topics which may need going over again with their class. In this example, Q2 stands out as a possibility as only 37% of the students got it right.

Now, at the same time, a coordinator at a national maths curriculum centre might look at Figure 5 and like to have a breakdown of results over all of the country’s geographic zones.

This can be done using the code found in column 15 of the Test13 workbook (see Figure 1).

The “Item responses by groups” option in Lertap5 will do what the coordinator wants. It will produce the “Ibreaks1” report as captured in Figure 6.



Lertap5 item responses for score MathTest, grouped by Zone (4 groups).

Q1	1	2	3	4	Other	n	mean	s.d.
Zone1	5%	7%	**80%**	6%	2%	699	0.80	0.40
Zone2	3%	7%	**83%**	5%	2%	688	0.83	0.37
Zone3	13%	15%	**63%**	7%	2%	556	0.63	0.48
Zone4	4%	10%	**84%**	2%	1%	784	0.84	0.37
					F, sig, eta^2 :	*36.64*	*0.00*	*0.04*

Q2	1	2	3	4	Other	n	mean	s.d.
Zone1	**34%**	22%	18%	20%	6%	699	0.34	0.47
Zone2	**39%**	29%	13%	11%	9%	688	0.39	0.49
Zone3	**40%**	16%	16%	13%	15%	556	0.40	0.49
Zone4	**36%**	20%	26%	7%	11%	784	0.36	0.48
					F, sig, eta^2 :	*1.81*	*0.14*	*0.00*

Figure 6

The results in Figure 6 suggest a possible problem in Zone 3. The students tested in this zone were noticeably less proficient on item Q1 – only 63% got the item right in Zone 3 compared to 80% or more in the other zones.

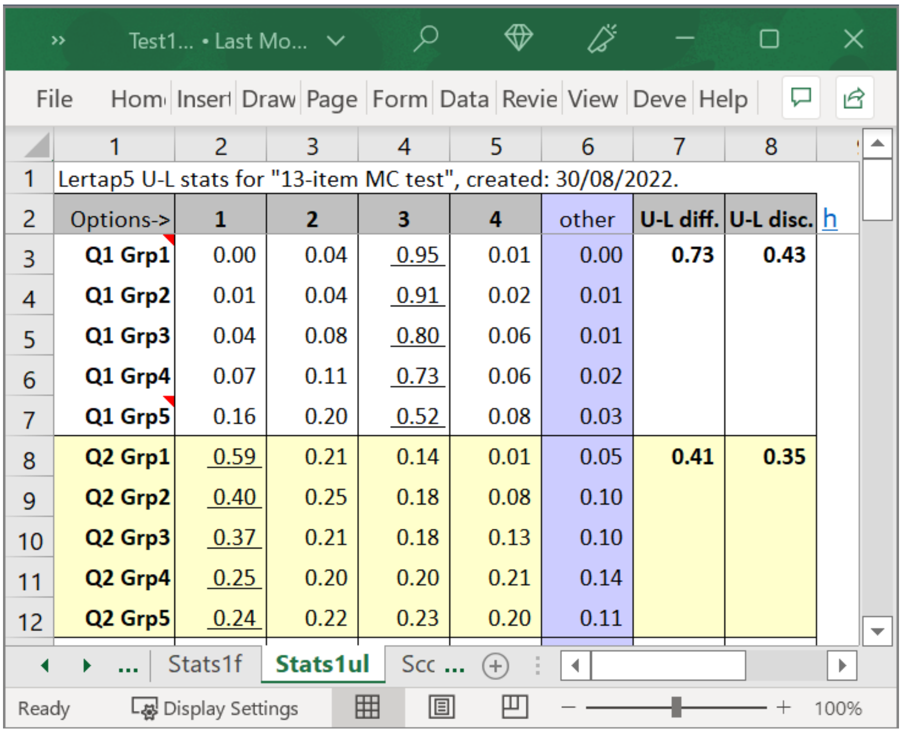
Q2: Another research question, a very common one: which of the 13 items can be used to pick out the strongest students?

As we know, often a test is used so that we have some means of identifying the "best", or the "strongest", students. When this is an objective, we want to have items with a demonstrated ability to discriminate, to give us the means of picking out the cream of the crop, and/or, to perhaps be able to single out students who might need remedial work.

