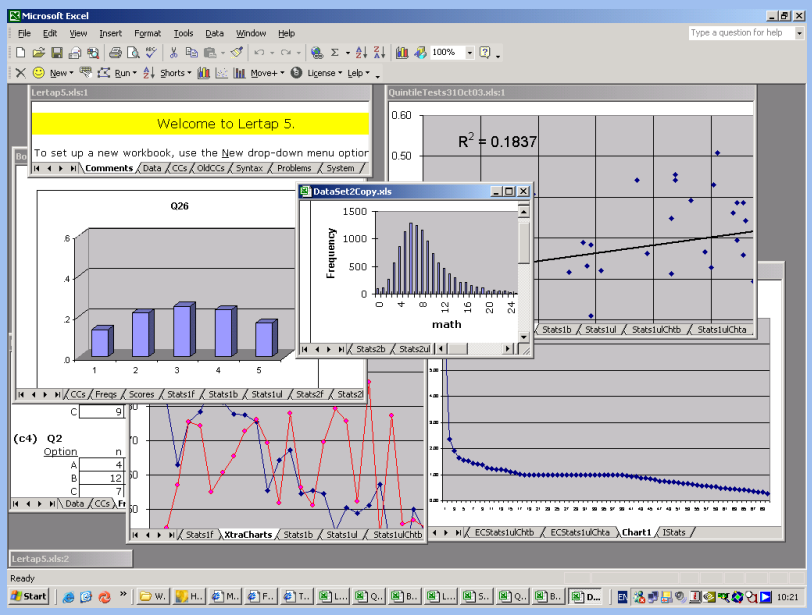


Tips for Lertap 5 with Excel

The "B Science Dataset"

Interactive PDF version

Made in Western Australia

Table of Contents

1 Introduction

If you have come here and yet have no idea of what "Lertap" is, see the link to introductory slides at point 2 below, and also have a look here, if you please.

I had several objectives in mind when I set about to create this document/website.

It is designed to be very practical, using real data and presenting actual issues which were faced by a very active testing site in 2012.

There are messages and suggestions. Foremost among them: examining data quality. It's easy to get Lertap to output lots of tables and graphs, but these will be of questionable validity if they're based on data which are not trustworthy. Excel and Lertap both have tools which let us look into the matter of data quality; these are demonstrated in the first topics. By and large these tools are easy to use.

Once the data have been poked and prodded, cleaned and caressed, it's on to looking at item and test quality before scores are used for reporting student results. In the dataset you're about to visit in this document, the responses students gave to two items signaled that they may have found the items to be ambiguous. There are steps that may be taken to counterbalance the ramifications of item ambiguity; you'll see them.

At times I make some effort in this document to cover matters which I suspect many Lertap users may be unaware of. From Excel, for example, I look thoroughly at the use of the data Filter, and make mention of pivot tables. From Lertap: the use of blank lines in CCs and Data worksheets; a tricky use of res= to count the number of unanswered questions; getting the histogrammer to work from a Breaks report; using packed quintile plots; interpreting item response flags; and a bit more.

The dataset itself is from a source which will remain anonymous. I should explain one term that may seem odd if you haven't seen it before: some countries refer to primary school grade levels as "Standards", and to high school grade levels as "Forms". The item responses contained in this dataset are from Standard 4 and Standard 5, which would generally correspond to what other countries might call Grade 4 and Grade 5.

The data were collected via the use of a scanner and imported into Excel.

Almost all of the screen shots in this document were taken while using Excel 2010; one or two are from Excel 2013. Lertap version 5.10.2 was used exclusively.

Links to other Lertap resources are listed below. Note that you can pick up some sample data from 4 -- the "M.Nursing" dataset, for example, exhibits problems similar to those found in this document's dataset. Lertap 5.10.2 is readily available at 5 and may be used for practice.

1 A PDF copy of this website's topics. A CHM copy (compiled help file for Windows). An iBook copy, ready for reading on an iPad or an iPhone. A link to the website itself.

2 A small set of PowerPoint slides with a quick introduction to Lertap. These are also available as a PDF file.

3 The main Lertap website. Has more examples and samples, with links to videos, the manual, and a variety of *riveting* technical papers (also known as "erudite epistles").

4 The online help system for Lertap. A primary source for finding out how to obtain Lertap, how to get it running, and understanding features added after the manual was printed.

5 Sample datasets for downloading. These also show off special features and showcase some of the most popular Lertap charts.

6 Larry's QUIA website, our developmental site. At times has special tidbits and morsels, especially for instructors and students.

7 The e-store for Lertap 5, the place which sells licenses for Lertap 5 users when they have more than 50 cases to process.

Please direct questions or comments to: lertap5@gmail.com

Last update: **27 May 2021**

1.1 Document sections

There are three main "themes" to this document, organized as "Parts".

Part 1 is given to data "snooping".

> The dataset has a few categorical variables, such as age and gender. I show how to use Excel and Lertap to check on data integrity.

Part 2 involves snooping of another sort.

> In this section I take a look at missing data, especially the many unanswered questions that are a "feature" of this particular dataset. There are so many unanswered questions that test reliability was affected, but in a way you might

not have anticipated. This section also looks at some grade-level and age-level differences in test scores, and in item responses too.

Part 3 gets into Lertap's real forte: item analysis.

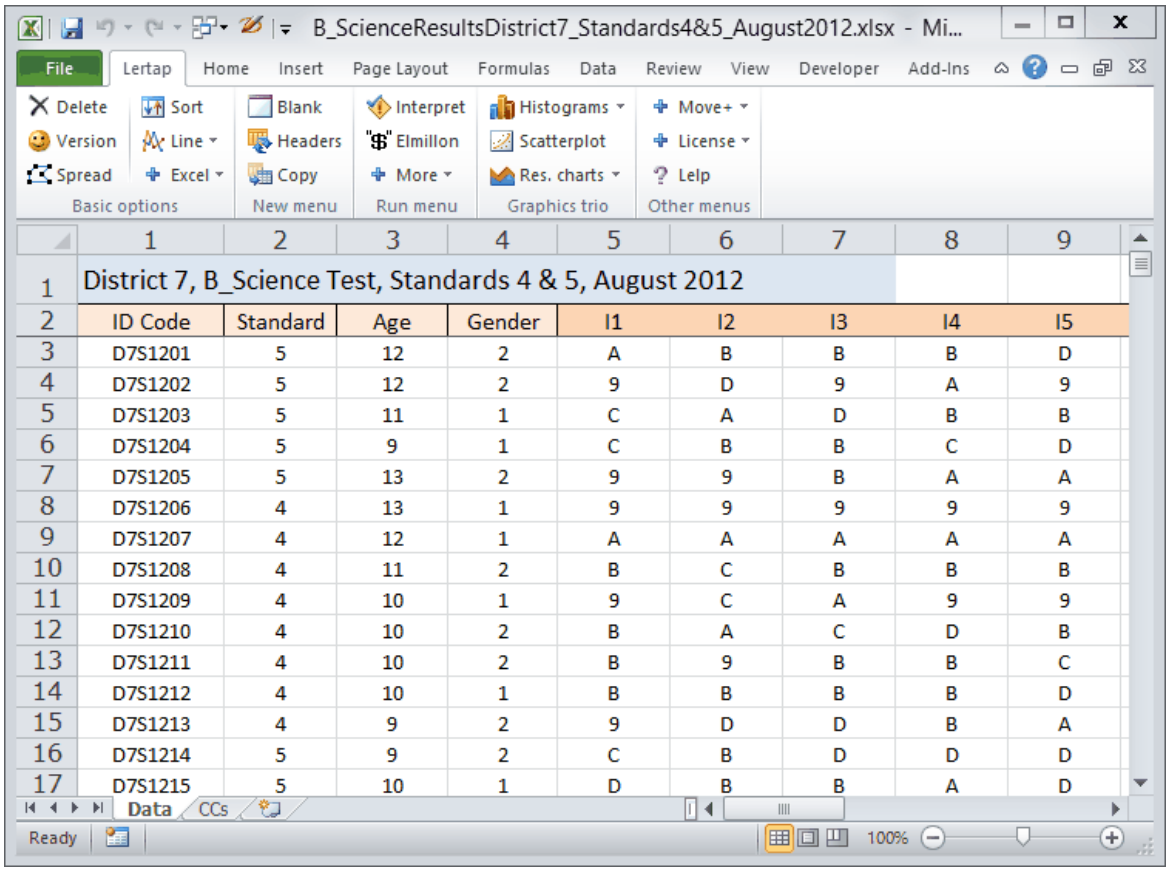
This part exemplifies how I myself go about item and test analysis. As is very common, the test I've looked at had some items which did not perform well, having some apparent ambiguity. In "double keying items", I show how to fine tune test scoring so that students are not disadvantaged by this oft-encountered problem.

I suspect that many readers will want to focus on the third part. But I suggest that the other parts are at least worthy of a good, relaxed browse. I say this as I've attempted to point out some Excel and Lertap tools which may well come in handy someday.

Get yourself up and running by first having a look at the Data and CCs worksheets, coming up next.

1.2 Data

The **Data** worksheet for this test is on display here in Excel 2010. The Excel ribbon is showing the Lertap tab with its five groups of icons, from "Basic options" to "Other menus". (More information on Lertap Data worksheets is here.)



	1	2	3	4	5	6	7	8	9
1	District 7, B_Science Test, Standards 4 & 5, August 2012								
2	ID Code	Standard	Age	Gender	I1	I2	I3	I4	I5
3	D7S1201	5	12	2	A	B	B	B	D
4	D7S1202	5	12	2	9	D	9	A	9
5	D7S1203	5	11	1	C	A	D	B	B
6	D7S1204	5	9	1	C	B	B	C	D
7	D7S1205	5	13	2	9	9	B	A	A
8	D7S1206	4	13	1	9	9	9	9	9
9	D7S1207	4	12	1	A	A	A	A	A
10	D7S1208	4	11	2	B	C	B	B	B
11	D7S1209	4	10	1	9	C	A	9	9
12	D7S1210	4	10	2	B	A	C	D	B
13	D7S1211	4	10	2	B	9	B	B	C
14	D7S1212	4	10	1	B	B	B	B	D
15	D7S1213	4	9	2	9	D	D	B	A
16	D7S1214	5	9	2	C	B	D	D	D
17	D7S1215	5	10	1	D	B	B	A	D

There were two nominal, or categorical, variables: Standard (column 2) and Gender (column 4). A student's age in years is found in column 3. Item responses begin in column 5 and extend through column 28.

Each item presented four options to the students. The response codes used for the options were {A,B,C,D}. If a student did not answer an item, or selected more than one option, a "response" code of 9 was used.

Above we can see that the first student selected option A for the first item (I1), B for the second (I2), third (I3), and fourth (I4) items, and, on I5, D.

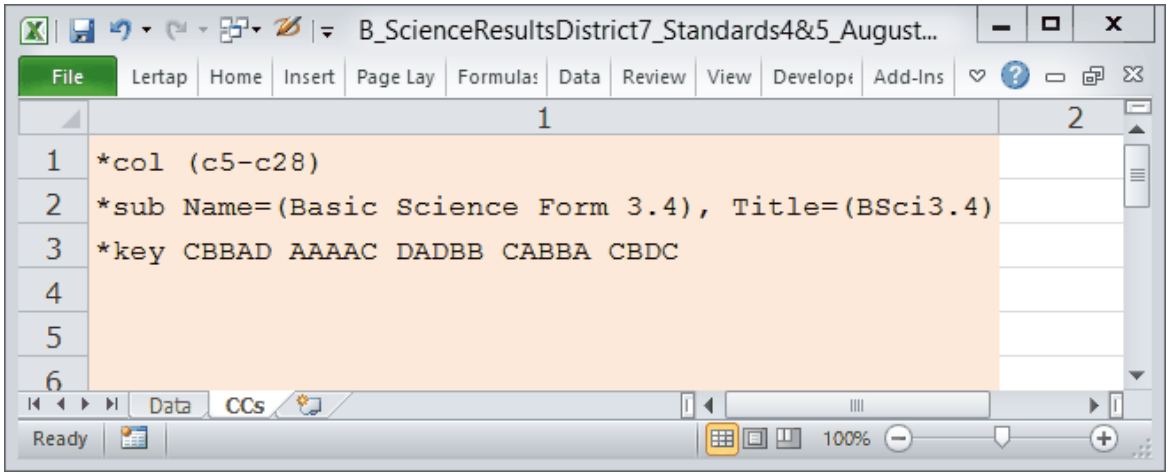
The second student did not answer I1, I3, and I9. The sixth student, D7S1206, did not answer any of the first five items.

There were 3,389 students when I started out. This dropped to under 3,000 after I had cleaned the data -- you'll see what I did a bit later.

Next up, the CCs worksheet.

1.3 CCs

Lertap uses a special worksheet to tell Excel how to process student responses. It's called "**CCs**".



The *col "card" (or line) indicates that item responses are located in columns 5 through 28 of the Data worksheet.

The *sub card provides Name and Title information which will be used when Lertap creates its various reports. Name and Title are optional fields on the *sub line; this being the case, this CCs worksheet could have had just two lines: *col and *key -- Lertap will use its innate intelligence to supply what it thinks might be a suitable Name and Title if you don't supply this information.

The *key card indicates the "keyed-correct answer" for each of the 24 items. The correct answer to the first item was C; it was B for the second and third items; A for the fourth item; D for the fifth; A for the sixth, and so on.

The spaces seen in this card are not important. A long string, *key CBBADAAACDADBBCABBACBDC, would have been quite okay. Here at Lertap Central we usually group item keys by fives as it makes finding the 14th item's key a lot easier (for example).

Is it possible for an item to have more than one keyed-correct answer? Yes, you bet. I demonstrate how in a later topic. If that's not enough, you can read much more about CCs worksheets and "control cards" by clicking here.

Jump to the next topic with a click here.

2 Part 1: snooping

With ready Data and CCs worksheets, we might ask Lertap for some results.

In this regard, I know that quite a number of people run Lertap in "production mode". When the production mode option has been turned on, getting Lertap to produce its copious number of "reports" is almost always just a one-click operation. (Read about production mode here.)

The advantage of using production mode? Time savings. A Data worksheet with results from thousands of students can take several minutes to fully process. (More if you're using Excel 2013; see our time trials.) Using production mode means you can let Lertap do all its things without having to manually direct it from one task to the next.

The disadvantage of using production mode? Your data might have errors; Lertap's results might be based on bad data. Before getting Lertap to make its many captivating tables and graphs, it's best to have a roll with the data first.

Back in the "good old days", whenever they were, researchers with fresh data in hand usually undertook a series of steps to the check on the integrity of their data before starting to get into data analysis. Errors in data processing are much more common than you might imagine.

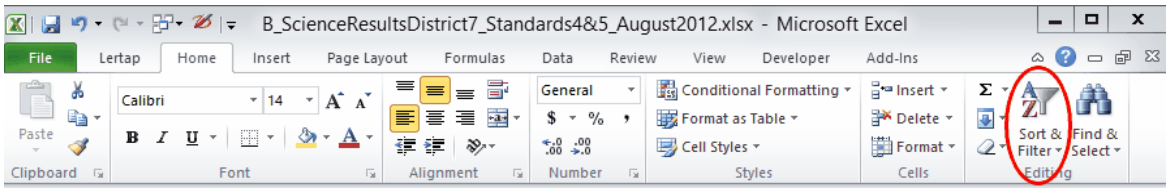
Lertap's main aide in this regard is its "Freqs" report. However, Excel is no slouch when it comes to data snooping. It can readily be used to check data quality, as I discuss in the next topic.

2.1 Snooping with Excel

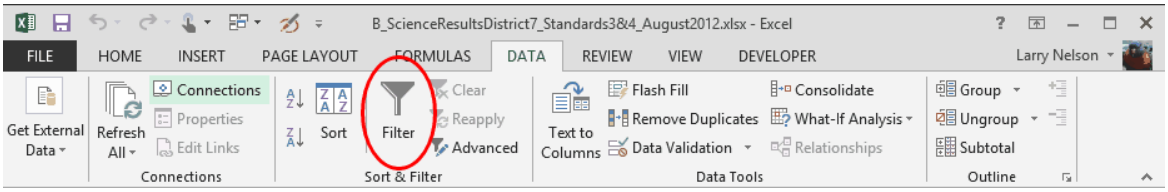
As mentioned, before I jump in and ask Lertap to spin out some results, I always always make an effort to see how the data look. I use readily-available tools to snoop the data, some in Excel, one in Lertap.

The snooping tools in Excel are excellent, excellent. The one I use all the time is the "**Filter**" option.

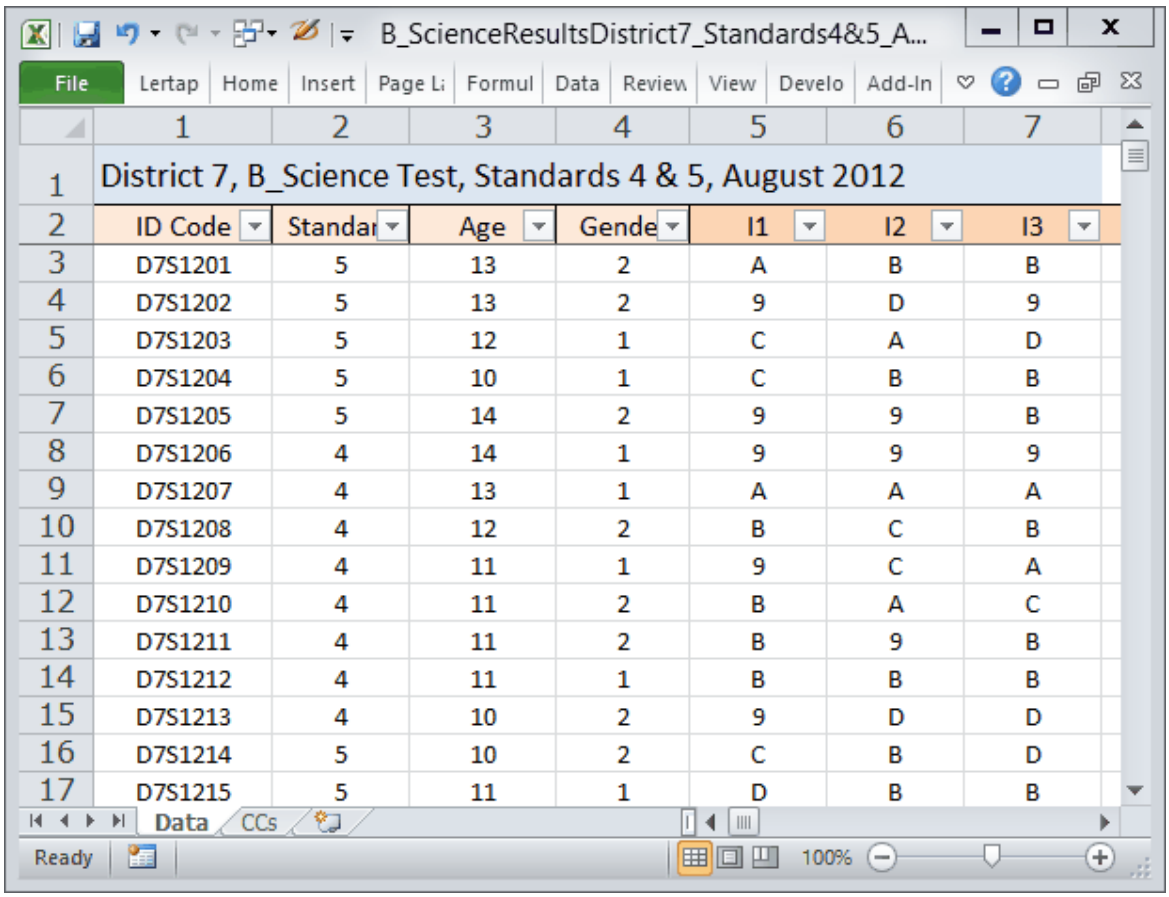
In **Excel 2010**, Filter is available on Excel's Home tab, in the Editing group, where it's referred to as "Sort & Filter":



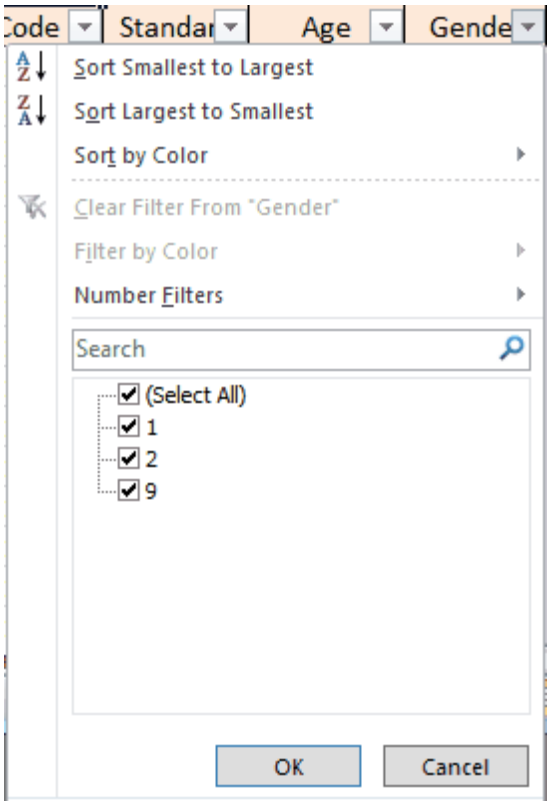
Things are different in **Excel 2013**, where Filter lives on the Data tab, in the Sort & Filter group:



This screen snapshot indicates how column headings change when the Filter option is selected in Excel 2010 -- little arrowheads appear at the top of each column:

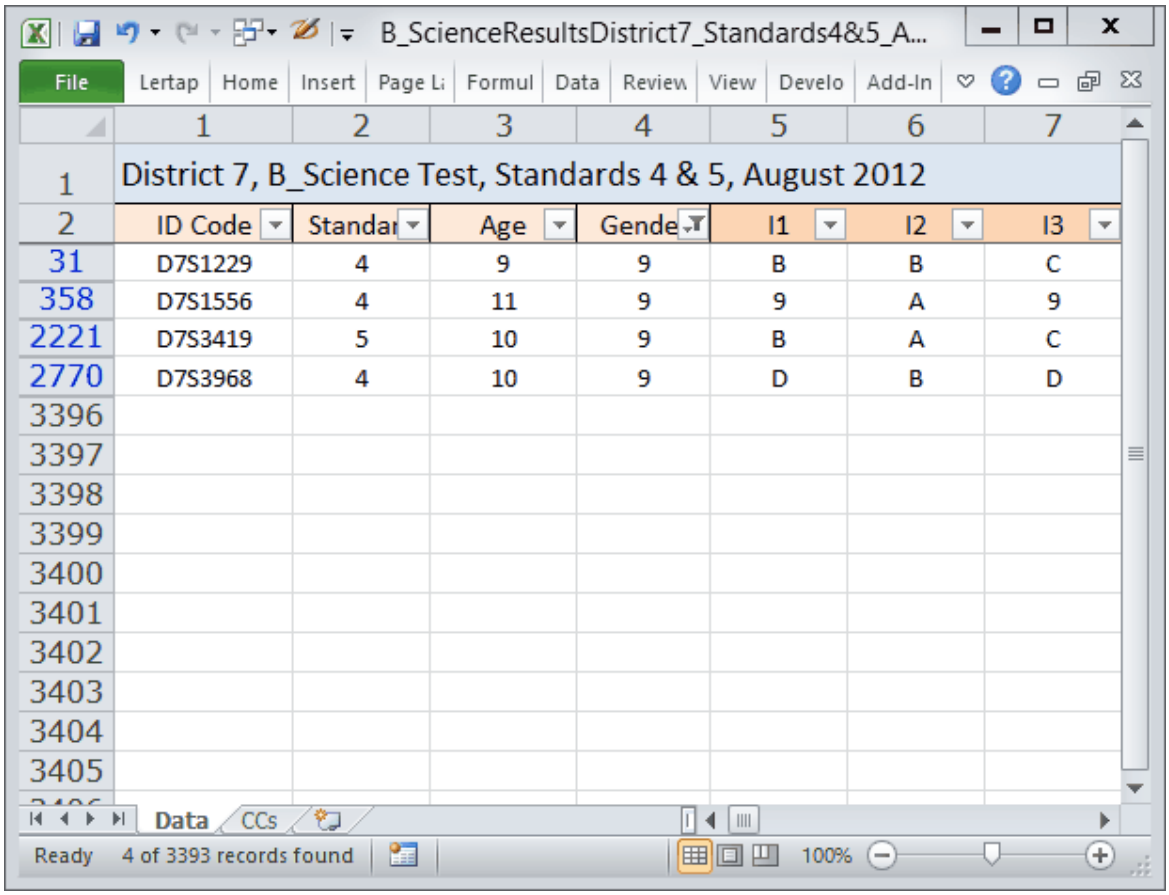


Let's say I'm interested in snooping the Gender column. When I click on the Gender column's Filter arrowhead, I get the following drop-down box of options:



Filter is telling me that the column has 1s and 2s and, what?, 9s? It's not supposed to have 9s, I should have only two codes for Gender, not three. Where are these 9s coming from? I need to snoop.