One way to index the discriminating ability of an item is to have beforehand a sample of students with known proficiency levels. We might have, for example, two proficiency levels in the sample: known low achievers and known high achievers. A discriminating item will be one correctly answered by only the strongest students.

However, it is not all that common to have a student sample with known proficiency levels. What Lertap5 and other item analysis packages do is create a test score for all students and use that score to identify achievement/proficiency levels. We'll give the test, score it, sort the scores from highest to lowest, and then, for example, pick out the top 20% of the scores, the next 20%, the middle 20%, the next-to-bottom 20%, and the bottom 20%. Then for each item we'll say that we have a discriminating item when only the top students get the item right.

The Stats1ul report in Lertap5 does this. Figure 7 is an example of results for two of Test13's items, Q1 and Q2.



Lertap5 U-L stats for "13-item MC test", created: 30/08/2022.

Options->	1	2	3	4	other	U-L diff.	U-L disc.	h
Q1 Grp1	0.00	0.04	0.95	0.01	0.00	0.73	0.43	
Q1 Grp2	0.01	0.04	0.91	0.02	0.01			
Q1 Grp3	0.04	0.08	0.80	0.06	0.01			
Q1 Grp4	0.07	0.11	0.73	0.06	0.02			
Q1 Grp5	0.16	0.20	0.52	0.08	0.03			
Q2 Grp1	0.59	0.21	0.14	0.01	0.05	0.41	0.35	
Q2 Grp2	0.40	0.25	0.18	0.08	0.10			
Q2 Grp3	0.37	0.21	0.18	0.13	0.10			
Q2 Grp4	0.25	0.20	0.20	0.21	0.14			
Q2 Grp5	0.24	0.22	0.23	0.20	0.11			

Figure 7

In a Stats1ul report, the top 20% is referred to as "Grp1", the bottom as "Grp5". The correct answer to an item will have its proportion-correct figures underlined.

We would say that Q1 has shown some ability to discriminate among the students tested. In Grp1, the group with the highest test scores, 95% got Q1 correct, dropping to 52% in Grp5, the lowest group.

The **U-L disc.**[2] statistic for Q1 is 0.43, being 0.95 (Grp1's proportion correct), minus 0.52 (Grp5's proportion correct). Q2's value isn't that far behind but note that the proportion correct for the best group is only 0.59 – students found Q2 to be a difficult item.

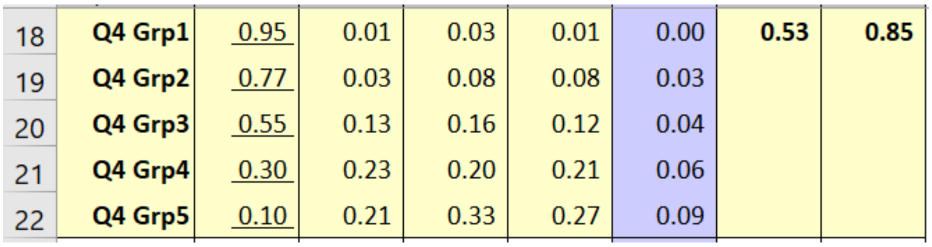
A quick way to assess item discrimination in a Stats1ul report is to scan down the **U-L disc. col**. It turns out that the best discriminator was Q4 with a value of 0.85 – 95% of the strongest students answered Q4 correctly, going down to just 10% in the bottom group. See Figure 8.

The two little plots in Figure 9 demonstrate item discrimination. The blue lines trace the proportion correct for each of the five groups while the other lines trace the proportions for the three distractors; the light-blue line with asterisks tracks the proportion of missing data for each item over the five groups.

These are Lertap5 "quintile plots". These plots are rather striking examples of what item discrimination means. Q4 is a much better discriminator than Q2 as the trace line for the correct answer (the blue line in each plot) begins at a low value on the left and rapidly rises to nearly reach the maximum possible value of 1.00 on the right. The slope of Q4's correct answer trace line is much greater than the corresponding trace line for Q2's correct answer.

The Grp5 proportions for both items, as seen in the plots, are rather similar. But note (Figure 9) how the proportions for Q4's distractors fall to just about zero as we move left to right across the plot, from the weakest students to the strongest (that's not the case for Q2).

As seen in the plots, in the case of Q4, almost all students in the top group, Grp1, were able to pick the correct answer while in the case of Q2 some students in the top group selected one of the item's distractors – that's not desired if test items are meant to discriminate the strong from the weak – we don't expect the top students to select an incorrect answer (a distractor).



18	**Q4 Grp1**	0.95	0.01	0.03	0.01	0.00	**0.53**	**0.85**
19	**Q4 Grp2**	0.77	0.03	0.08	0.08	0.03		
20	**Q4 Grp3**	0.55	0.13	0.16	0.12	0.04		
21	**Q4 Grp4**	0.30	0.23	0.20	0.21	0.06		
22	**Q4 Grp5**	0.10	0.21	0.33	0.27	0.09		

Figure 8

[2] U-L for "upper-lower"

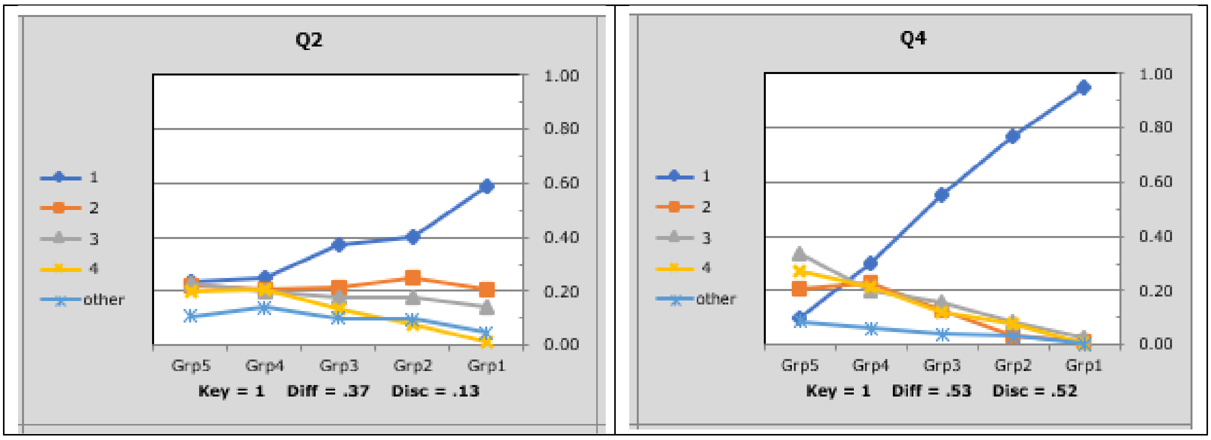
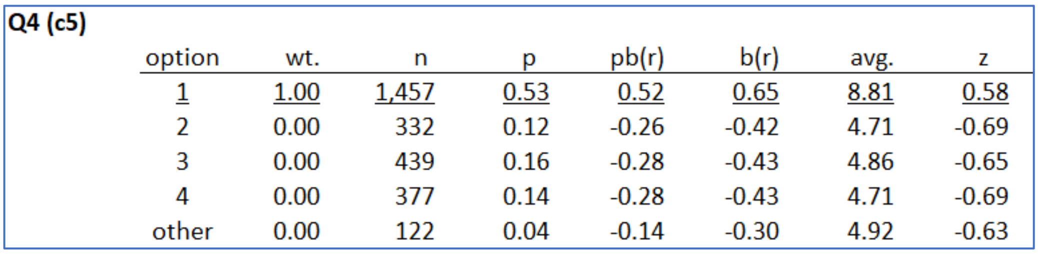


Figure 9

The Diff and Disc figures seen at the base of each quintile plot are the classic measures of item difficulty and discrimination found in Stats1b reports (see Figure 5).

In turn, the difficulty and discrimination figures in Stats1b are really lifted from the most detailed of Lertap5's various statistical reports, Stats1f.

Figure 10 displays Stats1f's results for Q4.