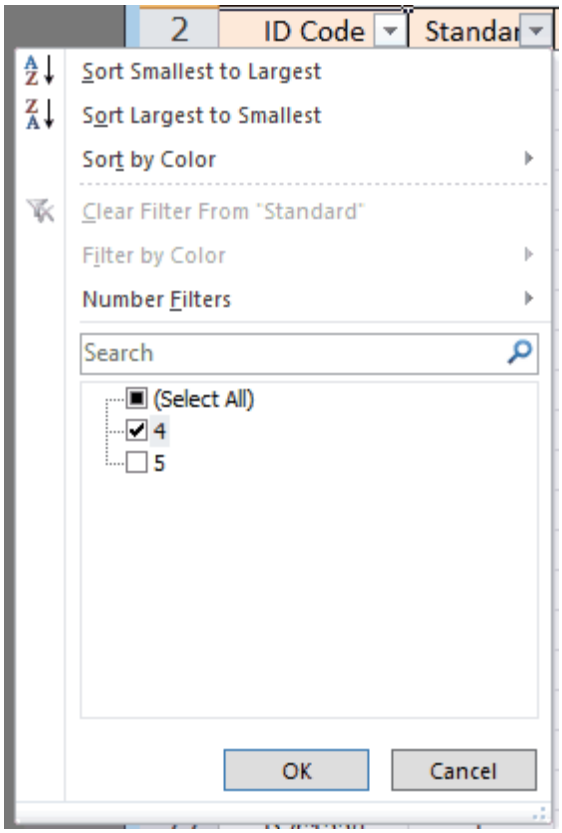
To find the records with a code of 9 in the Gender column, I click on the (Select All) option to clear all of the Search boxes. Then I click on the box for 9 to indicate that I want to see just those records with a Gender code of 9. This is what I see:



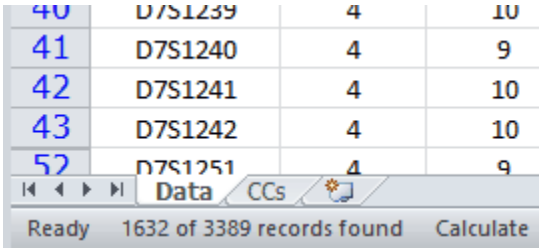
Filter tells me that there are four records with a Gender code of 9, and it indicates the corresponding row numbers in the Data worksheet: rows 31, 358, 2221, and 2770 in this case. If possible I'd get on the phone to the school district to see if they could supply correct gender information for the students with the ID Codes showing, after which I'd make changes in these four records.

Once I've finished snooping Gender, I turn off Filter simply by clicking it again (see the red circles way above).

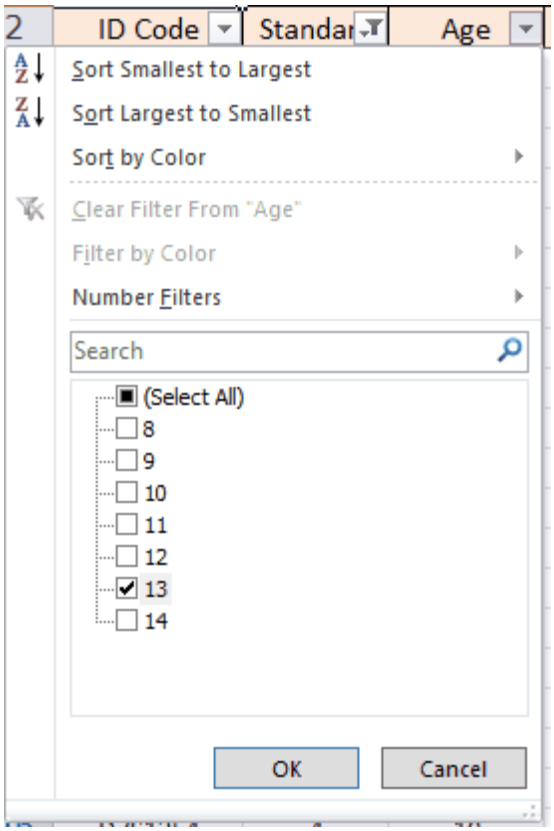
Note something else about Filter: it's possible to drill down. Let's suppose, for example, that I want to find how many Standard 4 students had an age of 13. I begin by clicking on Filter (red circles way above). The little arrowheads are then displayed at the top of the columns. I go to the column with Standard in it, and click on the arrowhead. I select Standard 4 as seen here:



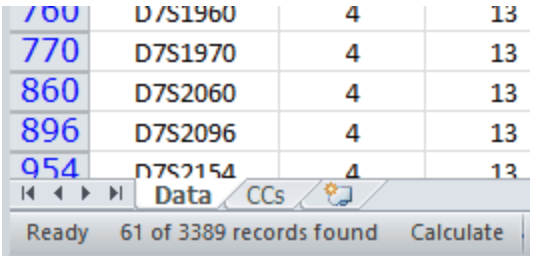
I click the OK button, and Excel filters the Data records, showing only those with Standard = 4. Down on the Status line, Excel tells me that it found 1,632 records having a value of 4 in the Standard column:



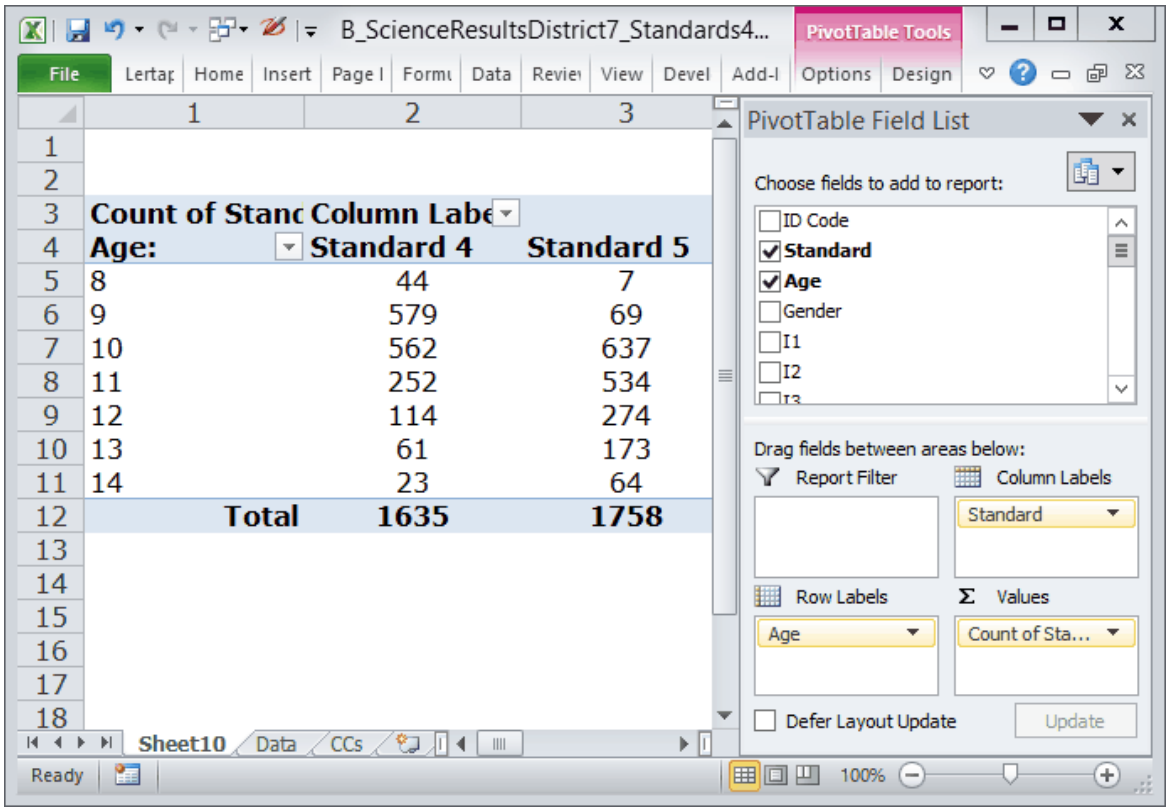
Next I click on the Age column's arrowhead, and select only the 13-year-olds:



I then look down at the Status line, and find that there are 61 students in Standard 4 with Age = 13:



Excel will make "pivot tables". With a little practice they can also be real handy. Here's one that took about 20 seconds to create in Excel 2010:



See what I've found out? Most of the 8-year-olds, 44 of them, are in Standard 4. But some of them (rather unexpectedly) are in Standard 5. And I note there are quite a number of older kids in Standard 4. How many are 12 years of age, or older, in Standard 4? 198 (sum 114, 61, and 23). More than 10% of the young students in Standard 4 were not so young. If I go into some of the Standard 4 classrooms, I'll find little kids, and some big ones, too.

Here I might comment that the data in this example are from a country in the "developing world". Not all students began their primary school education at the same age, and, in some cases, students were held back to repeat a school year. Knowing this, I still wonder how we might see seven 8-year-olds in Standard 5 -- perhaps children found to be gifted were allowed to skip grade levels (my wife went to primary school in Myanmar and says that this was called "double promotion"; I went to primary school in Wisconsin and there some of the "most capable" young students could skip a grade providing they had special short summer school sessions with bridging topics).

So, what do I make of this bit of snooping? Well, I've got four bad Gender codes, and also a real age spread in Standard 4 and Standard 5. The four Gender codes of 9 are obviously errors, but I don't have sufficient knowledge of the country to know if the age spreads in the two grade levels are unusual or not. I suspect they're not -- in

remote Australian schools, where indigenous children are involved, age spreads like these might be observed.

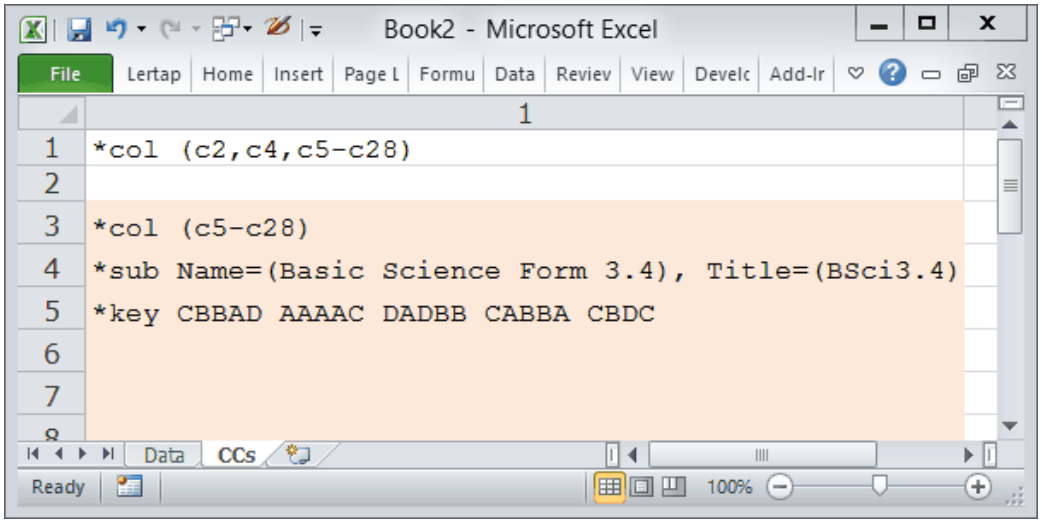
Erratum: astute readers may have noted a discrepancy. I actually made the pivot table before I did all the snooping using Filter. The pivot table indicates that there are 1,635 students in Standard 4. However, when I used Filter to find records with Standard = 4, it found 1,632. I had taken out the four records with Gender = 9 before I asked Filter to look at the Standard column. Three of these were from Standard 4.

Next topic.

2.1.1 Snooping with Freqs

I've already mentioned that Lertap's Freqs report is a useful tool for checking on data quality. In some cases it's not as "snooper" as Excel, but it can still be quite useful.

Let me show you. I modified the original CCs worksheet by inserting two lines above the original first row. Look:

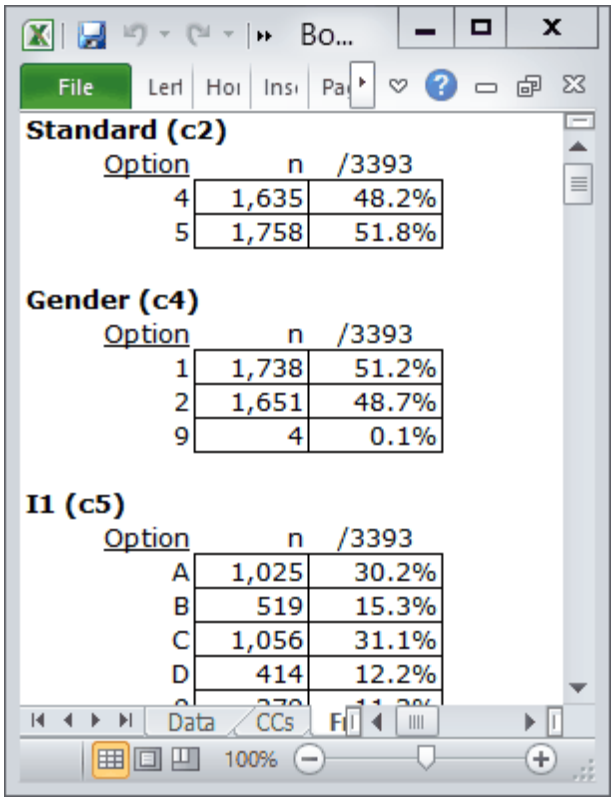


Now my *col line mentions two other columns. Column 2, or "c2", contains the code for Standard, while column 4, "c4", has data for Gender.

The second line is completely blank. Whenever Lertap encounters a row in a CCs or a Data worksheet whose first column is empty, or "blank", it'll think that it has reached the end of the world, and will then proceed to process results. This is what I do when all I want from Lertap is its Freqs report. I have the *col line, followed by an empty line, or row. After the empty row I can have other rows with stuff in them; Lertap will not read these rows, but I don't want it to. (This is handy with the Data worksheet too: when I want to see if my CCs lines are going to do all that I want, I'll

test them out with just the first 50 data records by inserting a blank row after the 52nd row in the Data sheet -- remember, the first two rows in the Data worksheet are reserved for other information; student responses begin in row 3.)

If I now click on Lertap's Interpret option, the resultant Freqs report will include information for columns 2 and 4. Behold:



Standard (c2)

Option	n	/3393
4	1,635	48.2%
5	1,758	51.8%

Gender (c4)

Option	n	/3393
1	1,738	51.2%
2	1,651	48.7%
9	4	0.1%

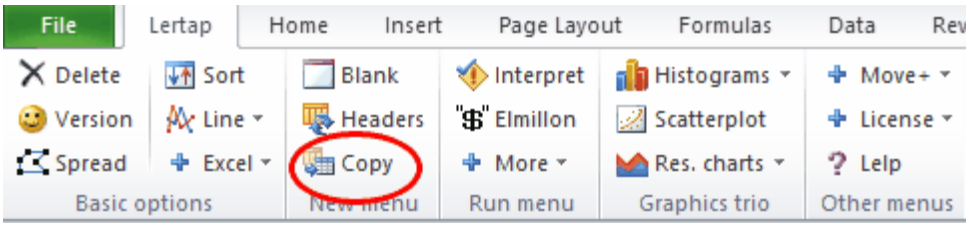
I1 (c5)

Option	n	/3393
A	1,025	30.2%
B	519	15.3%
C	1,056	31.1%
D	414	12.2%

Ah-ha! The values for Standard are as expected, just 4 and 5, but the Gender column has four 9s, which were not expected.

Okay then, thank you Freqs, you've pointed out that of the 3,393 data records, four of them have a missing Gender code. What next? Well, it depends -- if it's not too much trouble to get the missing data from the school district, I'd give it a try. Otherwise, I'd make a copy of the workbook, delete the four records from its Data worksheet, and carry on.

You do know that making a copy of a Lertap workbook is exceptionally easy, don't you? Just click on the Copy icon:



Why didn't I include the third column, "c3", in my special-purpose *col row discussed above? Because I know what will happen: Lertap will toot its horn and announce that its Interpret option does not like to process Data columns if they have more than one character. The third column, c3, has Age in years, and, unless all students are less than 10 years old, there will be entries in this column with two "characters", such as 10, 11, 12, and 13. The Interpret option will, cough, sputter, and say sorry, no can do column 3.

So, okay, I've shown that Freqs can be used in a manner analogous to Excel's Filter option when it comes to Data columns with single-character fields. This is handy, but note that Excel's Filter can actually show me which Data records have strange codes -- Freqs cannot.

When it comes to Data columns with more than one character in them, Freqs fails. Filter does not. But Lertap does have another tool which would, for example, let me look at the distribution of age by Standard -- it's the "Breakout score by groups" option. When combined with the Histograms option, and the "Box and whiskers" option, I can get very enlightening graphs. Later on I'll show you how to use these tools; for the moment let me just say that what I'd have to do is first copy the Age column from the Data worksheet to the Scores worksheet, after which it's pretty clear sailing. (There are options on the Move+ menu which let me move columns back and forth between Data and Scores.)

Next.

2.2 Adding a score

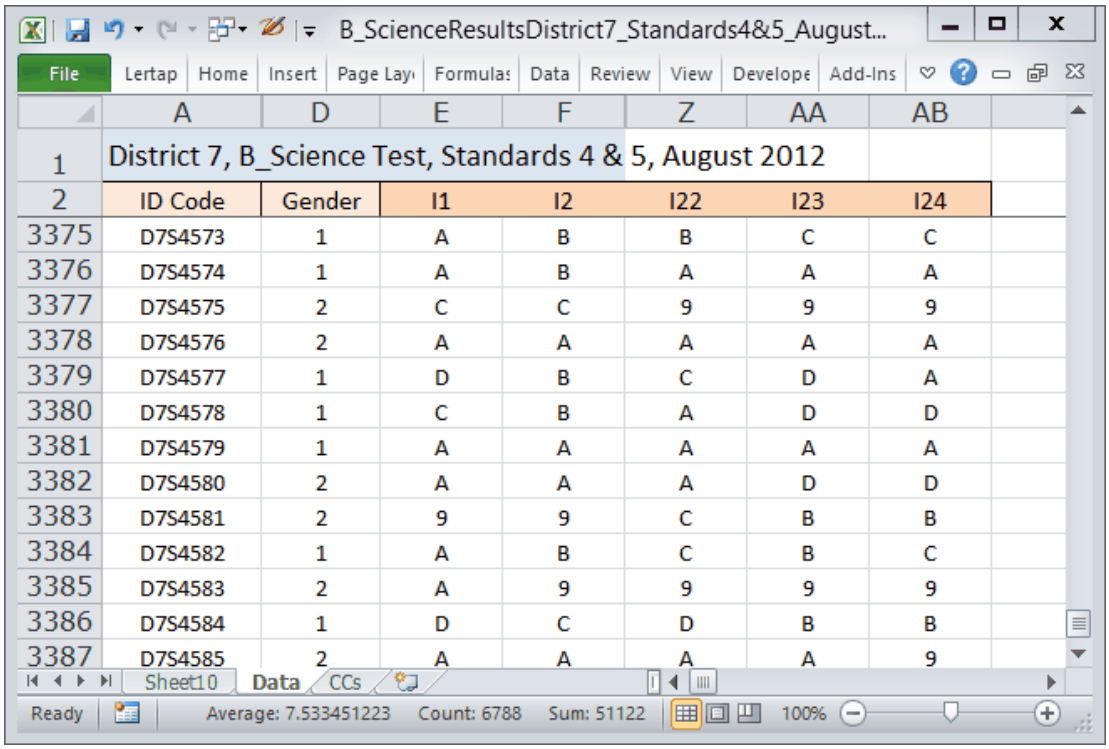
At this stage, I know things about this dataset that you don't. While you were watching the Australian Wallabies crush the British Lions by one point in a rugby match, I was scrolling through the Data worksheet. I sensed that it will be useful to add a field to the Data worksheet which counts the number of test items that a student did not answer. This means I want to sum the number of 9s found for each student in Data columns 5 through 28. To do this I can use a special Excel function, "COUNTIF", or I can use Lertap. In this topic I'll use COUNTIF.

The first student's results are found in row 3, column 5 of the Data worksheet. Now, I know that the COUNTIF function is one of many in Excel that likes to have letters in the column headers instead of numbers (*this is not strictly correct*, COUNTIF will also be happy using "rc" notation, an example is seen below). For example, the first

column will be "A", not "1". Lertap much prefers to have numbers as column headers and it'll almost always try to put them there. How to change column headers to letters instead so that I will be able to COUNTIF?

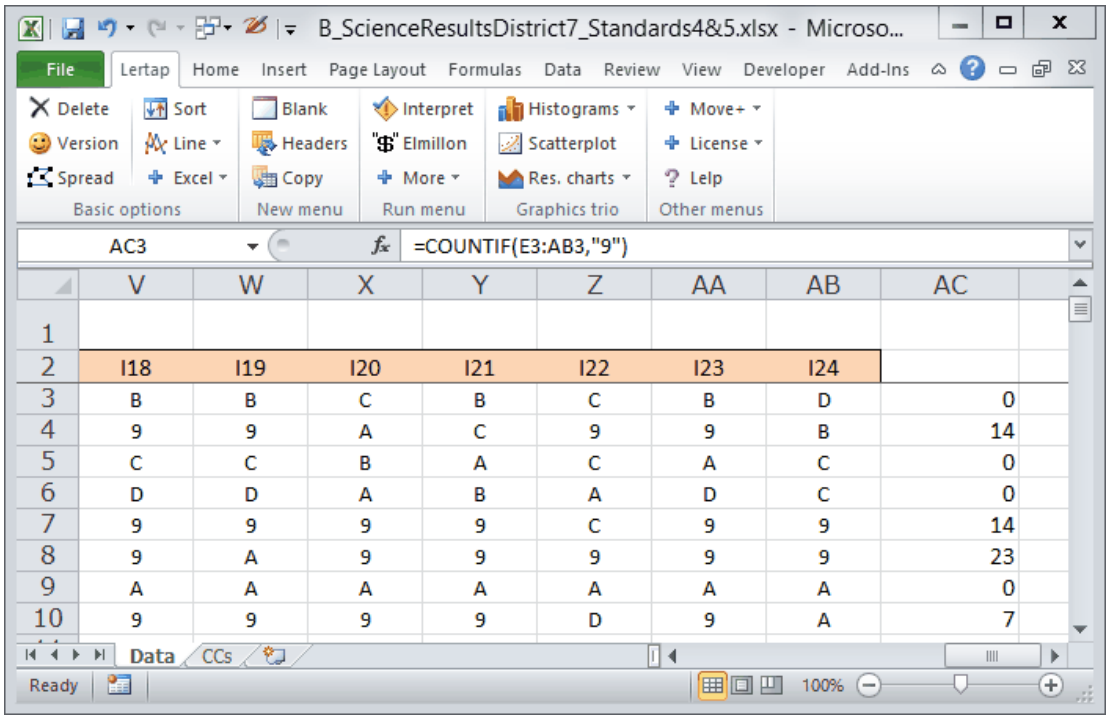
Most easy. I make use of a Lertap "Excel shortcut", the one called "Ref. style". Once I've done this, the column headers become letters, and I see that item responses, formerly found in columns 5 through 28, are now in columns E through AB.

I've taken another screen snapshot, hiding some of the columns so that you can see for yourself. The responses for the first item, I1, are in column E, while the responses for the last item, I24, are in column AB, see:



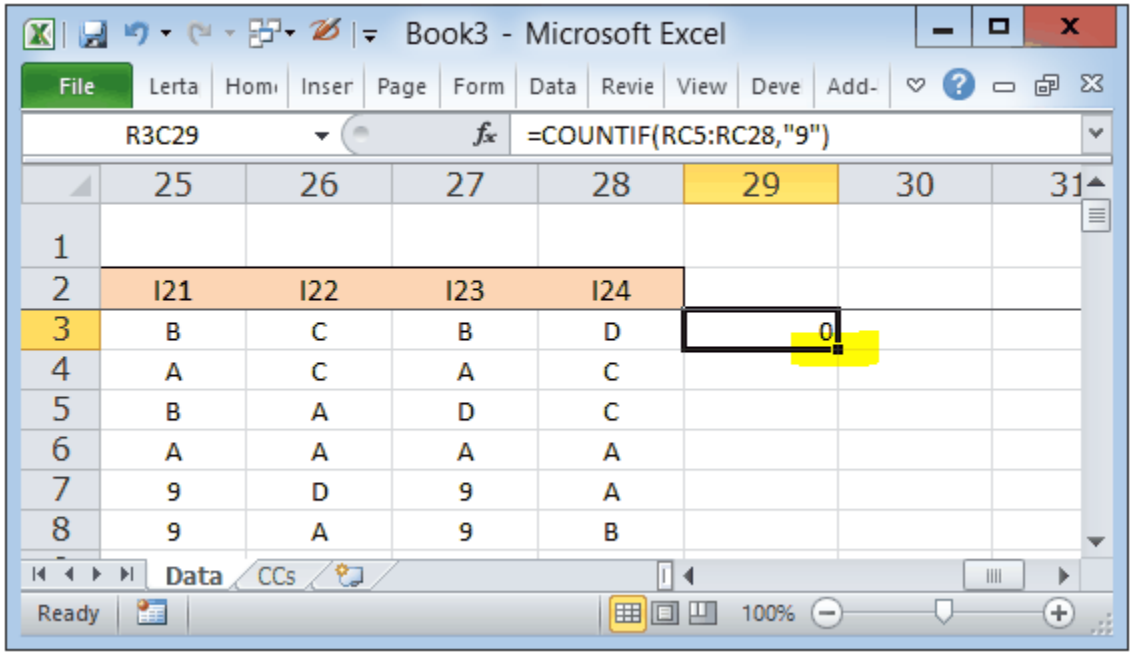
	A	D	E	F	Z	AA	AB
1	District 7, B_Science Test, Standards 4 & 5, August 2012						
2	ID Code	Gender	I1	I2	I22	I23	I24
3375	D7S4573	1	A	B	B	C	C
3376	D7S4574	1	A	B	A	A	A
3377	D7S4575	2	C	C	9	9	9
3378	D7S4576	2	A	A	A	A	A
3379	D7S4577	1	D	B	C	D	A
3380	D7S4578	1	C	B	A	D	D
3381	D7S4579	1	A	A	A	A	A
3382	D7S4580	2	A	A	A	D	D
3383	D7S4581	2	9	9	C	B	B
3384	D7S4582	1	A	B	C	B	C
3385	D7S4583	2	A	9	9	9	9
3386	D7S4584	1	D	C	D	B	B
3387	D7S4585	2	A	A	A	A	9

Once I have letters as the column headers, I am poised to use Excel's COUNTIF function, and a little beauty it is.



	V	W	X	Y	Z	AA	AB	AC
1								
2	I18	I19	I20	I21	I22	I23	I24	
3	B	B	C	B	C	B	D	0
4	9	9	A	C	9	9	B	14
5	C	C	B	A	C	A	C	0
6	D	D	A	B	A	D	C	0
7	9	9	9	9	C	9	9	14
8	9	A	9	9	9	9	9	23
9	A	A	A	A	A	A	A	0
10	9	9	9	9	D	9	A	7

In cell AC3, that is, in column AC, row 3, I have entered a formula which will count the number of 9s for the first student, as found in cells E3 through AB3:
=COUNTIF(E3:AB3,"9")



	25	26	27	28	29	30	31
1							
2	I21	I22	I23	I24			
3	B	C	B	D	0		
4	A	C	A	C			
5	B	A	D	C			
6	A	A	A	A			
7	9	D	9	A			
8	9	A	9	B			

A reader wrote in to say that he'd suggest using **"rc" notation** instead of having to switch column headers to letters. I show this in the snapshot above. ("rc" stands for "row-column"; if there's no number after the R, or after the C, that tells Excel to apply the formula to the whole row, or to the whole column.)

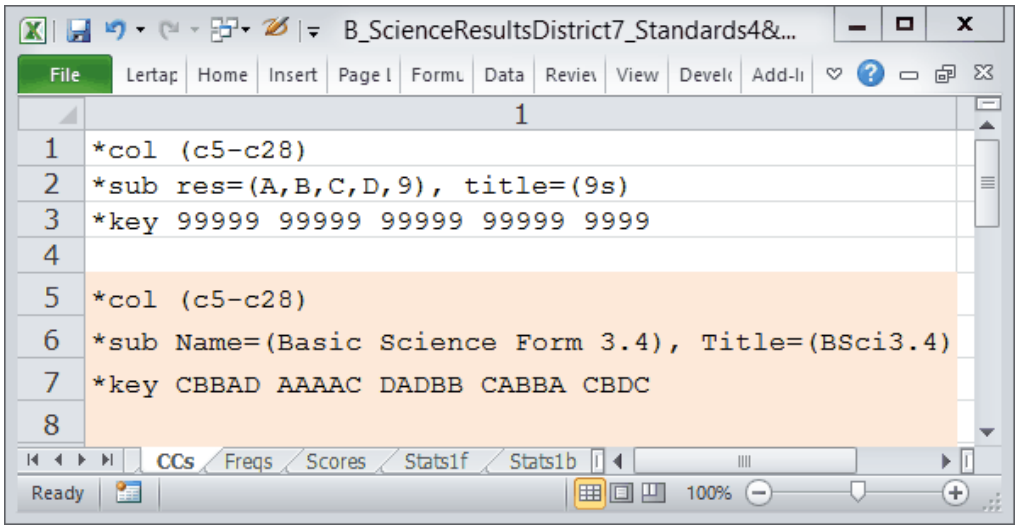
To have my formula copied to all following rows is easy: see the yellow highlight above? See the little black square that's in the yellow highlight? A double-click on that little black square will see that my formula is copied to all following rows.

Yes, I know, you're sitting there in seat 751U of your Airbus 380 flight to Waukesha, and wondering why oh why have I done this. You'll see two topics down. Let me first show you how you could get Lertap to count the number of 9s. I will do so in the next topic.

2.2.1 Lertap COUNTIF

My objective in this topic is to show you how Lertap can be coaxed into counting the number of 9s for each student.

Look at this CCs sheet (why is row 4 blank?):



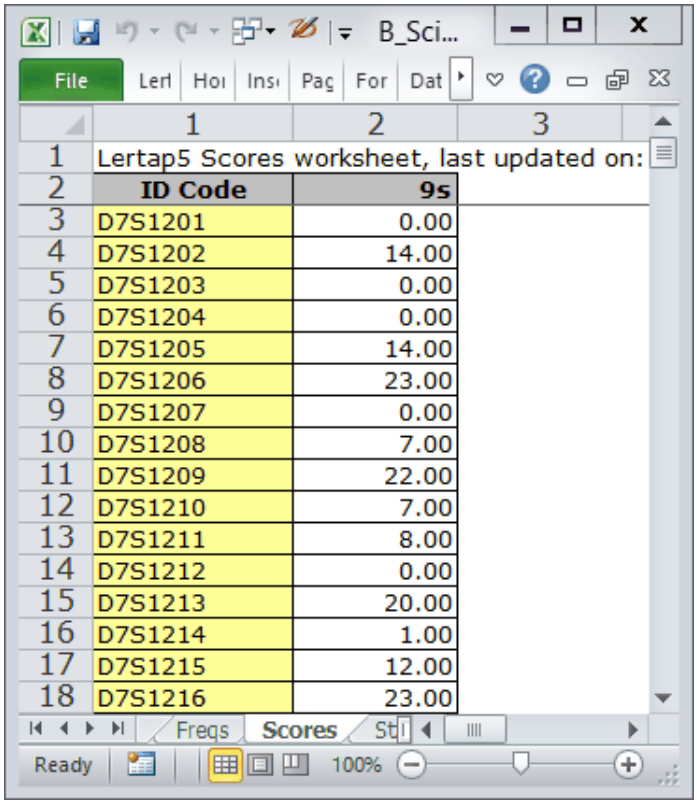
Ordinarily, Lertap thinks that cognitive test items will have four options, {A,B,C,D}. This is Lertap's "default" assumption for cognitive items.

Now, by using an "res=" declaration on the *sub "card", I am telling Lertap that my items have five possible valid responses, {A,B,C,D,9}.

The *key card says that 9 is the correct answer for each item. Each student will get one point each time she or he has a 9 as an item "response", effectively counting the

number of 9s. (Is it possible for an item response to get more than one point? Yes indeed, but of course, there's an example coming later. It's not relevant to this topic, but *mws and *wgs cards are the things to use when you want an item response to be worth something other than one point. You can even use *mws cards to award points for more than one item response; for example, you may decide that options B and D are each worthy of a point or, perhaps, half a point.)

I click on the Interpret option, and then on the Elmillon option. Look at my Scores worksheet:



Pretty neat, eh? The 9s column now has a count of the number of 9s for each student, the number of items the student did not answer.

As I glance down the 9s column in the screen snapshot above, I begin to feel a bit queasy. It looks like there may be quite a few students who did not answer a fair number of the questions. I've got 16 student results showing, and four of these students did not answer 20 or more items. That's 25% of this little sample. Is this characteristic of the greater picture? Of the three thousand plus students, did 25% of them answer just four questions, or less? Houston, we may have a problem should this be the case! (We'll find out the true state of affairs in a later topic.)

But wait. For what I have in mind, where is it best to have my column of 9s? I've just put it into the Scores worksheet. I'd rather have it in the Data worksheet, at least for the moment. So, I move it there using an option on the Move menu. Then I click on the Delete option which will eliminate all the worksheets added by using the Interpret and Elmillon options, leaving me with just Data and CCs worksheets, much as if I were just starting out. But now I have an extra column in the Data worksheet, the number of unanswered items for each student. This will be handy. Good on me.

Next topic.

3 Part 2: missingness

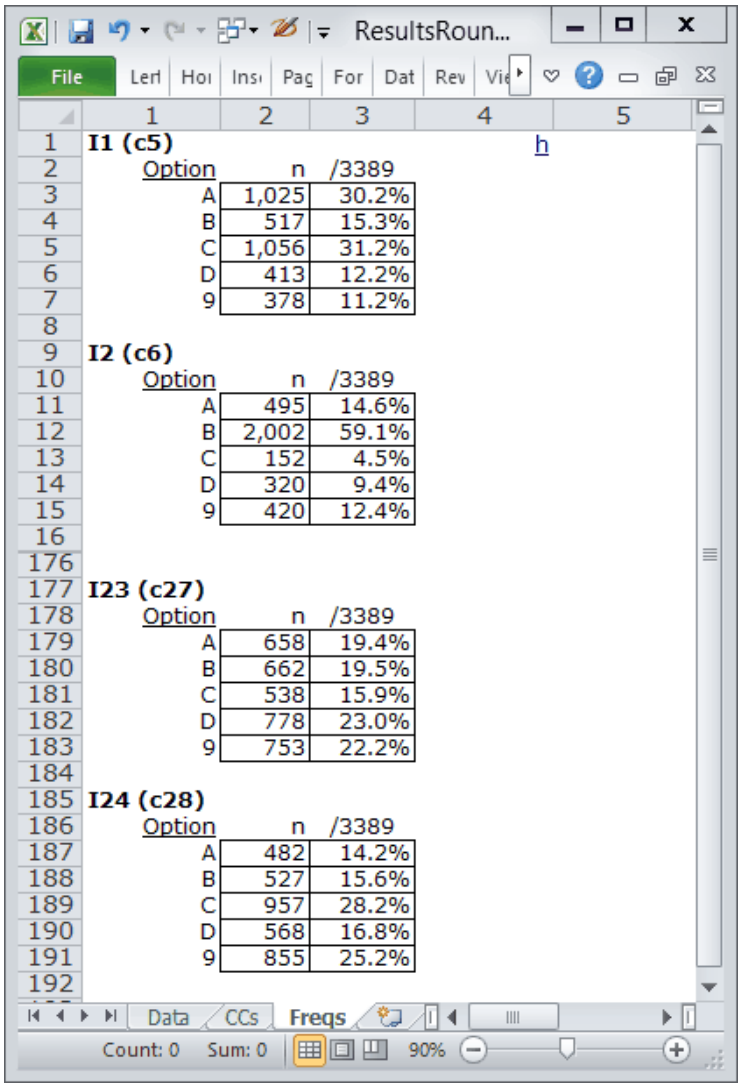
I have had an initial go with the Data worksheet, checking on data quality, snooping here, snooping there. I found four records with bad Gender codes, and I eliminated them from Data when you weren't paying attention. I had reason to suspect that a fair number of the children did not answer a fair number of them test items, so I used Excel's COUNTIF function to add an extra column to the Data worksheet which indicated the number of items not answered by a student. I also demonstrated how Lertap may be used to count the number of 9s.

Continue? Ready to roll? Okay.

Note that I am still not running in production mode. I much prefer to step through Lertap options as I need them, one by one, pausing after each step to ponder what Lertap and Excel reveal.

I have the Data and CCs worksheets as shown earlier (the new 9s column in the Data worksheet doesn't show in my screen snapshots, but it's there, in column 29, also known by Excel as column AC).

I click on Interpret and end up with Freqs:



ResultsRoun...

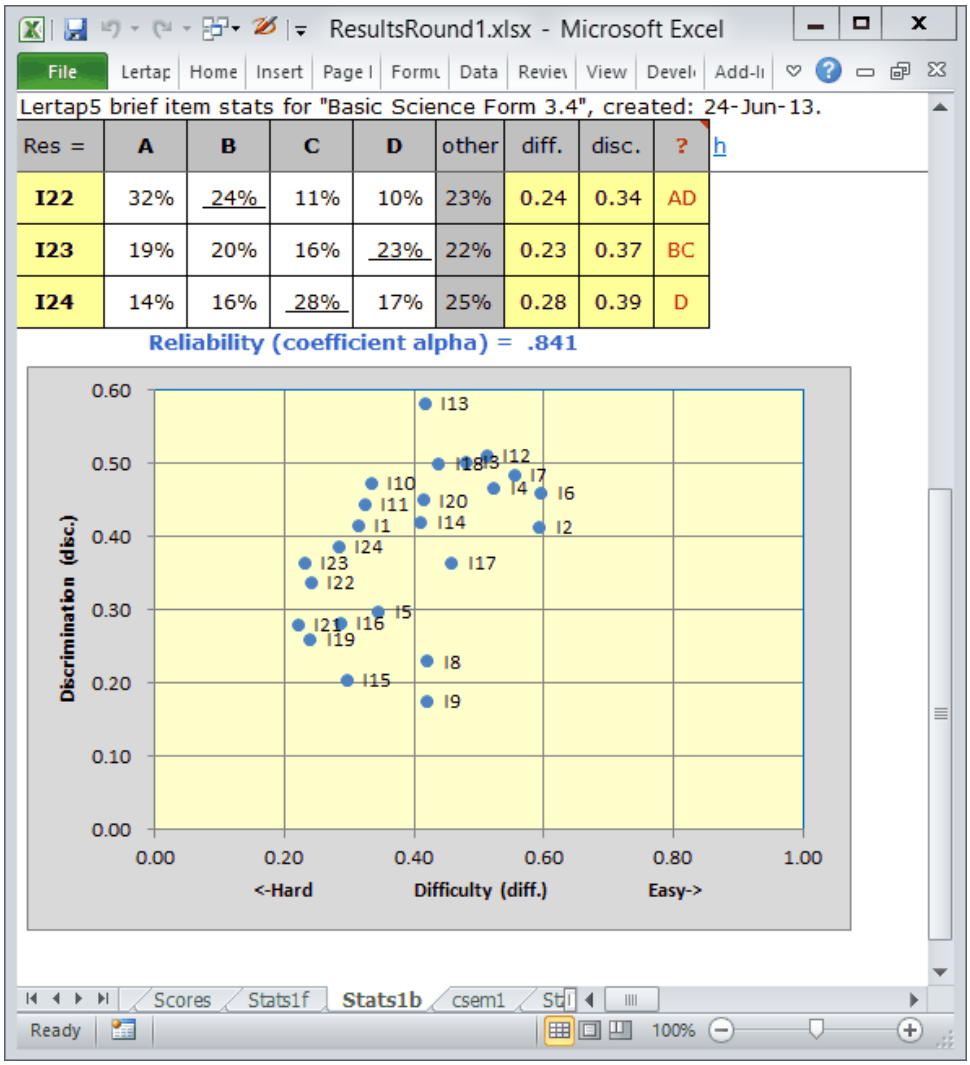
	1	2	3	4	5
1	**I1 (c5)**			h	
2	Option	n	/3389		
3	A	1,025	30.2%		
4	B	517	15.3%		
5	C	1,056	31.2%		
6	D	413	12.2%		
7	9	378	11.2%		
8					
9	**I2 (c6)**				
10	Option	n	/3389		
11	A	495	14.6%		
12	B	2,002	59.1%		
13	C	152	4.5%		
14	D	320	9.4%		
15	9	420	12.4%		
16					
176					
177	**I23 (c27)**				
178	Option	n	/3389		
179	A	658	19.4%		
180	B	662	19.5%		
181	C	538	15.9%		
182	D	778	23.0%		
183	9	753	22.2%		
184					
185	**I24 (c28)**				
186	Option	n	/3389		
187	A	482	14.2%		
188	B	527	15.6%		
189	C	957	28.2%		
190	D	568	16.8%		
191	9	855	25.2%		
192					

Data CCs Freqs

Count: 0 Sum: 0 90%

I have hidden many of the rows in Freqs so that you can see a couple of things. Note, first of all, if you would, that there were lots of 9s for the items. Remember what the 9s mean? Unanswered questions. On I1, 378 kids did not answer. By the time I get to the last two items, I23 and I24, more than 20% of the students did not answer. Hmmmm ....

I click on Elmillon and end up with Scores, Stats1f, Stats1b, csem1, and Stats1ul. Look at the little scatterplot at the bottom of Stats1b:



Res =	A	B	C	D	other	diff.	disc.	?
I22	32%	24%	11%	10%	23%	0.24	0.34	AD
I23	19%	20%	16%	23%	22%	0.23	0.37	BC
I24	14%	16%	28%	17%	25%	0.28	0.39	D

You must not look at the reliability figure of .841. Well, alright, look at it. For a test with a relatively small number of items (24), this woud be regarded as a reasonable result. I might ordinarily agree, but there may be more going on here -- I'm not going to be happy until I know more about all those unanswered questions.