Q4 (c5)

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	1.00	1,457	0.53	0.52	0.65	8.81	0.58
2	0.00	332	0.12	-0.26	-0.42	4.71	-0.69
3	0.00	439	0.16	-0.28	-0.43	4.86	-0.65
4	0.00	377	0.14	-0.28	-0.43	4.71	-0.69
other	0.00	122	0.04	-0.14	-0.30	4.92	-0.63

Figure 10

The p column in Figure 10 gives the proportion of students selecting each of the item's options. The p value for the keyed-correct option (underlined) is the classic index of item difficulty. The pb(r) column is the point-biserial correlation of each option with the total test score and is the classic index of item discrimination. The avg. figure corresponds to the average test score for all those students who selected each option. When an item is a discriminating one, the avg. for the keyed-correct option (underlined) will be greater than the avg. values for each of the distractors.

This webpage has much more information about Stats1f.

Answering Q2: Which of the 13 items can be used to pick out the strongest students?

The question is asking for the item having the greatest discrimination value. We can answer it by scanning down the Discrimination column in the Stats1b report (Figure 5), or by scanning down the U-L disc column in Stats1ul (Figure 7).

Q4 had the highest discrimination (0.52) in the Stats1b report, and also the highest U-L disc value (0.85) in Stats1ul. These figures differ as they estimate discrimination in different ways. The pb(r) point-biserial correlation computed by Lertap5 and found in Stats1f and Stats1b reports is by far the most common classic discrimination index.

A recommended reference is Chapter 7 of the Lertap user manual; page 10 discusses pb(r). This reference has more information about quintile plots, such as those see in above in Figure 9.

Not mentioned to this point is the importance of item distractors. They will always be present in multiple-choice tests. Their job is to pose plausible alternatives to the correct answer. If they fail in this crucial task, the chance of weak students getting an item right increases.

Both the Stats1b report and the Stats1f report have “flags” that appear whenever an item’s distractors appear to have failed. See this topic.

Q3: What about Test13’s reliability?

The reliability of a measuring instrument has to do with its accuracy. If I measure something twice, are the two measures exactly the same? For example, if I use a scale to measure the weight of a package of apples purchased at the supermarket, record the weight, remove the bag from the scale, and then weigh it again, how close are the two readings? How consistent is the scale and my use of it? If I use a different scale, will it agree?

Similar questions pertain to the use of tests and surveys. If I were to retest the students who answered Test13 yesterday, would they get the same test scores?

They may very well do so. They may remember the items and the answers they gave yesterday.

That might be likely for many of the students, but let’s imagine this: I change Test13 items so that they’re not exactly the same.

Item1’s stem in Appendix A is 7097 + 1903. I’ll change it to 1952 + 7048. Then I’ll also alter the other items so that the exercises are the same but not *precisely* identical. I’d most likely change the distractors too. Once I have done this, I’ll have what’s called a parallel version of Test13, known as a “parallel form”.

I’ll have the same students answer my parallel form and then compare results with the original test scores. In doing so I will be said to be looking at the parallel forms reliability of Test13. The test’s reliability will be the correlation between the two test scores. Perfect reliability will be a correlation of 1.00 .

Parallel forms are not necessarily easy to create[3]. The example I’ve suggested above for Item 1 may seem reasonable, but that would not likely be the case for many of the other items. So, what I could do instead is divide the test into two parts, correlate the scores from each part and take that as an initial estimate of reliability, an approach known as “split-halves”.

There are some item analysis apps that output split-half reliability estimates, but not many[4]; coefficient alpha, also known as Cronbach’s alpha, is the preferred and universal method used to derive a reliability estimate for both tests and surveys. It’s used in Lertap5 and reported towards the bottom of Stats1f reports. The Reliability section in Chapter 7 of the Lertap manual (pp. 11-13) is a recommended read.

For Test13, alpha came out to be 0.76; were we able to test the students again, perhaps with a parallel form, this is an estimate of how their test scores might correlate with those from the initial testing. For a short test, such as Test13, a reliability of 0.76 could be acceptable for some purposes.

[3] An example of a parallel forms reliability study is here.