You can see that the items tended to be on the difficult side: the scatterplot has many blips to the left of the **diff**. 0.40 line. These are the items where fewer than 40% of the children were able to pick out the right answer.

A problem is that **diff**., the item difficulty index, may have been lowered by the large number of students who did not answer the items. This could also impact on the reliability estimate: coefficient alpha is an internal consistency index; the more a student consistently gets items correct, or consistently gets them incorrect, the better

alpha might be. And, Lertap generally interprets unanswered questions as "incorrect". (But not always. See this discussion on did-not-see items and Lertap.)

The time has come to look into the extent of unansweredness in this dataset. Please topic forward.

3.1 9s R Us

Back a few topics I added a column to the Data worksheet which counted the number of 9s per student. This is the number of unanswered questions for each student. Now I'm going to have a much more detailed look at the 9s.

These are the questions I'd like to answer:

1. How many students have lots of 9s? As there are 24 items, at the start I'll say that "lots of 9s" will mean any student with at least twelve 9s, half the number of items.

2. Is there any apparent pattern to the 9s? If students find the item to be difficult, many times the 9s will come towards the end -- the students run out of time and are unable to take a stab at all questions.

3. Are there more 9s in the lower grade level, Standard 4?

4. How are the 9s affecting test reliability?

Get yourself a fresh cup of tea, and page forward.

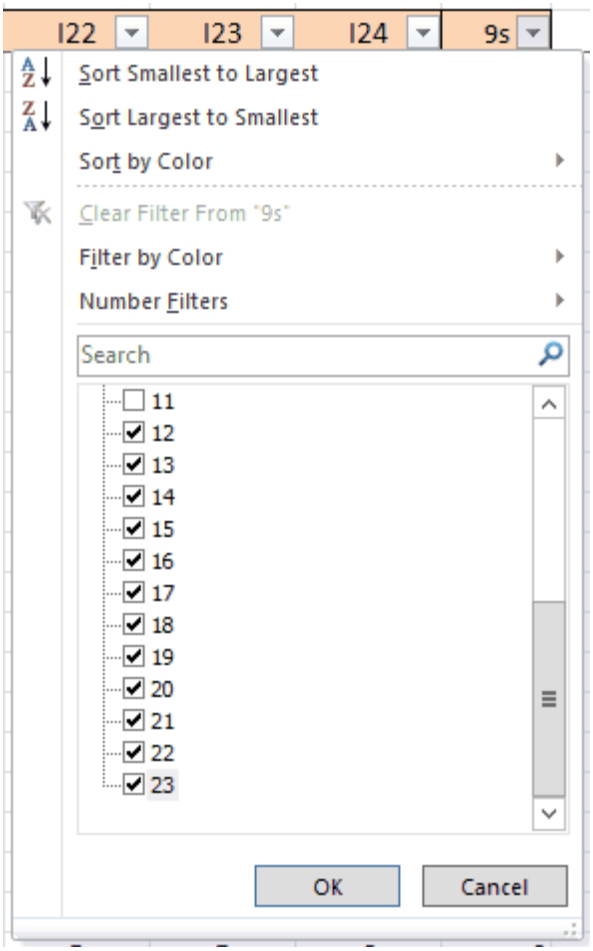
3.1.1 Lots of 9s?

This topic is devoted to the first two of my questions.

1. How many students have lots of 9s? As there are 24 items, I'll say that "lots of 9s" will mean any student with at least twelve 9s, half the number of items.

2. Is there any apparent pattern to the 9s? If students find the item to be difficult, many times the 9s will come towards the end -- the students run out of time and are unable to take a stab at all questions.

Remember that my 9s are found in column 29 of the Data worksheet. I'll use Excel's Filter tool on this column, as seen here:



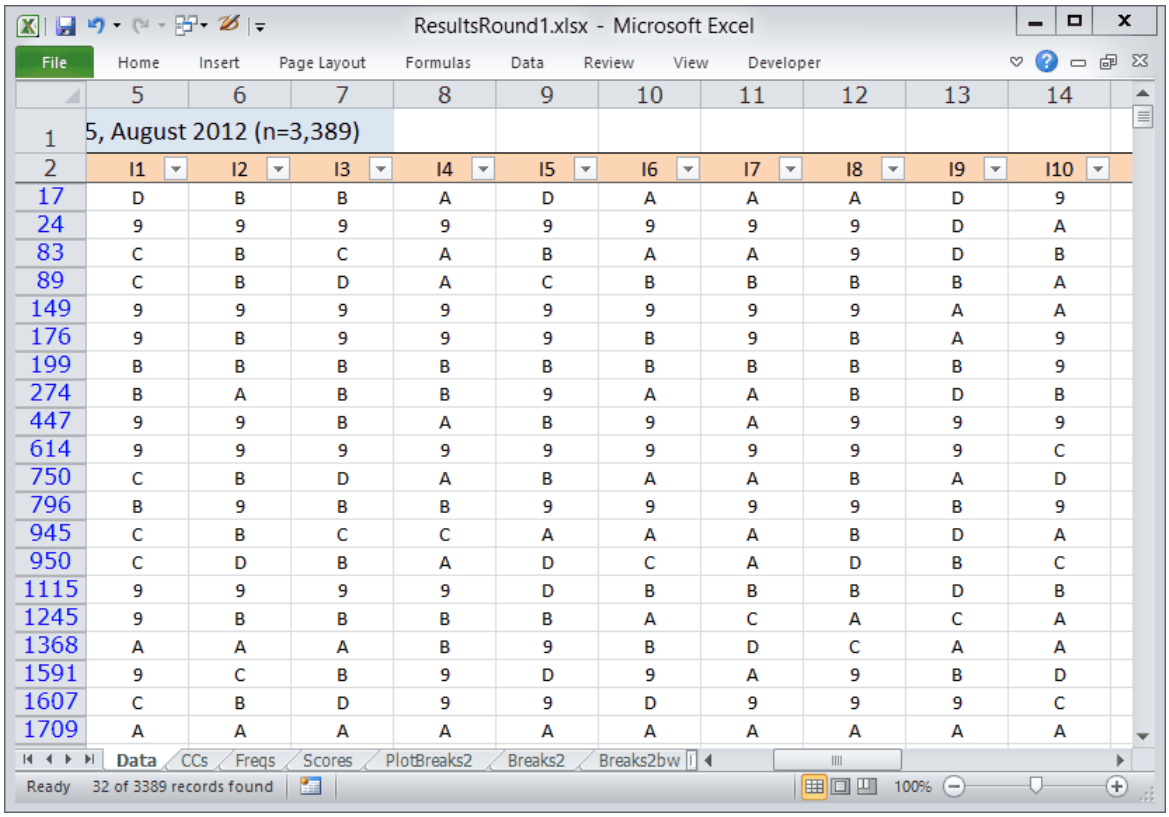
Excel reports that there are 485 records having from twelve to twenty-three 9s. This is about 14% of the total number of students. To me, this is indeed "lots of 9s".

What about a pattern? Is it the case that students start to answer the items, but then, for some reason, stop?

To answer this question, I'm going to use Filter to look at those students who have just twelve 9s.

Filter reports that there are 32 students who left twelve items unanswered.

Have a look at the first ten item responses for the first twenty of these students:



	5	6	7	8	9	10	11	12	13	14
1	5, August 2012 (n=3,389)									
2	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
17	D	B	B	A	D	A	A	A	D	9
24	9	9	9	9	9	9	9	9	D	A
83	C	B	C	A	B	A	A	9	D	B
89	C	B	D	A	C	B	B	B	B	A
149	9	9	9	9	9	9	9	9	A	A
176	9	B	9	9	9	B	9	B	A	9
199	B	B	B	B	B	B	B	B	B	9
274	B	A	B	B	9	A	A	B	D	B
447	9	9	B	A	B	9	A	9	9	9
614	9	9	9	9	9	9	9	9	9	C
750	C	B	D	A	B	A	A	B	A	D
796	B	9	B	B	9	9	9	9	B	9
945	C	B	C	C	A	A	A	B	D	A
950	C	D	B	A	D	C	A	D	B	C
1115	9	9	9	9	D	B	B	B	D	B
1245	9	B	B	B	B	A	C	A	C	A
1368	A	A	A	B	9	B	D	C	A	A
1591	9	C	B	9	D	9	A	9	B	D
1607	C	B	D	9	9	D	9	9	9	C
1709	A	A	A	A	A	A	A	A	A	A

Of the twenty students seen in this screen capture, eight have a 9 on the first item, I1 (column 5).

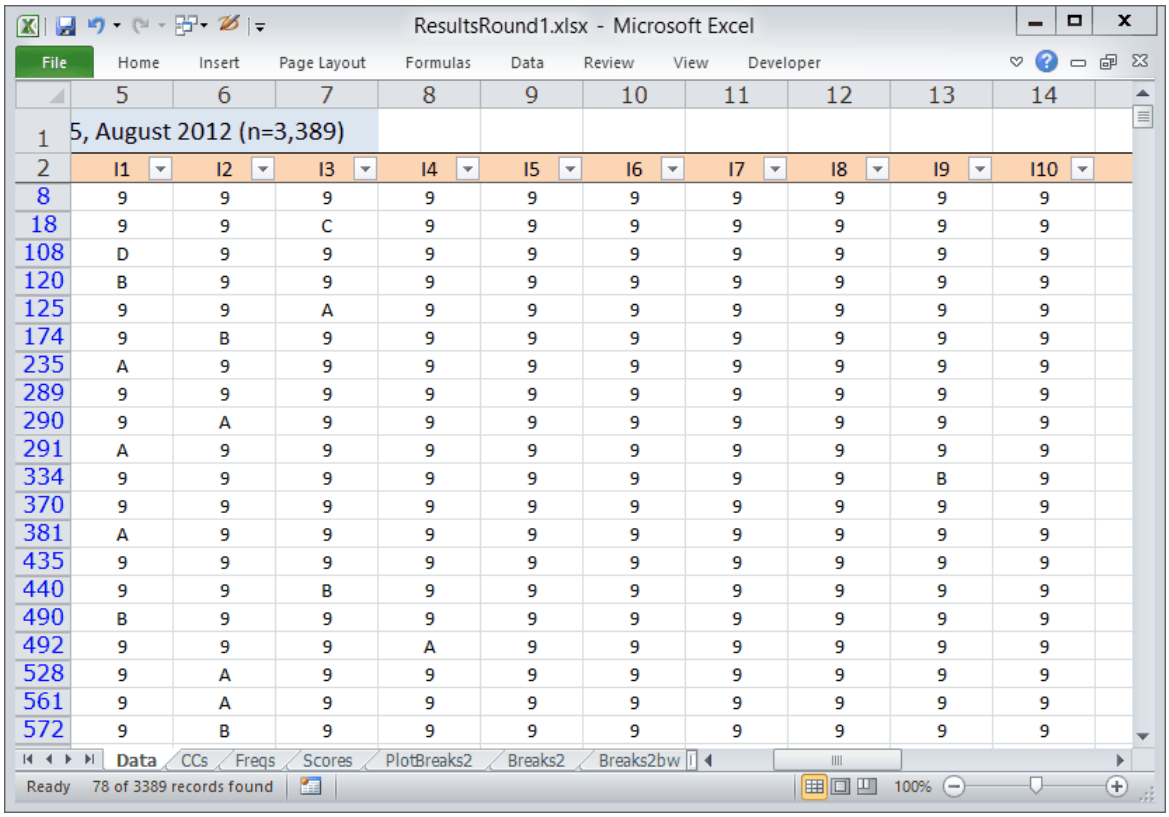
Look at the last record, row 1709. This student started out selecting option A on the first twelve items, after which he or she left the remaining twelve items unanswered. (No, that's right, you can't see this by looking at the screen snapshot above as I've only snapped the first ten columns. But trust me; I used to sell used cars.)

Look at the first record, row 17. This student appeared to make a good start at answering the items: the first nine questions were all answered. However, the student did not answer the next seven items (I10 through I16). Then, he or she did answer I17, I18, and I19, but left all remaining items unanswered.

A few students left initial items unanswered, but then, for some reason, began to answer the questions. See, for example, rows 24,149, 614, and 1115.

I do not see a pattern here. Things are helter-skelter.

I'll do one more thing: get Filter to show responses for all those students who did not answer twenty-three of the items. Here are the first twenty of seventy-eight records identified by Filter:



	5	6	7	8	9	10	11	12	13	14
1	5, August 2012 (n=3,389)									
2	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
8	9	9	9	9	9	9	9	9	9	9
18	9	9	C	9	9	9	9	9	9	9
108	D	9	9	9	9	9	9	9	9	9
120	B	9	9	9	9	9	9	9	9	9
125	9	9	A	9	9	9	9	9	9	9
174	9	B	9	9	9	9	9	9	9	9
235	A	9	9	9	9	9	9	9	9	9
289	9	9	9	9	9	9	9	9	9	9
290	9	A	9	9	9	9	9	9	9	9
291	A	9	9	9	9	9	9	9	9	9
334	9	9	9	9	9	9	9	9	B	9
370	9	9	9	9	9	9	9	9	9	9
381	A	9	9	9	9	9	9	9	9	9
435	9	9	9	9	9	9	9	9	9	9
440	9	9	B	9	9	9	9	9	9	9
490	B	9	9	9	9	9	9	9	9	9
492	9	9	9	A	9	9	9	9	9	9
528	9	A	9	9	9	9	9	9	9	9
561	9	A	9	9	9	9	9	9	9	9
572	9	B	9	9	9	9	9	9	9	9

What I would understand is students answering one of the first three items, and then giving up. Quite a few of these students did that (for example, rows 108, 120, 125,174, 235, 290, 291, 381, 440, 490, 528, 561, and 572). However, some students didn't provide their answer until well into the questions (8, 289, 334, 370, and 435).

Next, I'll drill down with these results. There were seventy-eight students who answered just one item (that is, had a 23 in column 29). I'll ask Filter to tell me how many of these answered the first item. Answer = 12. How many answered the second item? Answer = 32. How many answered the third item? Answer = 7. This is a bit better if I'm looking for a pattern: 51 of the 78 students who answered just a single item did so early in the test -- they gave it a bit of a go at the start, and then appeared to give up (which is what we might expect). Still, 27 of these students left the first items unanswered.

I'll come back to my questions:

1. How many students have lots of 9s? As there are 24 items, at the start I'll say that "lots of 9s" will mean any student with at least twelve 9s, half the number of items.

2. Is there any apparent pattern to the 9s? If students find the items to be difficult, many times the 9s will come towards the end -- the students run out of time and are unable to take a stab at all questions.

Yes, there were lots of 9s. About 14% of the students, 485 of them, did not answer at least half of the items.

There wasn't a clear pattern to me except, possibly, among those seventy-eight students who only answered one of the 24 items. The majority of these students, 51 of 78, gave their answer on one of the first three items. They seemed to give it a whirl at the beginning of the test, but quickly surrendered.

Click for next.

3.1.2 More 9s in Standard 4?

This is my third question:

3. Are there more 9s in the lower grade level, Standard 4?

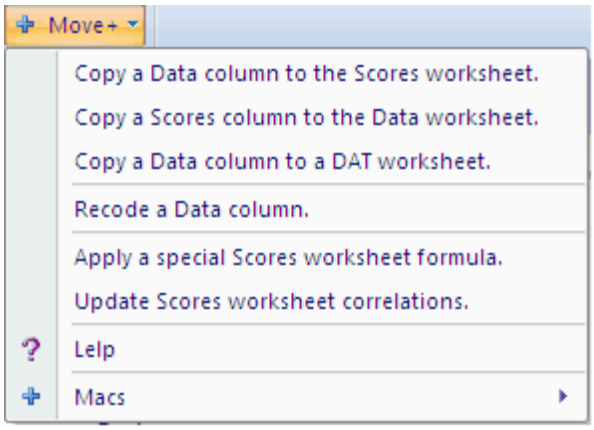
I'm going to use Lertap's "Breakout score by groups" option to answer this one.

The groups will be grade levels, Standard 4 and Standard 5, coded in column 2 of the Data worksheet.

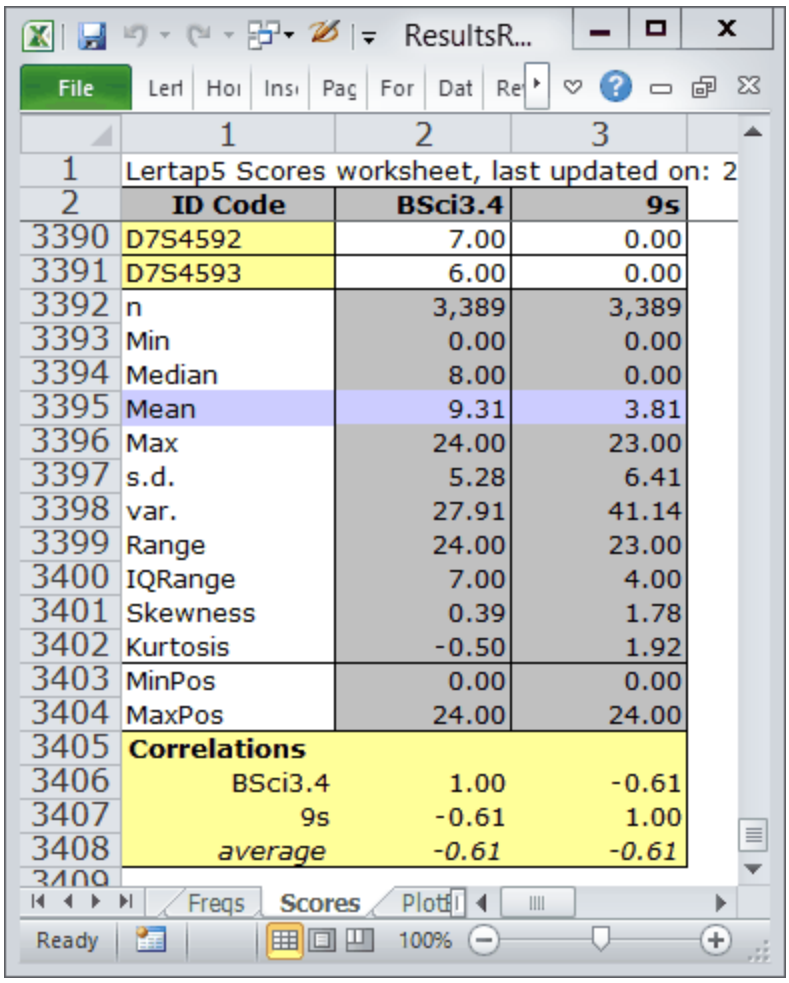
The "score" will be the number of unanswered questions. I've called this the "9s" field; it's found in column 29 of the Data worksheet.

The breakout option is happy to have the "groups" information in the Data worksheet, but it wants the "score" to come from the Scores worksheet. I need to copy Data column 29 to the Scores worksheet.

Is this hard to do? Is having a fresh, strong coffee first thing in the morning hard to do? Answer to both these Qs: no.



I take the first option from the Move menu, and before you can recite the names of my neighbor's children, dog, and egg-laying hen, my Scores worksheet looks like this:

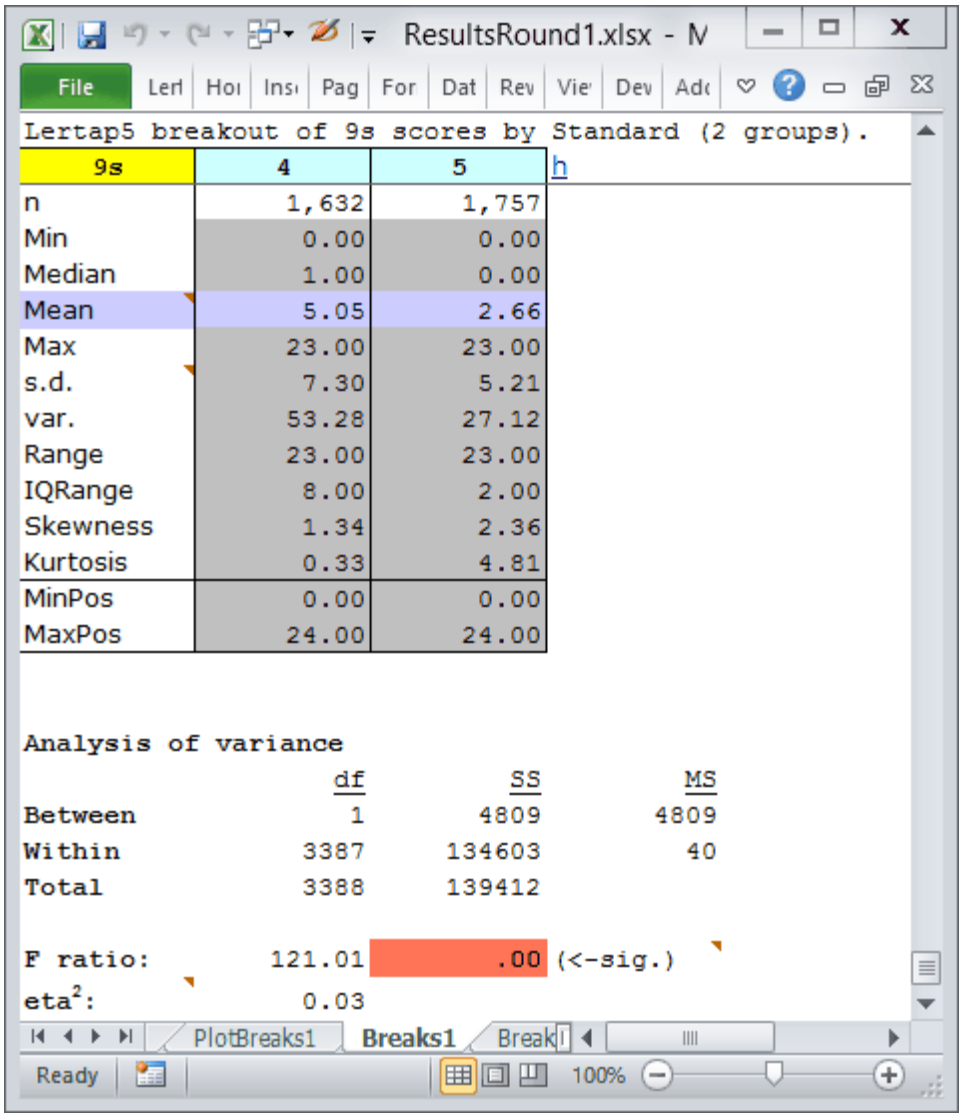


The first score, "**BSci3.4**", is test score; in this case, it is the number of items answered correctly. Since there are 24 items, the "**MaxPos**" (maximum possible) score is 24.00.