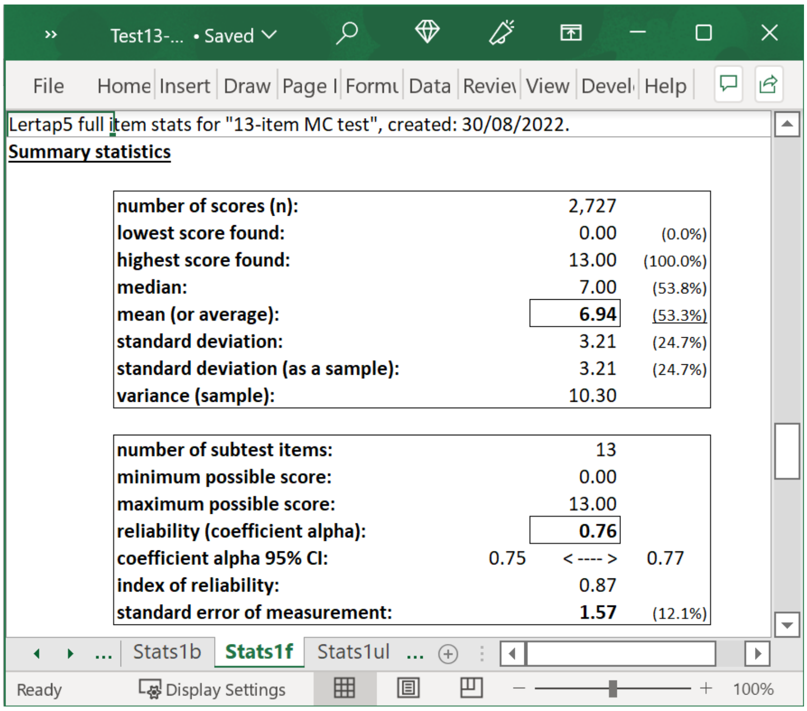
[4] The Iteman program from Assessment Systems is one. jMetrik is another (recommended).

But if test scores were to be used for grading purposes, we'd want to have an alpha value at or above 0.85, hopefully even getting close to 0.90.

Q4: Measurement error

The fact that we do not expect scores from parallel forms to be the same acknowledges that they will *vary*. We're not using exactly the same items. They're parallel in the sense that they mirror the original items, but item stems, keyed-correct responses, and distractors will all have changed.

Given a particular test, such as our Test13, in classical test theory a student's "true score" is said to be the average score as obtained over repeated administrations of the test, or over (in theory) an infinite number of parallel forms. A student's actual score on Test13 is their "observed score". The difference between the observed score and the true score is referred to as measurement error.



Lertap5 full item stats for "13-item MC test", created: 30/08/2022.

Summary statistics

number of scores (n):	2,727	
lowest score found:	0.00	(0.0%)
highest score found:	13.00	(100.0%)
median:	7.00	(53.8%)
mean (or average):	**6.94**	(53.3%)
standard deviation:	3.21	(24.7%)
standard deviation (as a sample):	3.21	(24.7%)
variance (sample):	10.30	

number of subtest items:		13	
minimum possible score:		0.00	
maximum possible score:		13.00	
reliability (coefficient alpha):		**0.76**	
coefficient alpha 95% CI:	0.75	< ---- >	0.77
index of reliability:		0.87	
standard error of measurement:		**1.57**	(12.1%)

Figure 11

The "Summary statistics" section of a Stats1f report is shown in Figure 11.

The standard error of measurement (SEM), seen at the bottom of the section, may be used to form a confidence interval around a student's Test13 observed score which we believe will include their true score at a given probability level.

In this case SEM was 1.57. For a student with an observed score of 8, adding and subtracting 1.57 gives an interval which in theory would have a 68% chance of including the student's true score. In this case the interval would be about 6.4 to 9.6. A 95% confidence interval would involve a range of just below 5.0 to just above 11.0. Page 12 in Chapter 7 of the Lertap manual has more details about the formation and use of such confidence intervals. The greater the reliability of a test, the smaller the standard error of measurement, and the shorter the 68% and 95% confidence intervals.

Q5: How might Test13's reliability be improved?

Have a look again at Appendix A and at Q2 in particular. Assume that the objective of Test13 is to assess competency in arithmetic skills.

Getting the right answer to Q2 requires students to know the number of days in the months of June and July. This is curious indeed. Why was this item in the test? We'll never know, but we might well be interested in knowing what the reliability of the test would be if Q2 were omitted.

You could try this for yourself: add an "exc" line to the CCs sheet in the Test13 workbook:

```
*col (c2-c14)
*sub res=(1,2,3,4), Name=(13-item MCtest), Title=(MathTest)
*key 31312 13223 325
*exc (c3)
```

The *exc line means "exclude the item found in column 3 of the Data worksheet".