The second score, "**9s**", is the number of unanswered questions. The "MaxPos" is also 24 for this score, but I note that "**Max**", the maximum 9s score, was 23.00, meaning that all of these students answered at least one of the 24 items.

Pause for a moment to consider the correlation between the two scores. It's -0.61. This makes a lot of sense; students getting the higher test scores tended to have a lower 9s score, that is to say, unsurprisingly, students who failed to answer items tended to have a lower test score.

Now I am set to use that mighty "Breakout score by groups" option. It makes this little report for me:


ResultsRound1.xlsx - M

File | Lert | Hor | Ins | Pag | For | Dat | Rev | Vie | Dev | Ad

Lertap5 breakout of 9s scores by Standard (2 groups).

9s	4	5
n	1,632	1,757
Min	0.00	0.00
Median	1.00	0.00
Mean	5.05	2.66
Max	23.00	23.00
s.d.	7.30	5.21
var.	53.28	27.12
Range	23.00	23.00
IQRange	8.00	2.00
Skewness	1.34	2.36
Kurtosis	0.33	4.81
MinPos	0.00	0.00
MaxPos	24.00	24.00

Analysis of variance

	df	SS	MS
Between	1	4809	4809
Within	3387	134603	40
Total	3388	139412	

F ratio: 121.01 .00 (<-sig.)

eta^2: 0.03

PlotBreaks1 | Breaks1 | Break

Ready 100%

I see that the average number of unanswered questions in Standard 4 was 5.05, compared to 2.66 in Standard 5. Yes, it seems there were more 9s in the lower grade level.

But I'm not satisfied. I know that I can get Lertap to give me more information. Please, Lertap, could I have a histogram of the 9s score for each group?

With the Breaks1 report showing (as above), I click on "Histograms".

Here are the results for Standard 4 (I have not copied the whole histogram; some of the frequency bars extend much more to the right):

z	score	f	%	cf	c%	h
-0.69	0.00	707	43.3%	707	43.3%	||||||||||||||||||
-0.55	1.00	181	11.1%	888	54.4%	||||||||||||||||||
-0.42	2.00	123	7.5%	1,011	61.9%	||||||||||||||||||
-0.28	3.00	60	3.7%	1,071	65.6%	||||||||||||||
-0.14	4.00	44	2.7%	1,115	68.3%	|||||||||||
-0.01	5.00	33	2.0%	1,148	70.3%	||||||||
0.13	6.00	33	2.0%	1,181	72.4%	||||||||
0.27	7.00	31	1.9%	1,212	74.3%	||||||||
0.40	8.00	39	2.4%	1,251	76.7%	||||||||||
0.54	9.00	23	1.4%	1,274	78.1%	||||||
0.68	10.00	17	1.0%	1,291	79.1%	||||
0.82	11.00	19	1.2%	1,310	80.3%	|||||
0.95	12.00	14	0.9%	1,324	81.1%	||||
1.09	13.00	17	1.0%	1,341	82.2%	||||
1.23	14.00	17	1.0%	1,358	83.2%	||||
1.36	15.00	28	1.7%	1,386	84.9%	|||||||
1.50	16.00	36	2.2%	1,422	87.1%	|||||||||
1.64	17.00	14	0.9%	1,436	88.0%	||||
1.77	18.00	23	1.4%	1,459	89.4%	||||||
1.91	19.00	26	1.6%	1,485	91.0%	|||||||
2.05	20.00	31	1.9%	1,516	92.9%	||||||||
2.19	21.00	25	1.5%	1,541	94.4%	||||||
2.32	22.00	33	2.0%	1,574	96.4%	||||||||
2.46	23.00	58	3.6%	1,632	100.0%	||||||||||||||

How many Standard 4 students left 12 or more items unanswered? I drag my mouse down the "f" column, from a score of 12 to a score of 23. When I do this, Excel adds "Sum: 322" to its Status bar, so Answer = 322. This is equal to 19.7% of the Standard 4 students.

Here are the results for Standard 5:

z	score	f	%	cf	c%	h
-0.51	0.00	1,018	57.9%	1,018	57.9%	||||||||||||||||||
-0.32	1.00	202	11.5%	1,220	69.4%	||||||||||||||||||
-0.13	2.00	121	6.9%	1,341	76.3%	||||||||||||||||||
0.06	3.00	58	3.3%	1,399	79.6%	||||||||||
0.26	4.00	43	2.4%	1,442	82.1%	||||||||
0.45	5.00	30	1.7%	1,472	83.8%	||||||
0.64	6.00	36	2.0%	1,508	85.8%	||||||
0.83	7.00	19	1.1%	1,527	86.9%	||||
1.03	8.00	26	1.5%	1,553	88.4%	|||||
1.22	9.00	23	1.3%	1,576	89.7%	|||||
1.41	10.00	12	0.7%	1,588	90.4%	||
1.60	11.00	6	0.3%	1,594	90.7%		
1.79	12.00	18	1.0%	1,612	91.7%	||||
1.99	13.00	15	0.9%	1,627	92.6%	|||
2.18	14.00	16	0.9%	1,643	93.5%	|||
2.37	15.00	15	0.9%	1,658	94.4%	|||
2.56	16.00	18	1.0%	1,676	95.4%	||||
2.75	17.00	11	0.6%	1,687	96.0%	||
2.95	18.00	9	0.5%	1,696	96.5%	||
3.14	19.00	14	0.8%	1,710	97.3%	|||
3.33	20.00	10	0.6%	1,720	97.9%	||
3.52	21.00	10	0.6%	1,730	98.5%	||
3.71	22.00	7	0.4%	1,737	98.9%		
3.91	23.00	20	1.1%	1,757	100.0%	||||

How many Standard 5 students left twelve or more items unanswered? I drag my mouse down the "f" column, the score frequency column, from a score of 12 to a score of 23, and, looking at Excel's Status bar for "Sum:", I find that Answer = 181. This is equal to 10.3% of the Standard 5 students.

3. Are there more 9s in the lower grade level, Standard 4?

You bet! Compare the two percentages: approximately 20% of the Standard 4s left at least twelve questions unanswered, compared to about 10% of the Standard 5s.

Reliability is next.

3.1.3 Reliability vs. 9s

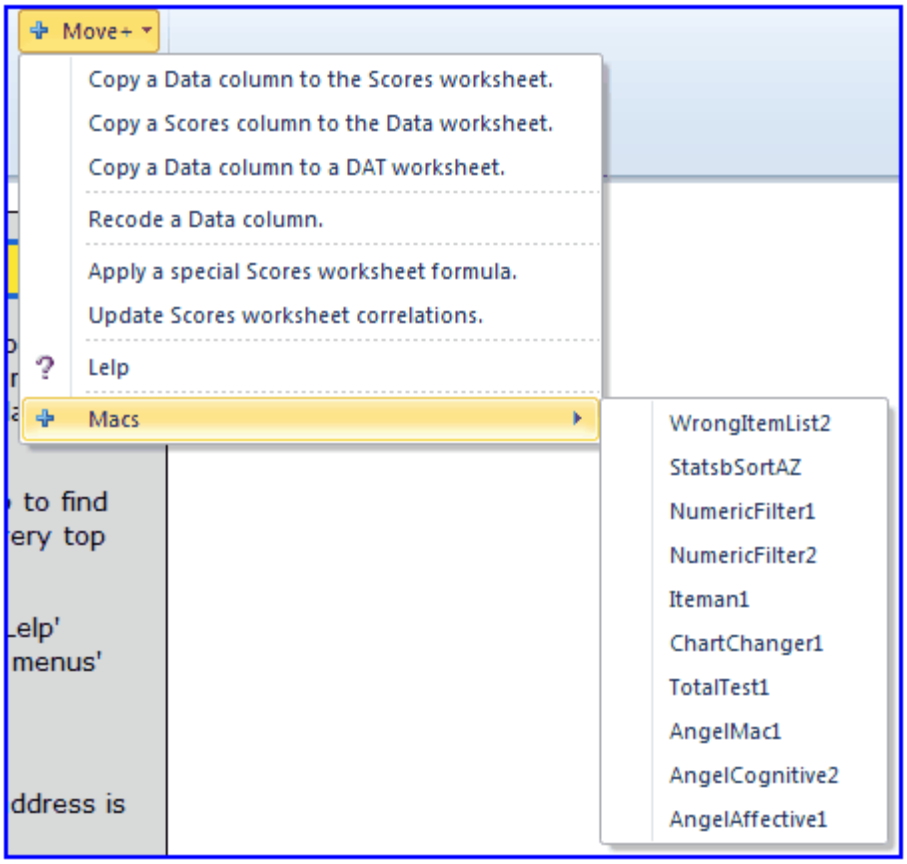
My fourth question:

4. How are the 9s affecting test reliability?

Back a ways, I found a reliability (coefficient alpha) of .841 for this administration of the 24-item test.

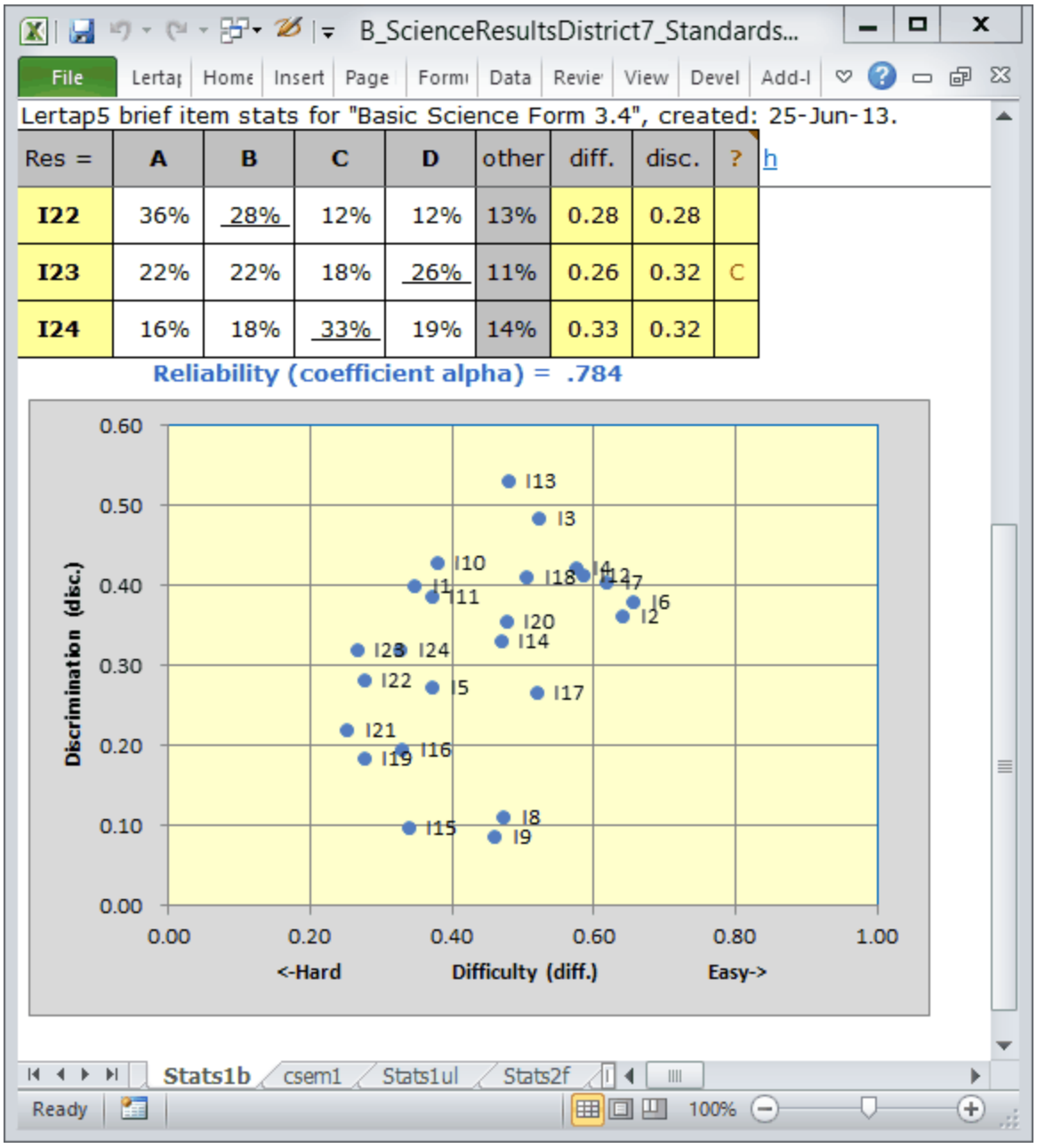
Know what I'm going to do now? Knock out all those students with a 9s score of 12 or more from the dataset. I want to have a Data worksheet wherein all of the students answered at least half of the items.

How will I do it? I'll use the "**Numeric Filter 1**" macro, available from the Macs menu:



After this macro has run, I'm left with a Data worksheet with 2,904 students.

Next, I take Lertap's Interpret and Elmillon options, as per usual. I look at the bottom of the Stats1b report, get out my screen snapshotter, take a picture, and present results here:



Reliability, as indexed by coefficient alpha, has gone down. It was .841 with all of the students; now, after taking out the 485 students who left twelve or more items unanswered, it's .784. This is quite a drop.

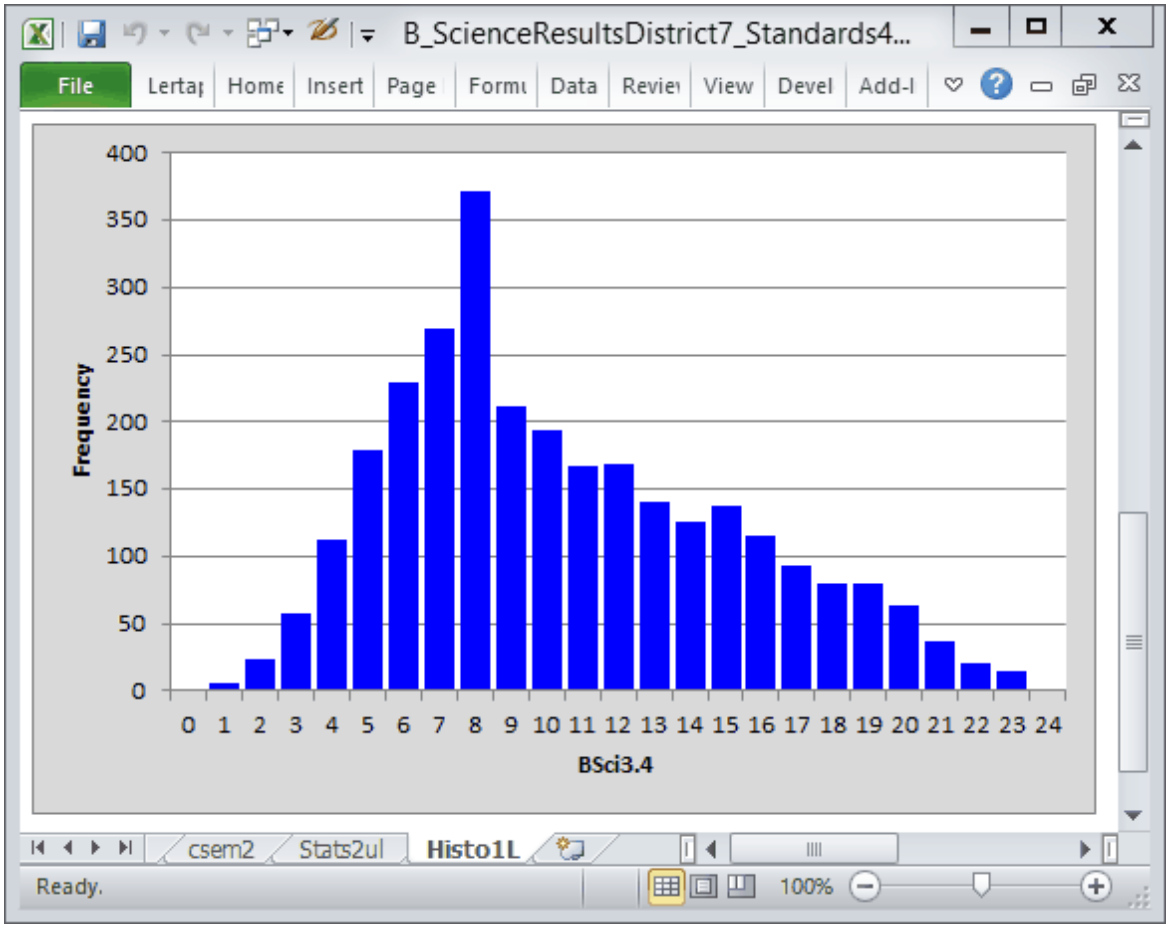
In taking out those students, I have removed a block of students whose item responses were pretty consistent: they all tended to "answer" 9. Coefficient alpha,

an internal consistency statistic, has taken a hit, it's gone down without this block of consistent responses.

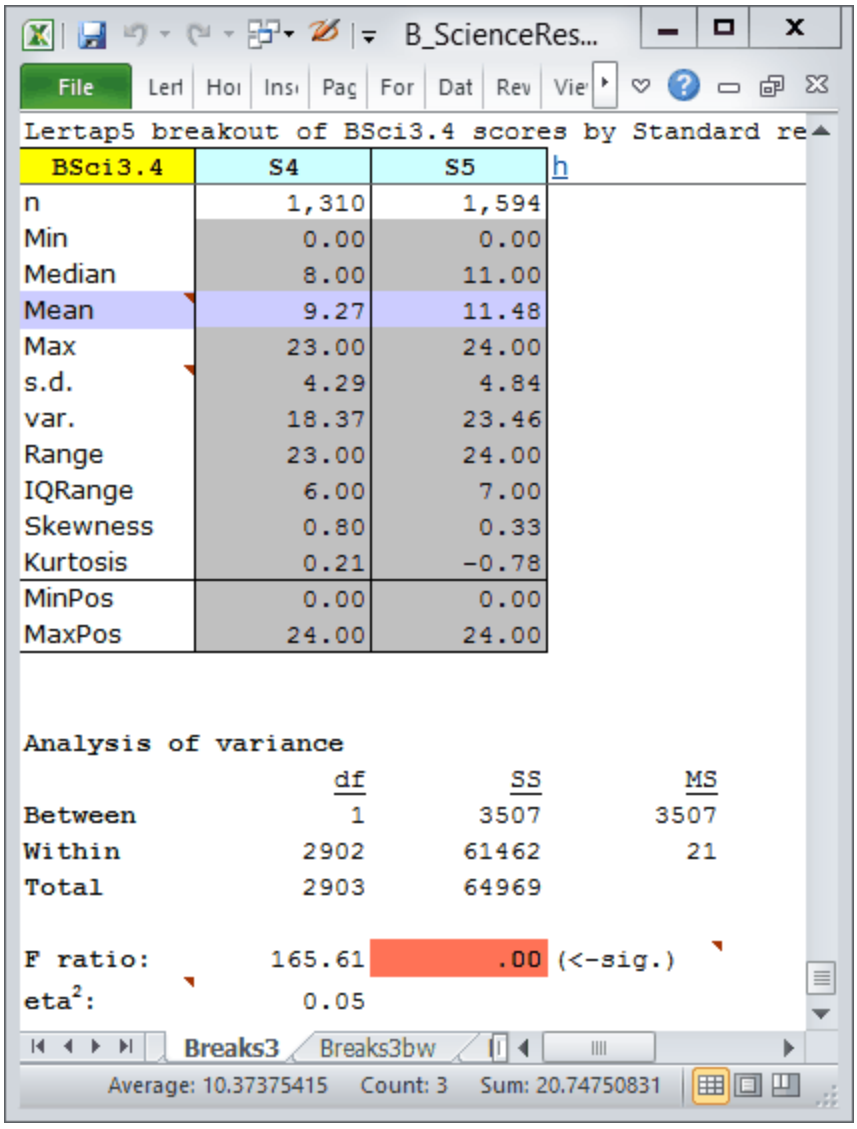
There are some other observations I could make. The items are somewhat easier now that the 9-ers have been removed. Compare the scatterplot above with that seen before the 9-ers were removed: the "blips" have shifted to the right, there are now fewer items with difficulty values below 0.40.

The average test score has gone up. Now, without the 9-ers, it is 10.48. It was 9.31 before.

One more thing. Got time to see a histogram of the test scores?



Got time for one more one more thing? Good -- let's now compare results from the two grade levels, using that whiz-bang "Breakout score by groups" option, getting a **Boxplot** and a couple of score histograms, one for each grade, or "Standard".



B_ScienceRes...

File | Lert | Hor | Ins | Pag | For | Dat | Rev | Vie

Lertap5 breakout of BSci3.4 scores by Standard re

BSci3.4	S4	S5	h
n	1,310	1,594	
Min	0.00	0.00	
Median	8.00	11.00	
Mean	9.27	11.48	
Max	23.00	24.00	
s.d.	4.29	4.84	
var.	18.37	23.46	
Range	23.00	24.00	
IQRange	6.00	7.00	
Skewness	0.80	0.33	
Kurtosis	0.21	-0.78	
MinPos	0.00	0.00	
MaxPos	24.00	24.00	

Analysis of variance

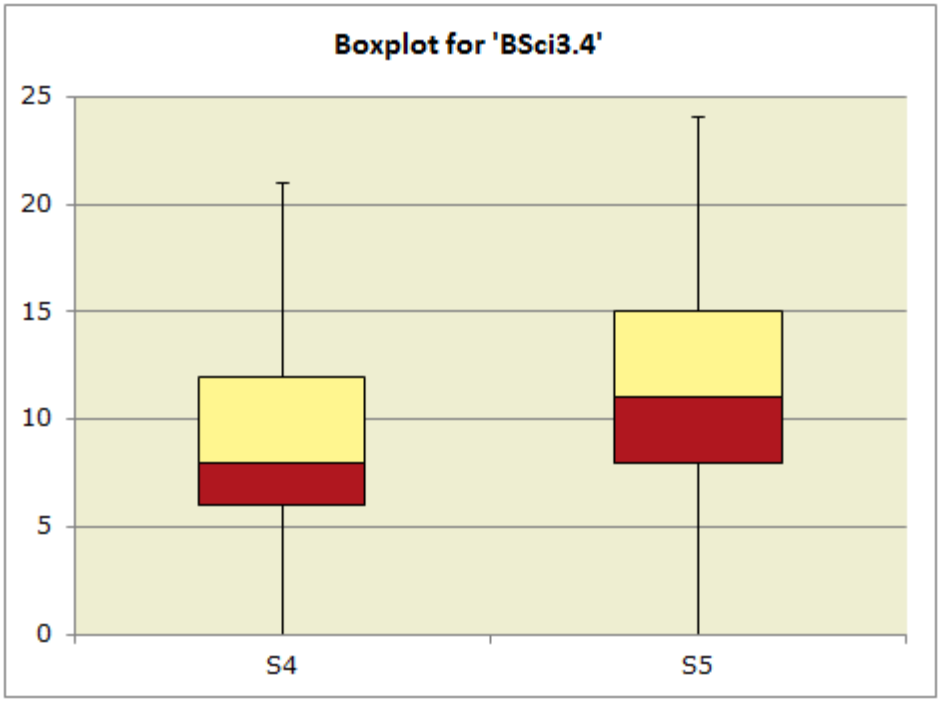
	df	SS	MS
Between	1	3507	3507
Within	2902	61462	21
Total	2903	64969	

F ratio: 165.61 .00 (<-sig.)