That's where student responses to Q2 are found – See Figure 1 above.

Take Lertap5's Interpret and Elmillon options again.

The test reliability with Q2 omitted is 0.77, an increase over the 0.76 reliability figure for all 13 items. And now the standard error of measurement has gone down to 1.49 from the previous value of 1.57.

These may seem like modest gains, but in this business they're not insubstantial. We've removed a "bad item" from the test and have seen benefits.

Q6: How to identify problem items?

Q2 may be referred to as a "problem" item because its inclusion in the test has lowered test reliability (coefficient alpha).

Now, keep in mind that when we talk about test reliability and measurement error, we're dealing with overall test scores. It just might be that the creators and users of Test13 may not have been at all interested in test scores. Their interest may have been simply in item level results as seen, for example, in the first five columns of the Stats1b report, Figure 5 above, and in district results such as those seen in Figure 6.

At the item level there might not be "bad" and "good" items. Items which turned out to be easy, correctly answered by perhaps at least 75% of the students, may indicate satisfactory student mastery of whatever the item was meant to cover, something that, in turn, may reflect well on an instructor's teaching. Items which turned out to be difficult, say with less than half of the students getting them wrong, may point to a need for more teaching. Perhaps even Q2 was an item included in the district's math syllabus, as odd as that might seem to us.

But if we want reliable test scores with minimal measurement error, as mentioned above, we want coefficient alpha to be at least approaching 0.85. We might then use the scores for grading purposes. We might confidently use them to identify students needing help.

There are numerous spots in Lertap5's portfolio of reports that help identify problematic items. My favourite, my personal "go-to" report, is the plot included in Statsb worksheets, such as that seen in Figure 12.

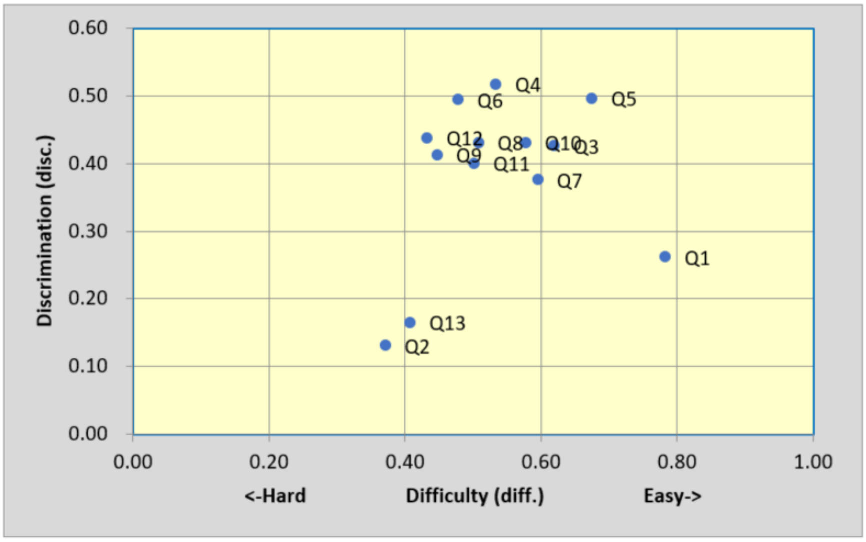


Figure 12

There are three "outliers" in Figure 12: Q1, Q2, and Q13. All three have item discrimination figures below 0.30, a cutoff point I tend to use when looking for problem items (other item analysts may lower this cutoff to 0.20).

It was the results seen in Figure 12 that led me to look at excluding Q2 from test scoring. Indeed, just looking at Q2 in Appendix A had already tended to tip me off. It seemed markedly different to the other items.

I didn't suspect problems with Q13 when looking at Appendix A, but, sure enough, excluding Q13 and Q2 boosted alpha to 0.78, and lowered the standard error of measurement to 1.40.

Anther Lertap5 worksheet useful for identifying poorly performing items is "IStats": weak items will be those with low squared multiple correlations, "SMCs" (see this topic). The "Closed-form lambda estimates" at the very bottom of IStats may also be useful at times (this topic).

Summary comments

This document has presented the terms, statistics, and common objectives found in classical item analysis. Two levels of analysis have been discussed: the item level and the test score level.

Item level statistics are fairly simple – our focus is on the number of students getting each item correct. We might not mind if nearly all students are able to pick out the keyed-correct answer to an item; such an outcome might indicate that our teaching has been adequate. At the same time, the items have distractors for a good reason – if the correct answer is very obvious there will be little challenge for students, and perhaps little useful feedback for instructors – looking at the performance of item distractors is consequently likely to be an important part of our review of the results. We want distractors to distract some of the students; that's why they're called distractors.

Getting an overall test score for each student takes us to another level. First of all, we seek some assurance that such a score makes sense – here our main statistic will be the test reliability index, coefficient alpha. Should alpha be low, below 0.70 for example, the implication is likely to be that the test score may be of limited use. If we were able to retest students, their scores would be expected to differ noticeably from those observed in the first testing, an indication of low reliability, a caution flag waving to say that we ought to view results with limited confidence.

Test reliability will be high when items show an ability to "discriminate". This will result when item distractors are performing well.

The quintile plots seen above in Figure 9 might be one of the best and easiest to use of Lertap5's "reports" when it comes to assessing distractor performance. The "popularity" of every distractor should decline as we move across each plot from left to right.

In Grp5, the least-capable students, few should be able to identify the item's correct answer, while, over on the right, in the strongest group, few should fail to identify the correct answer.

The quintile plot for item Q4, as seen in Figure 9, shows the highly desired pattern for a discriminating item. The plot for Q2, on the other hand, shows that some the best students, some of those in Grp2 and Grp1, are showing some favour for two of Q2's distractors. Only 60% of the strongest students (Grp1) correctly answered Q2, compared to near 100% on Q4.

These results suggest that Q2 needs attention and revision. It's not working as desired. Above I showed what might well happen when poorly performing items are excluded from the total test score; reliability improved after Q2 had been taken out.

As a final observation, my discussion here has not addressed cases where we might want a "mastery test", one used for pass/fail decisions, a test where, for example, a student must get 70% of the items correct in order to pass. Such tests are fairly common. See this topic for more information and also perhaps this technical journal paper.

Notes: This document has more about analysing Test13 results.

Importing data from csv[5] files and other sources is discussed here.

[5] csv files are text files with commas used to separate information fields

Appendix A

Test13 Items

1. Solve the following.

```
  7097
+ 1903
```

What is the answer?

(1) 89100
(2) 8990
(3) 9000
(4) 8000

2. Mike reached Sydney on 13th June in the morning and left on 4th August in the night. For how many days did Mike stay in Sydney?

(1) 53 days
(2) 52 days
(3) 51 days
(4) 50 days

3. What is the place value of 3 in 683941 ?

(1) 3
(2) 300
(3) 3000
(4) 30000

4. What is the place value of 8 in 548762 ?

(1) 8000
(2) 800
(3) 80000
(4) 8

5. Solve the following

$$\begin{array}{r} 7895 \\ \underline{-\ \ 5704} \end{array}$$

What is the answer?

(1) 1191
(2) 2191
(3) 2101
(4) 1101

6. 8763 – 6998 = ?

(1) 1765
(2) 2910
(3) 2875
(4) 15761

7. If the cost of 6 shirts is $ 480, then what will be the cost of 8 shirts?

(1) $ 288
(2) $ 384
(3) $ 640
(4) $ 360

8. Bus fare for four people is $ 100. What will be the fare for 10 people for the same journey?

(1) $ 25
(2) $ 250
(3) $ 400
(4) $ 1000

9. The price of one toy motor is $ 15.50. What will be price of 10 such motors?

(1) $ 1550
(2) $ 155
(3) $ 155.50
(4) $ 1550.50

10. Janice bought onions for $ 10, potatoes for $ 8, and tomatoes for $ 5. She gave $50 to the shopkeeper. How many dollars will she get back as change?

(1) $ 37
(2) $ 32
(3) $ 27
(4) $ 26

11. A box can contain 50 apples. How many boxes are needed to contain 2550 apples?

(1) 501
(2) 510
(3) 51
(4) 105

12. 20000 ÷ 200 = ?

(1) 1000
(2) 100
(3) 200
(4) 2000

13. 2016 ÷ 2 = ?

(1) 18
(2) 108
(3) 1003
(4) 1008