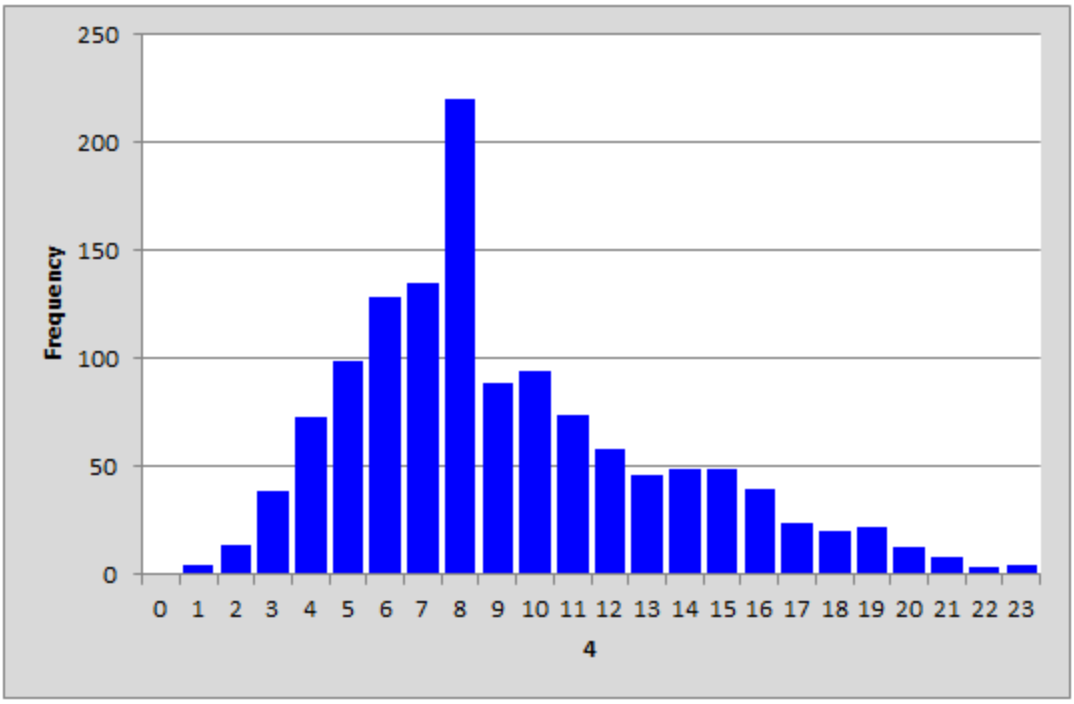
eta^2: 0.05

Breaks3 | Breaks3bw

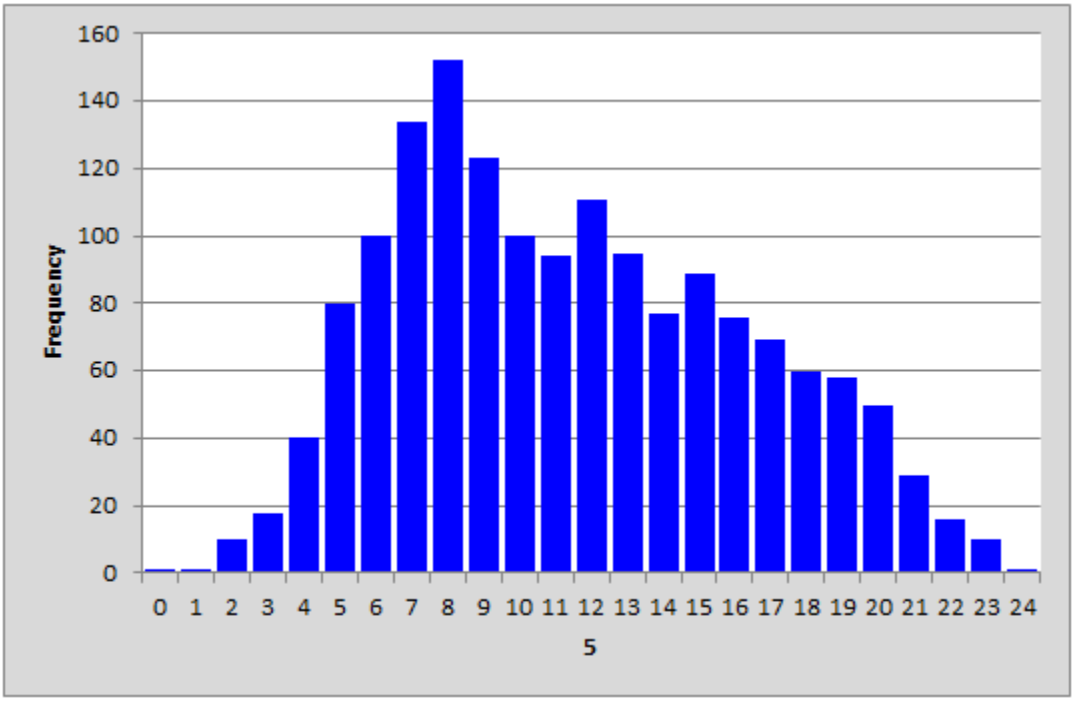
Average: 10.37375415 Count: 3 Sum: 20.74750831



The score histogram for Standard 4:



The score histogram for Standard 5:



The Boxplot and the two histograms were made using the Breaks3 report (usually I have a Breaks1 report, but, in the background, I had been messing around with some other breaks before I came to this spot in the story; each time the "Breakout score by groups" option is taken, it looks to see if there are already some Breaks reports, and, if so, it finds the number of the last one, such as Breaks1, and increases the report number by one, making Breaks2, or Breaks3, or ...).

Read more about Boxplots here, and how to make histograms from a Breaks report here.

Back to my question for this topic: how did the 9s affect reliability? They inflated it. With the 9s in the picture, reliability (coefficient alpha) was .841. It went down to .784 once I had extracted the 9s.

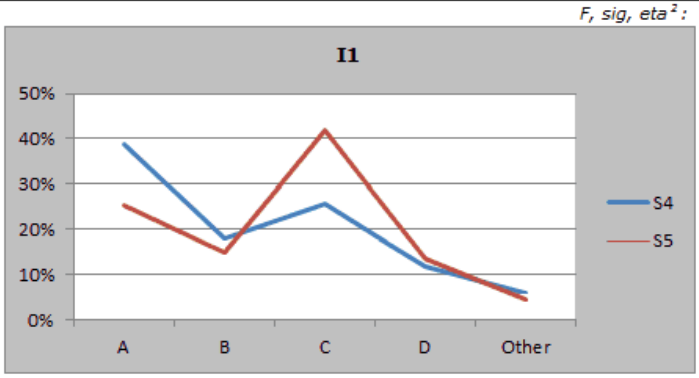
I extended this topic by looking at grade-level differences in the test score. The higher grade, Standard 5, had better test scores; the test was harder for the Standard 4 students. Almost as an aside, I also looked for item-level differences between the grade levels. Results are in the next topic.

3.1.3.1 Item-level results

The "item responses by groups" option will let me compare Standard 4 and Standard 5 differences by item.

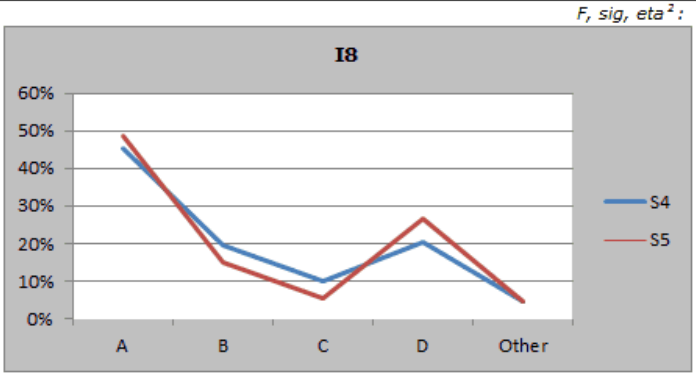
Here are some examples.

I1	A	B	C	D	Other	n	mean	s.d.
S4	39%	18%	**25%**	12%	6%	1310	0.25	0.44
S5	25%	15%	**42%**	14%	5%	1594	0.42	0.49
					F, sig, eta²:	*86.94*	*0.00*	*0.03*



S5 (Standard 5) students were more capable of identifying the correct answer on the first item, where 42% of S5 kids got the right answer, compared to 25% of the S4 students. Other items where the difference was at least 10 percentage points: I2, I3, I5, I7, I10, I11, I13, I14, I18, and I24. The S5 students did better on all of these items by at least 10 percentage points.

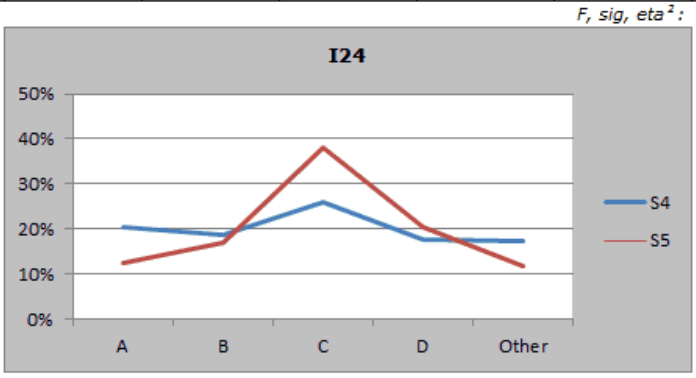
I8	A	B	C	D	Other	n	mean	s.d.
S4	**45%**	20%	10%	20%	5%	1310	0.45	0.50
S5	**48%**	15%	6%	27%	5%	1594	0.48	0.50
					F, sig, eta²:	*3.01*	*0.08*	*0.00*



The eighth item, I8, was one of the few where response differences were slight. However, on all 24 items, S5 students had an edge; there was not a single item where the S4 students did better.

The "**Other**" column in these little tables indicates the percentage of students who had an "answer" of 9 to an item. For the first few items, these percentages were quite similar, as they are above for I1 and I8. But, towards the end of the test, S4 students were more likely to leave questions unanswered. Look at the results for the last item:

I24	A	B	C	D	Other	n	mean	s.d.
S4	20%	19%	**26%**	18%	17%	1310	0.26	0.44
S5	13%	17%	**38%**	20%	12%	1594	0.38	0.49
					F, sig, eta²:	*49.38*	*0.00*	*0.02*



Age a bit in the next topic.

3.1.3.2 Age breakout

I wish now that I had suggested a reason for saying that it might be useful to look for Age differences. I didn't, but now I'll sneak this question in: "Were there age differences in test scores?".

What I'm thinking of is the situation where students are held back a grade for one reason or another. Or, the case where some students started their schooling later than others. When this happens a classroom will of course have a mixture of ages -- indeed, back a few topics I got Excel to create a "pivot table", and there I saw considerable age ranges in both of the grade levels I've been looking at (Standard 4 and Standard 5).

Where in the Data worksheet is the Age variable, or field? It's in column 3 (see it here).

That pivot table indicated that ages ranged from 8 years to 14. Okay, now I'd like to see how test scores varied over the age levels.

Where are the test scores? In the Scores worksheet. They're labeled as "**BSci3.4**". See here.

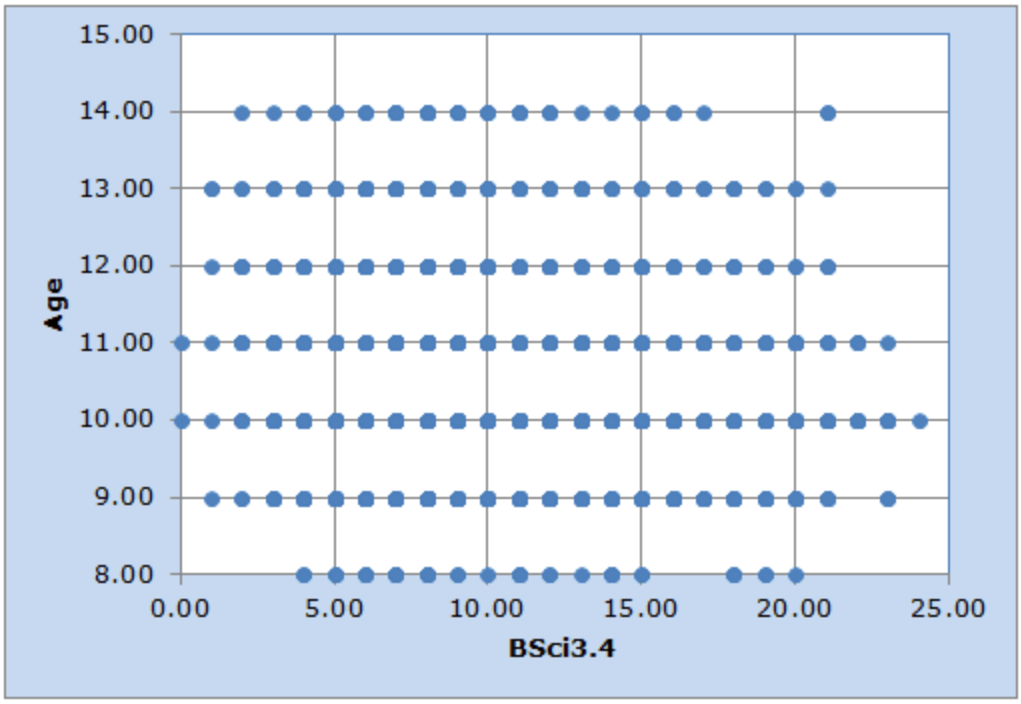
(Someone in the audience has sent a text message, asking why the test scores are known as BSci3.4. I gave the scores this title in the CCs worksheet, which you may see here.)

Were there age differences in test scores?

There are two ways I might go about answering this question with Lertap. The quickest way is to ask for a "scatterplot" of test score by age. Lertap's scatterplotter is one of the options in the "Graphics trio" on the Lertap Excel ribbon tab.

The scatterplotter wants to have both of its variables in the Scores worksheet, so I need to copy the Age column from the Data worksheet to the Scores worksheet. Easy peasy, I did this sort of thing earlier, when I wanted to move the 9s filed from Data to Scores.

Once I've Aged the Scores worksheet, I click on the Scatterplotter icon on the Lertap tab, and before I can go to the galley for another cup of tea, I see this:



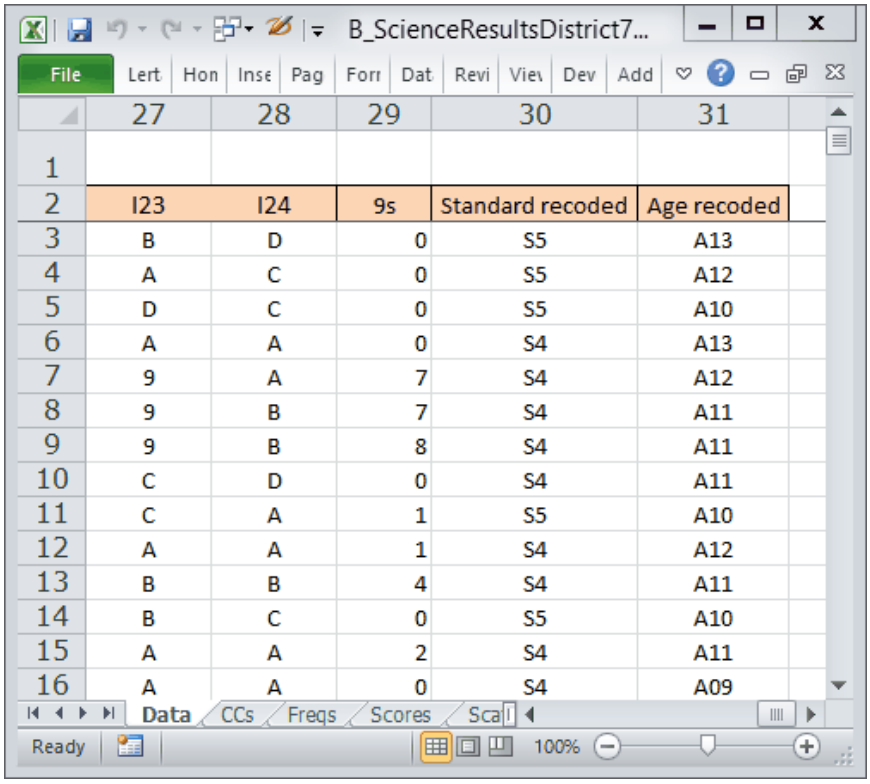
This graph may be pretty, but it's only marginally informative. The number of students behind each of the blips in the scatterplot is not at all apparent. Take, for example, the blip seen at Age = 10.00 and BSci3.4 = 5.00. There may be tens of, or perhaps hundreds of age 10 students with a test score of 5.00. We can't tell.

So, I'm going to take another option. I'll ask for a "Breakout score by groups" where, this time, the groups are defined by the Age variable.

However (alert! alert!), there can be a problem when Excel is asked to make some of its charts when the x-axis variable is not categorical. This problem is mentioned in the caveat seen here.

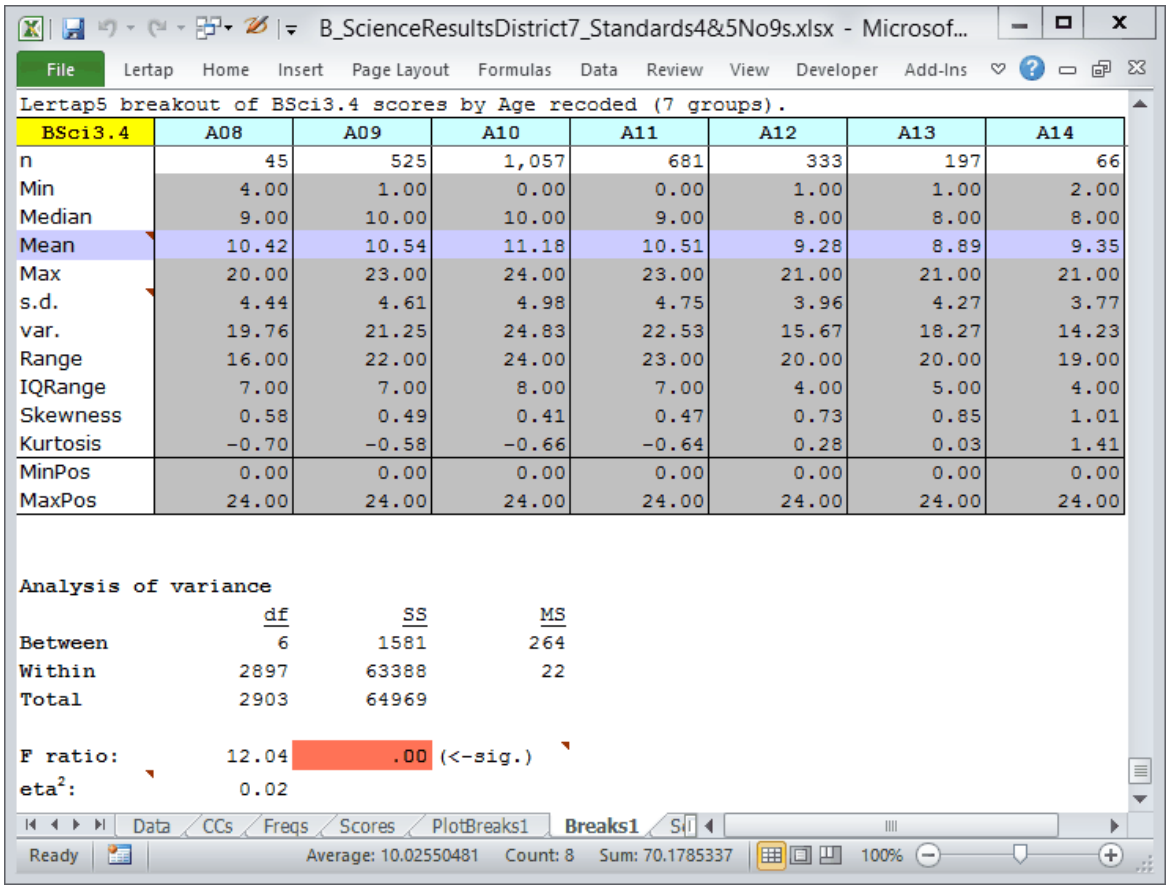
Accordingly, I make use of Lertap's fun-to-use recode option, and add another column to the Data sheet where the 8-year-olds are recoded as age "A08"; I'll also make new age codes of A09, A10, A11, A12, A 13 and A14.

Now the right-most side of my Data worksheet looks like this:



	27	28	29	30	31
1					
2	I23	I24	9s	Standard recoded	Age recoded
3	B	D	0	S5	A13
4	A	C	0	S5	A12
5	D	C	0	S5	A10
6	A	A	0	S4	A13
7	9	A	7	S4	A12
8	9	B	7	S4	A11
9	9	B	8	S4	A11
10	C	D	0	S4	A11
11	C	A	1	S5	A10
12	A	A	1	S4	A12
13	B	B	4	S4	A11
14	B	C	0	S5	A10
15	A	A	2	S4	A11
16	A	A	0	S4	A09

Ready? I take the "Breakout score by groups" option, and this is what I get:



B_ScienceResultsDistrict7_Standards4&5No9s.xlsx - Microsof...

File Lertap Home Insert Page Layout Formulas Data Review View Developer Add-Ins

Lertap5 breakout of BSci3.4 scores by Age recoded (7 groups).

BSci3.4	A08	A09	A10	A11	A12	A13	A14
n	45	525	1,057	681	333	197	66
Min	4.00	1.00	0.00	0.00	1.00	1.00	2.00
Median	9.00	10.00	10.00	9.00	8.00	8.00	8.00
Mean	10.42	10.54	11.18	10.51	9.28	8.89	9.35
Max	20.00	23.00	24.00	23.00	21.00	21.00	21.00
s.d.	4.44	4.61	4.98	4.75	3.96	4.27	3.77
var.	19.76	21.25	24.83	22.53	15.67	18.27	14.23
Range	16.00	22.00	24.00	23.00	20.00	20.00	19.00
IQRange	7.00	7.00	8.00	7.00	4.00	5.00	4.00
Skewness	0.58	0.49	0.41	0.47	0.73	0.85	1.01
Kurtosis	-0.70	-0.58	-0.66	-0.64	0.28	0.03	1.41
MinPos	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MaxPos	24.00	24.00	24.00	24.00	24.00	24.00	24.00

Analysis of variance

	df	SS	MS
Between	6	1581	264
Within	2897	63388	22
Total	2903	64969	

F ratio: 12.04 .00 (<-sig.)

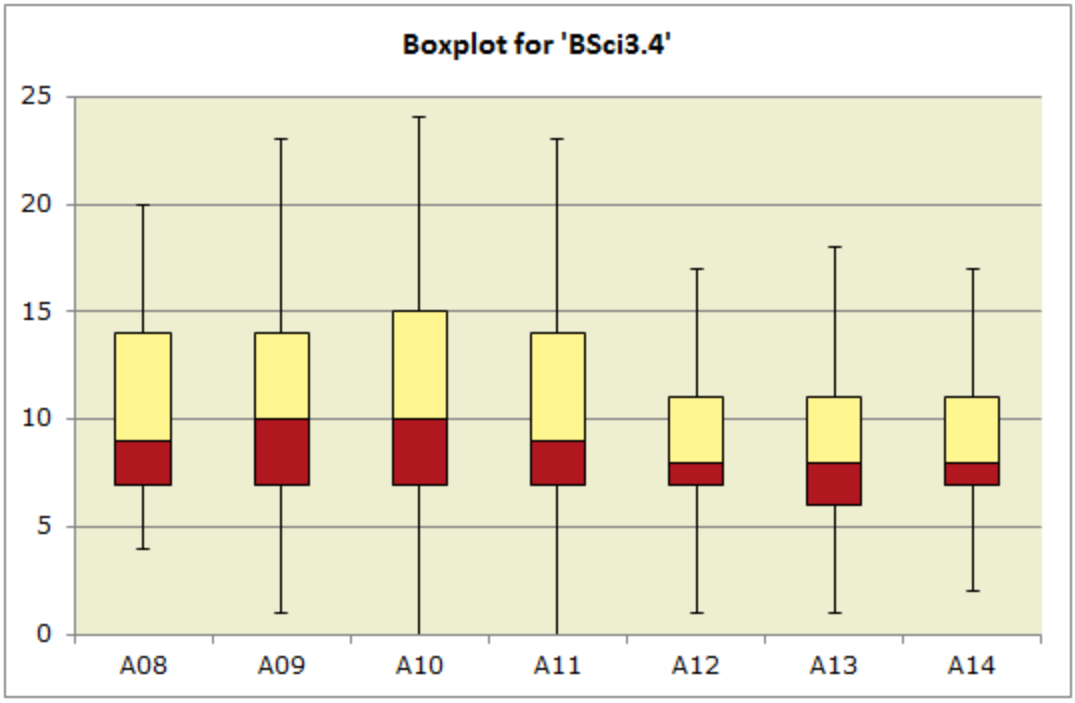
eta^2: 0.02

Data CCs Freqs Scores PlotBreaks1 Breaks1 Si

Ready Average: 10.02550481 Count: 8 Sum: 70.1785337 100%

Just looking at the Median row, at this point it does not seem to me that there are substantial score differences by age level. What is of some interest is the "IQRange" row (the inter-quartile range, the difference between the 75th and 25th percentiles of the score distribution). The scores seem to cluster closer together in the A12, A13, and A14 age levels.

I need a Boxplot.



Were there age differences in test scores?

Well, certainly! The groups have different medians, means, and score ranges. There are clearly differences.

The question must be refined: Were there *meaningful* age differences in test scores? As a principal, should I be concerned about the differences noted?

I might be. The A12, A13, and A14 students did not do as well as the others on this test. If I were to see the same pattern repeated on other tests, and in different subject areas, I might get the staff together and discuss the situation, asking if we need to perhaps lay on some remediation for the older students.

But note: this presupposes that the test had adequate statistical properties. If I've got a test with poor reliability, I would tend to downplay these results. Putting on remedial sessions takes resources, and my budget isn't that healthy. I want to be looking at test scores that have come from good quality tests.

Another note: isn't the value of Lertap/Excel's charts apparent here? The table of statistics is fairly lifeless when compared to the Boxplot.

Another another note. If you're familiar with analysis of variance methods, and with tests of "significance", you will have perhaps noted above that the differences in group means are statistically significant at the ".00" level. This is an artifact of

sample size in this case; hundreds of student test scores are involved here. What you might take in, something more useful, is the little eta^2 statistic. At .02, it's suggesting that the differences are not all that large (but they could nonetheless be meaningful to me as a principal). Read more about eta^2 at the end of this topic. Finally, if these students constitute the population of interest, and are not a carefully-selected sample from a population, then there's no need to look for statistical significance, the F ratio will have no relevance whatsoever -- but eta^2 will.

What's next?

4 Part 3: item analysis

Well here we are, at last. As I may have mentioned, I always check on the quality of the data I have on hand before getting into the various statistical and graphical summaries which Lertap and Excel produce.

I started out my data quality snooping in the original Data worksheet with 3,393 students. I found four students without a valid Gender code, and eliminated them, leaving 3,389 students.

I then noticed that some of the students appeared to be leaving many questions unanswered. I created a "9s score", a count of the number of questions a student did not answer. I found that almost 500 students had failed to answer twelve or more items, half the number of items on the test. I eliminated them (well no, not the students themselves, just their data records), leaving a final Data sheet with 2,904 students.

Now it is true that before arriving at this topic I went ahead and used the scores from this 24-item test in a couple of ways. For example, I looked for grade-level differences, and for age differences. I really shouldn't have done this before checking on the quality of the test itself. I'll turn to that now.

As I do, I will assume that the test is one that's meant to discriminate, that is, used to identify the strongest students, separating them from the weaker ones. Such tests are commonly used as part of the process of assigning an achievement descriptor to students, an indicator of how well they have done, such as "excellent", "good", "adequate", and "poor" (or, perhaps, A, B, C, and D grades). (Another common type of test is one which uses some sort of cut-off score to classify students as "masters"/"non-masters", or "pass"/"fail". Lertap has special tools for looking at these tests; such tools are described elsewhere.)

Tests meant to discriminate should have good reliability. For a test to have good reliability, its items have to have a demonstrated ability to discriminate; their discrimination "index" should be high.

The following topics will look at item discrimination and test reliability, using both tables and graphs. As you'll see, a couple of the test's items could have performed better, and, if they had, the test's reliability would have been a bit better.

Another look at reliability is next.

4.1 Reliability

I begin my investigation of item and test quality by clicking on Lertap's Interpret option.

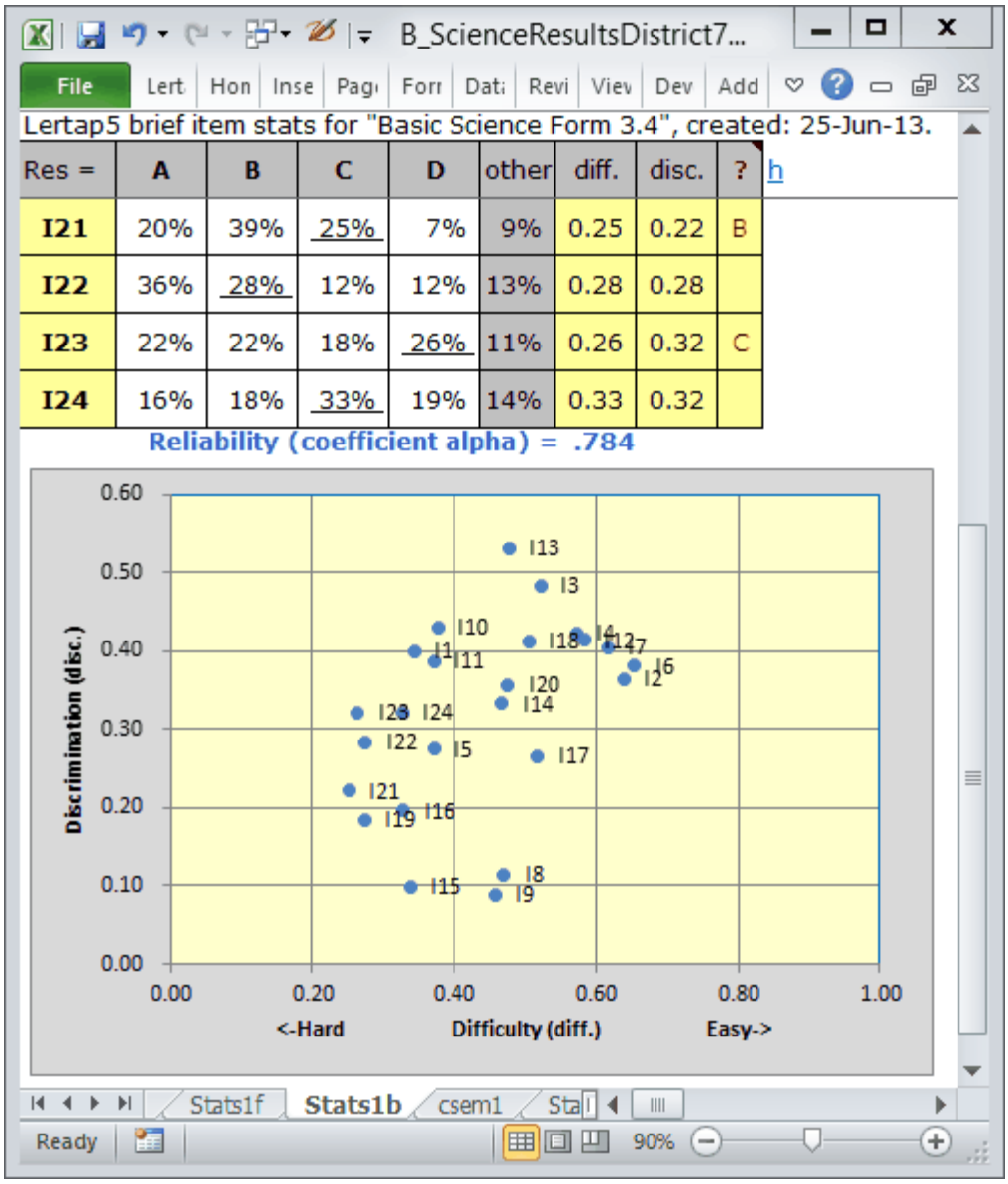
Lertap and Excel team up to make sure they can understand, or "interpret", the control lines I've placed in my CCs worksheet. All being okay, these two then get together to make the "Freqs" report. (There's actually a bit more that happens. If you want to be full bottle, read this.)

Mention has already been made of how I use Freqs as a check on data quality. It's invaluable to me, always my first stop, and the reason I don't run in "production mode": I want to make sure my data are free of data preparation and processing errors before going on to bigger and better things. Even when data records have been generated by the use of a scanner there are often errors, or, at least, unexpected results.

Next I go for Lertap's Elmillon option. It links to computer code which was originally written for the Venezuelan Ministry of Education in 1972. Elmillon has been enhanced over the years -- when it was born it produced just one statistical summary, now known as the "**Stats1f**" report (**f** for full). Later the "**Stats1b**" report was added (**b** for brief); it is a condensed version of Stats1f with a handy little chart at the bottom, a scatterplot of item difficulty by item discrimination. Still later another report, "**Stats1ul**", was put on basically as a means of getting item response plots, often referred to in Lertap as "quintile plots" (**ul** for upper-lower).

The information in these reports is not really unique. Each report looks at the same thing, how students have responded to the items, but they summarize the results in different ways. I know that not all users make use of all of these reports. That's fine; it is possible to find out how the items have performed by using just one of these three reports -- you choose. I use Stats1b as a starter, and then usually make packed quintile plots. After this I'll brew a fresh cup of coffee and go down the Stats1f report.

When Elmillon completes its run, Excel will focus on the bottom of the Stas1b report, as exemplified below:



Res =	A	B	C	D	other	diff.	disc.	?
I21	20%	39%	25%	7%	9%	0.25	0.22	B
I22	36%	28%	12%	12%	13%	0.28	0.28	
I23	22%	22%	18%	26%	11%	0.26	0.32	C
I24	16%	18%	33%	19%	14%	0.33	0.32	

The reliability figure, coefficient alpha, is low at .784. It should be over .80; ideally over .85; ideally ideally over .90.

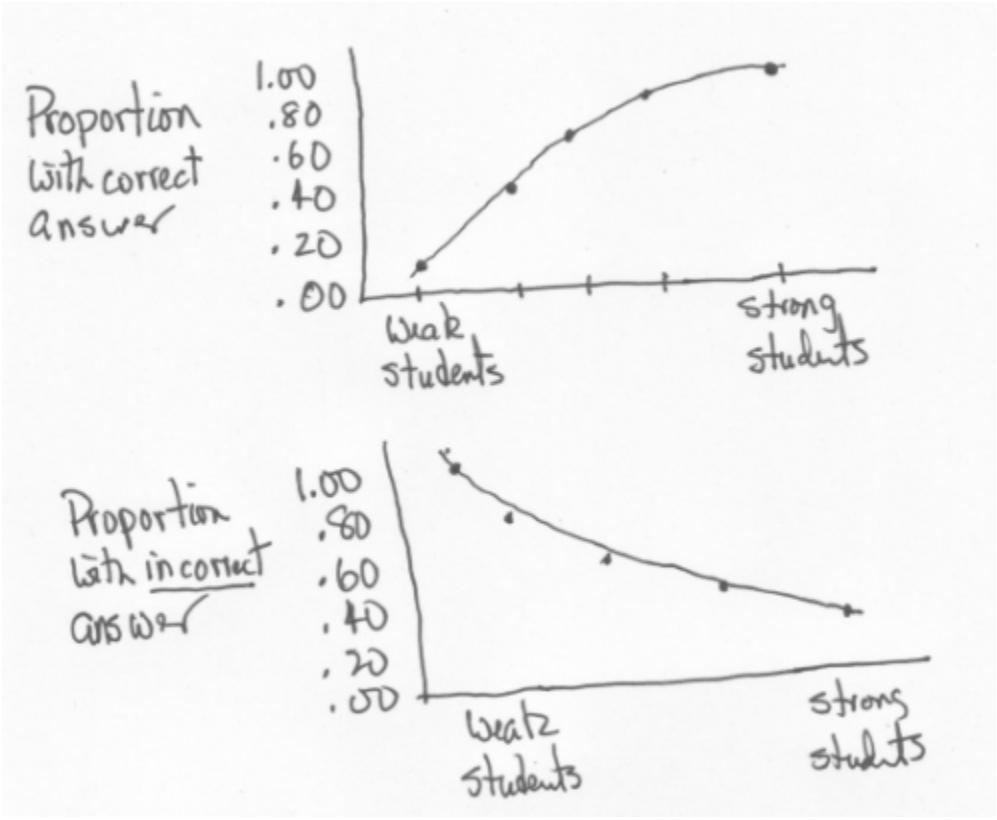
Can you make out the labels on the three lower blips in the chart? Items I8, I9, and I15 have the lowest item discrimination figures ("**disc.**"), with values around 0.10. These items have undoubtedly served to bring down coefficient alpha. When all of a test's items have discrimination figures at or above .30 we will, generally, be in better shape, and be rewarded with a higher reliability figure.

Later on I'll show how we can improve I8's discrimination so that it rises above .30. First, I want some "packed quintiles". I show how to get them in the next topic.

4.2 Packed quintiles

Let me come back to this matter of having "discriminating items". If the purpose of the test is to allow us to identify the best students (and, yes, also the weakest students as we might want to provide special tutorials for them), then, the students who get any given test item correct should be the strongest students. They should be the "most proficient" students. Other students should get the item wrong, especially the very weakest of the weakest students.

I step over to the whiteboard and, with a black marker pen, artfully create the following sketches for your admiration:



These two wonderfully hand-sculpted graphs are meant to display what we expect of a discriminating item.

Let's say we take a group of students and separate them into, say, five groups according to their proficiency level, which we somehow know in advance. Then we plot the proportion of students in each group who get the item correct (top sketch).

We make another plot, the proportion of students in each group who give an incorrect answer to the item (bottom sketch).

These sketches depict what we want of a discriminating item. The chances of a student getting an item correct depend on the student's ability level, on her (or his) degree of proficiency. The stronger the student, the more we expect him (or her) to get our item correct. On the other hand, those students who choose an incorrect answer should be the weaker students. (We assume that the items, as a whole, are neither too easy nor too difficult for the students.)

You okay with this? If not, take a break and come back later. I'm going to step up the action by showing you I13's profile in Lertap's Stats1ul report.

Before I do that let me tell you more about Stats1ul. The "ul" means upper-lower. Lertap's first step in creating this report is to sort all of the test scores from highest to lowest, from upper to lower. It then picks out the top 20% of the scores, and notes the top and bottom test score for the students in this 20%. This it calls the "upper" group. The top and bottom test scores are the score boundaries for the upper group.

B_ScienceResultsDistrict7_Standards4&5No9s.xlsx

Lertap5 U-L stats for "Basic Science Form 3.4", created: 25-Jun-13.

Res =	A	B	C	D	other	U-L diff.	U-L disc.	h

Summary group statistics

	n	avg.	avg%	s.d.	min.	mdn.	max.
upper	580	17.9	74%	2.1	15	18	24
2nd	580	12.9	54%	1.2	11	13	15
3rd	584	9.5	39%	0.9	8	9	11
4th	580	7.4	31%	0.6	6	7	8
lower	580	4.8	20%	1.2	0	5	6
everyone	2,904	10.5	44%	4.7	0	9	24

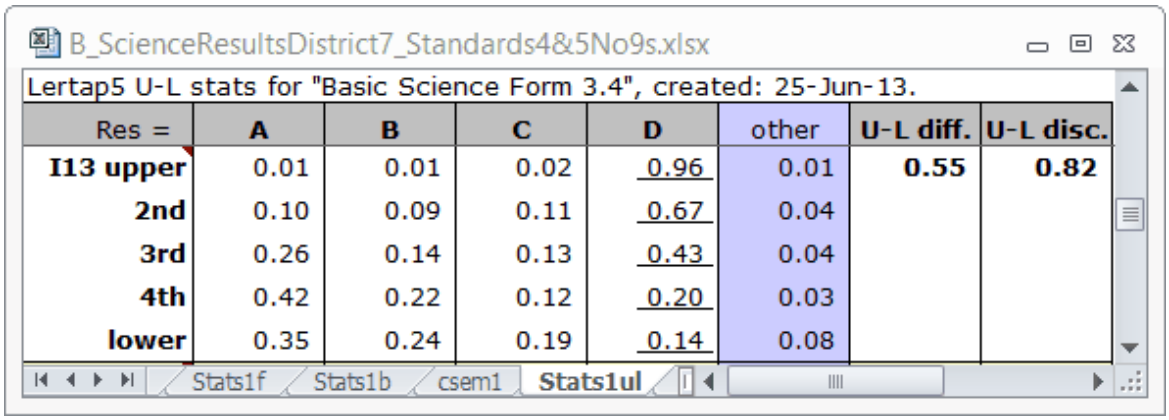
Stats1f | Stats1b | csem1 | **Stats1ul** | Sta

In the table above we see that there are 580 students in the "upper" group, about 20% of the 2,904 students tested. The lowest test score ("min.") in this group was 15, while the highest ("max.") was 24.

Having picked out the top 20%, and found corresponding "min." and "max." score boundaries, Lertap and Excel then get together to identify the next-best 20% of the students, which they call the "2nd" group. This continues into the "3rd", "4th", and "lower" groups.

The "avg." column in the table indicates the average test score found in each group; "avg%" expresses "avg." as a percentage of the maximum possible test score. The "mdn." column is the median test score for each group.

Right, then. Lertap's next step is to find response proportions for each item within each group, and to make another little table for us. Here's the table for I13:



B_ScienceResultsDistrict7_Standards4&5No9s.xlsx

Lertap5 U-L stats for "Basic Science Form 3.4", created: 25-Jun-13.

Res =	A	B	C	D	other	U-L diff.	U-L disc.
I13 upper	0.01	0.01	0.02	0.96	0.01	**0.55**	**0.82**
2nd	0.10	0.09	0.11	0.67	0.04		
3rd	0.26	0.14	0.13	0.43	0.04		
4th	0.42	0.22	0.12	0.20	0.03		
lower	0.35	0.24	0.19	0.14	0.08		

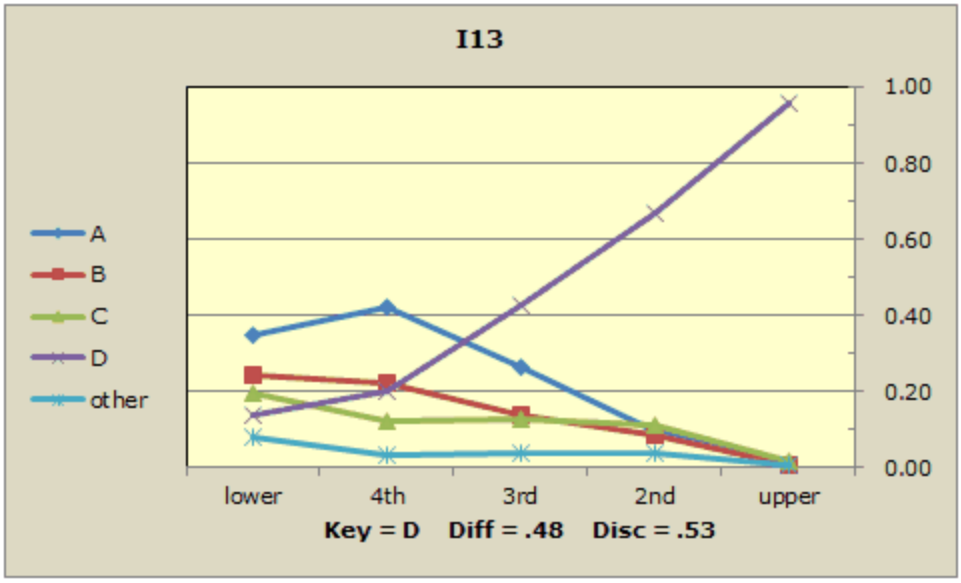
Stats1f | Stats1b | csem1 | **Stats1ul**

The correct answer to I13 was D. The D column is underlined to signify this. Now, look down this column, and refer back to my first sketch above. We've got the pattern wanted for a discriminating item. The weakest students, those in the "lower" group, had a proportion correct of 0.14. As we step up from the bottom of column D, going from weak to strong students, the proportion correct steadily increases. Almost all of the students in the top group, the "upper" group, got the item correct (proportion of 0.96).

The other columns in the table correspond to incorrect answers, generally called "distractors". Take one of these columns and step up from the bottom. Do you not observe the pattern seen in my second sketch?

I don't. The second sketch is for just the "incorrect answer". I13 has three "incorrect answers". To get a graph which resembles my second sketch, we should sum the proportions for the three incorrect answers. In the lower group the sum is 0.35 + 0.24 + 0.19, or 0.78. If we add to this the proportion who did not answer the item, the "other" column, we find a total proportion of 0.86. If you do this for each group, you'll find that the proportion incorrect goes from 0.81 in the lower group, to 0.80, to 0.57, to 0.34, and then, in the upper group, to just 0.04. We do end up with something akin to my second sketch.

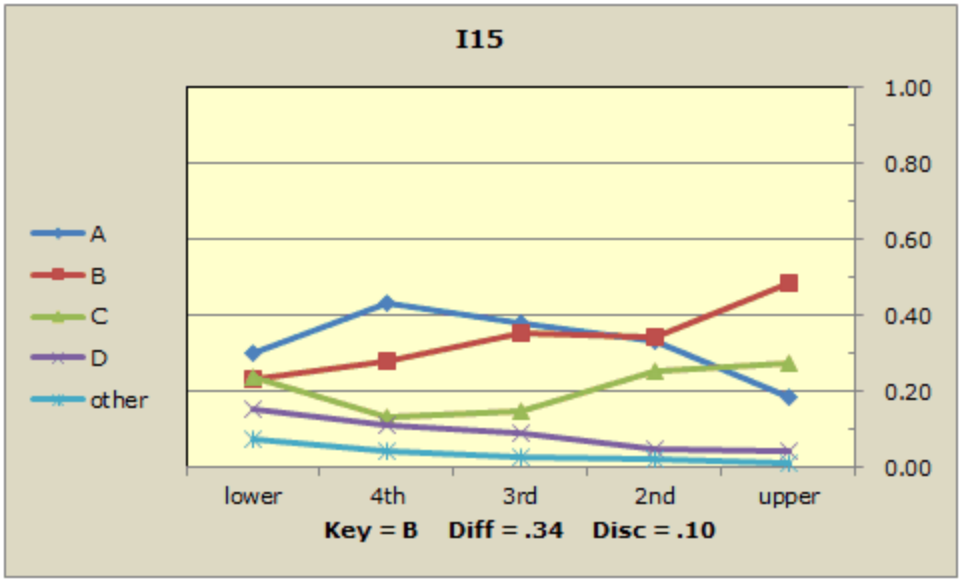
I get a bit bug-eyed looking at Stats1ul tables. To solve this problem I hardly ever look at them at all. Instead I look at the corresponding "quintile plot":



See how the response trace for the correct answer, D, starts low (not many in the lower group got the item correct) but, as student proficiency increases, rises to almost 1.00 in the upper group (where just about everyone got the item correct)? The other trace lines indicate the popularity of each of the distractors, the incorrect answers. Distractor A was especially popular in the lower group (0.35 proportion, or 35%), even more so in the next-lowest group ("4th" group, with 42%), but then, like an aging pop song, its popularity steadily declines. By the time we get to the top group, the "upper" group, only 1% selected this incorrect answer.

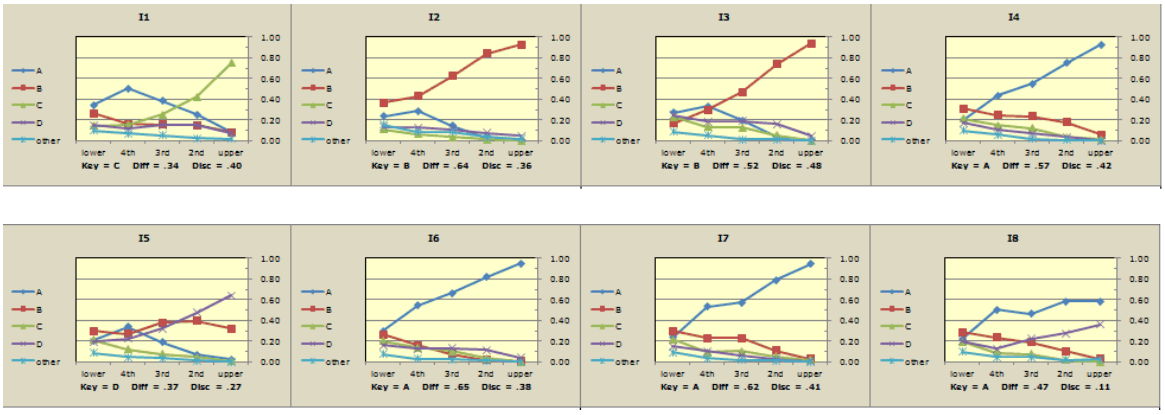
We want a test to discriminate? We want items like I13. The correct answer's popularity starts low but rises to just about 100% among the strongest students. The incorrect answers on the other hand, the distractors, will be relatively popular with the weaker students, but virtually ignored by the strongest. Distractor trace lines will fall off, declining as we go from left to right.

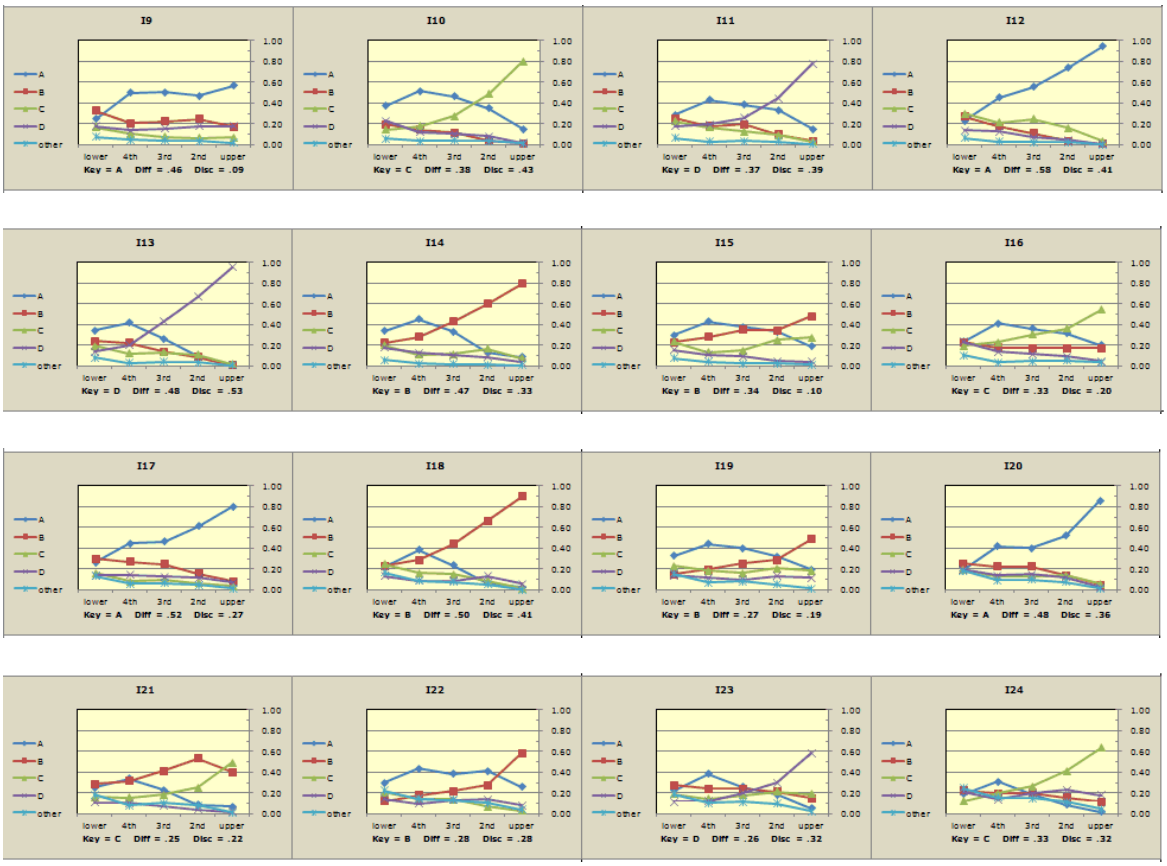
Now, compare this to the fifteenth item, I15:



Does the popularity of the correct answer (B) start low and rise to about 100%? No. It rises quite slowly, ending up at about 50% in the top group. Do the distractors, the incorrect answers, fall away as we go from left to right in the graph? A couple of them do, but one, C, definitely does not. It looks like just under 30% of the top students selected this distractor, thinking it to be the correct answer. Another distractor, A, does decline in popularity, but nonetheless was selected by almost 20% of our best students. I15 is not a good discriminator.

Know how I said I get somewhat bug-eyed looking at the tables seen in Stats1ul reports? A bit of the same happens when I look at quintile plots. I know that this feeling is not shared by many Lertap users, but me?, well, I very much prefer the gestalt offered by "packed quintiles".





Which of these 24 items display the response pattern desired of a discriminating test?

In my opinion, clear winners are items I1, I2, I3, I4, I6, I7, I10, I11, I12, I13, I14, I17, I18, and I20.

Clear losers might be I5 (one distractor is too popular in upper group), I8 (ditto), I9 (correct answer not high enough in upper group), and items I15, I16, I19, and I21 (all have distractors which fail to decline sufficiently).

The pattern seen in the last three items, I22, I23, and I24, is interesting. These items were probably just too difficult for the students who sat this test: all distractors are heading down on the right (they're "falling away"), and the response trace for the correct answer is rising nicely by the time we hit the upper group -- it's almost as if these three plots would come really good if we could get another group of students more proficient in the subject matter than those in the present upper group.

Our 24-item test would undoubtedly have had better reliability were it not for the seven items in my "clear losers" category. Before this test is used again, I'd get the item writer team together and go over things in a special workshop. Our goal would be to look at the "clear losers" and try to figure out what went wrong with them.

Before doing this, however, I'd make a bit more use of the Stats1b report in combination with Stats1f. Let me show you what I'd do. (Please see the next topic.)

4.3 Stats1b and 1f

The Stats1f report was the only summary of item and test quality made by the initial versions of Lertap. It goes back to the days when an attempt was made to train teachers in the development of multiple-choice tests, and in methods for assessing item and test quality. It used to be that most universities and teacher-training institutes required students to take at least one course in "tests and measurement". I think those days may be gone, but the Stats1f report, made as a tool for teachers to use back then, lives on (in an improved format).

It is certainly possible to gauge test quality by looking at the bottom of the Stats1b report to see test reliability, and the scatterplot of item difficulty by discrimination, followed by a study of quintile plots. I will not say that everyone should look at Stats1f.

But humor me for a moment. Let me Stats1f you.

Here's what Stats1f has to say about one of our star items, I13, found in column 17 of the Data worksheet:

I13 (c17)

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	659	0.23	-0.34	-0.47	7.53	-0.62
B	0.00	406	0.14	-0.25	-0.40	7.51	-0.63
C	0.00	335	0.12	-0.16	-0.27	8.33	-0.46
D	1.00	1,388	0.48	0.53	0.67	13.47	0.63
other	0.00	116	0.04	-0.11	-0.24	8.04	-0.52

The "**wt**." column indicates the number of points a student gets for choosing one of the item's options. The table says that those students who select option D will get one point. Option D is the correct answer for I13, and 1,388 students selected it. The "**p**" column indicates the proportion of students who selected each option; 48% (proportion of 0.48) of our students took option D, the correct answer.

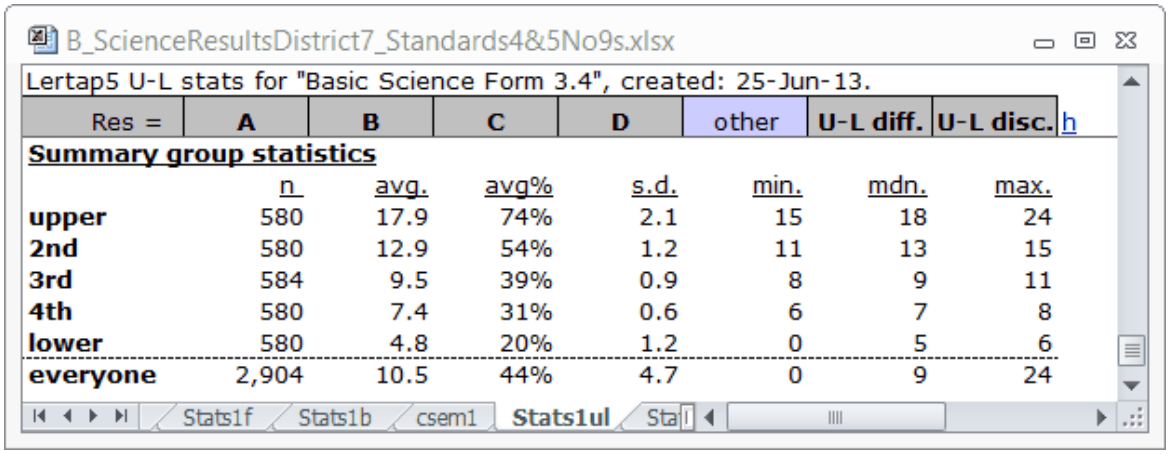
When there is but one right answer to an item, the "p" corresponding to it is the "**item difficulty**". I13's difficulty was 0.48.

I skip you over to the "**avg**." column now. It indicates the average "criterion score" for the students selecting each option. The criterion score is what Lertap uses to make its five groups in the quintile plots. It's usually just the test score, but it can be another score, a score often called an "external criterion".

What was the average test score for the 1,388 students who got I13 correct? It was 13.47, you can see it above. The adjacent column, "**z**" (for z-score), quickly lets me

know if this value is above or below the average score on the whole test. Here z = 0.63, meaning that these students, with their "avg." test score of 13.47, did well on the test as a whole.

I digress for a moment, taking you back to the Stats1ul table you loved so much when you saw it in an earlier topic:



B_ScienceResultsDistrict7_Standards4&5No9s.xlsx

Lertap5 U-L stats for "Basic Science Form 3.4", created: 25-Jun-13.

Res =	A	B	C	D	other	U-L diff.	U-L disc.	h

Summary group statistics

	n	avg.	avg%	s.d.	min.	mdn.	max.
upper	580	17.9	74%	2.1	15	18	24
2nd	580	12.9	54%	1.2	11	13	15
3rd	584	9.5	39%	0.9	8	9	11
4th	580	7.4	31%	0.6	6	7	8
lower	580	4.8	20%	1.2	0	5	6
everyone	2,904	10.5	44%	4.7	0	9	24

Stats1f Stats1b csem1 **Stats1ul** Sta

The "avg." score for the second-highest group, "**2nd**", was 12.9. I can reiterate that the 1,388 kids who got I13 correct, with their "avg." of 13.47, were capable students as their "avg." was above "avg." for the 2nd-highest group.

What was the average test score for "**everyone**", all 2,904 test takers? Lots of hands go up, thank you very much, and yes, you're all right: it was 10.5. Good going.

Back to I13's table of results above. The "avg." score for the 659 students selecting option A was 7.53; we can assume that most of these were in the "**4th**" group, next-to-bottom. A similar story holds for the 406 who went for option B, for the 335 who thought C was the correct answer, and even for the 116 students who did not answer I13.

Note that the "z" score corresponding to each of these options is negative. Items which discriminate will have a positive "z" score for the correct answer, and negative "z"s for all the incorrect answers, that is, for the distractors. The more positive the correct answer's "z", and the more negative the distractor "z"s, the better the item discriminates.

An **item's discrimination** value is "**pb(r)**" for the correct answer, the point-biserial correlation between the item and the criterion score. What is I13's discrimination? Not so many hands go up now as quite a few of you have left (a common happening when the going gets a bit tough). To the four people who remain, yes, thanks, you're spot-on mates, it's 0.53.

These discrimination creatures can have a minimum of zero (no discrimination), to a maximum of one (really terrific discrimination). A test will discriminate well if the great majority of its items have discrimination values of at least 0.30. The higher the better, of course. (Uh-oh, correction required: pb(r) values may be less than zero, they may go as low as negative one; distractors with negative pb(r) values are, in fact, what we want -- you can see 'em in the pb(r) column above. It's possible for the correct answer's pb(r) to be negative too, but, as you might suspect, that's not wanted at all; it will occur when the "avg." score for the correct answer is below the overall test's average. For an example of this highly unwanted but not really very uncommon event, see this interesting dataset when you have time.)

It can be shown that pb(r) is closely related to the differences among those "z"s; the higher the "z" for the correct answer, and the lower the "z"s for the distractors, the better the discrimination. If you don't care much for pb(r), just look at the "avg." column, and the "z" column. One of the "z"s should be positive; the others negative. One of the "avg." values should be higher than the others, hopefully much higher.

(For more about these things, click here for an internet page, and here for Chapter 7 of the Lertap manual, a great read, praised by numerous readers (not all of them members of my family or students looking for a good grade).)

I take you now to I15's summary in Stats1f:

I15 (c19)

option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	0.00	945	0.33	-0.12	-0.16	9.64	-0.18	
B	1.00	982	0.34	0.10	0.13	11.79	0.28	
C	0.00	610	0.21	0.07	0.10	11.13	0.14	<-aa
D	0.00	262	0.09	-0.15	-0.26	8.24	-0.47	
other	0.00	105	0.04	-0.11	-0.27	7.71	-0.59	

Is this a discriminating item? Is one "avg." much higher than the others? No. Is only one "z" positive and all the rest negative? No.

Is this a discriminating item? No.

Look what's happened. We have above-average students selecting the correct answer, B, which is good.

But look, we also have above-average students selecting one of the distractors, C. This we do not want to happen. The students who go for the distractors, the incorrect answers, should be weak students, below average. Two of I15's distractors worked well enough, but that's not good enough. All of the distractors should have "avg." values lower, hopefully much lower, than the "avg." value corresponding to the correct answer. All of the distractors should have negative "z"s.

We're getting to crunch time now. We have, I suggest, a more precise idea of what's wrong with I15; 610 students with an above-average test score selected option C, a distractor.

Look at I8:

I8 (c12)

option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	1.00	1,366	0.47	0.11	0.14	11.57	0.23	
B	0.00	493	0.17	-0.24	-0.36	7.92	-0.54	
C	0.00	220	0.08	-0.23	-0.42	6.71	-0.80	
D	0.00	689	0.24	0.16	0.22	11.85	0.29	<-aa
other	0.00	136	0.05	-0.12	-0.25	8.01	-0.52	

This is worse than I15! The 689 students who selected one of the distractors, D, had better test scores than the 1,366 students who selected the correct answer, A.

Crunch time for the workshop charged with reviewing the test items. Have we scored I8 and I15 correctly? Why are good students selecting bad answers? Are the bad answers not bad?

This is likely. It is ever so possible that the students read something into these two items that the item writers did not see. What the item writers saw as unambiguous, clear-cut items and options, could in fact be murky and ambiguous. We need to discuss this in our workshop: who can see the ambiguity? Often times a staff member can; more often a conversation with the students will uncover the problem. (For more about ambiguity in this context, I suggest searching the internet for "ambiguous multiple-choice questions" -- you should find some sample test items with ambiguities of one sort or another.)

Meanwhile, have the students who selected one of these poorly-functioning distractors been put at a disadvantage? They did not get a point for their answer, but, if there is ambiguity, if the answer they chose has some possibility of being a correct answer, well, they've been robbed of a point. Should we address this issue and give them a point, and say that each of these two questions has two correct answers? Would that not be only fair?

We discuss this in our little workshop. Yes. We decide it would indeed be just to the students to acknowledge potential problems in these two items, ambiguities are there. We'll "double-key" I8 and I15, re-score the test, and tag these items as "do-not-use-again" until we've fixed them.

I show how to double-key test items in the following topic.

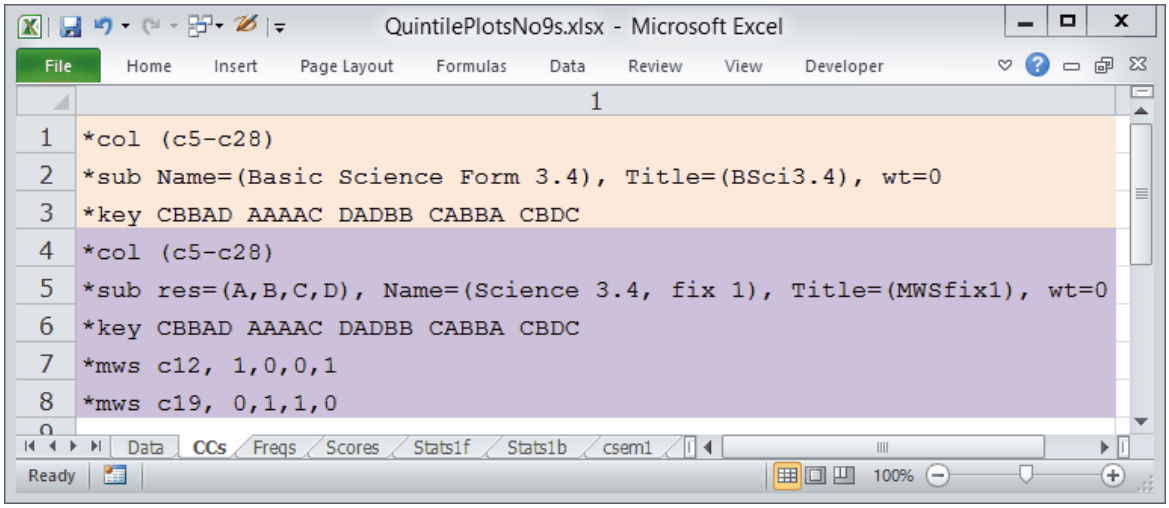
4.3.1 Double keying items

I have two items whose scoring is to be changed: I8 in column 12 (c12) of the Data worksheet, and I15 in c19.

I8 is to have two correct answers, A and D -- a student selecting either of these answers will get a point.

I15 is also to have two correct answers, B and C -- a student selecting either of these answers will get a point.

Ready, set, go. I add some new lines to my CCs worksheet. Behold:



Relax, stay with me, the bell has not rung yet. I'll explain what I've done.

The first three rows are the same as before, except for the little wt=0 at the end of *sub. These three rows define what Lertap will automatically call "Subtest1". Its title is BSci3.4.

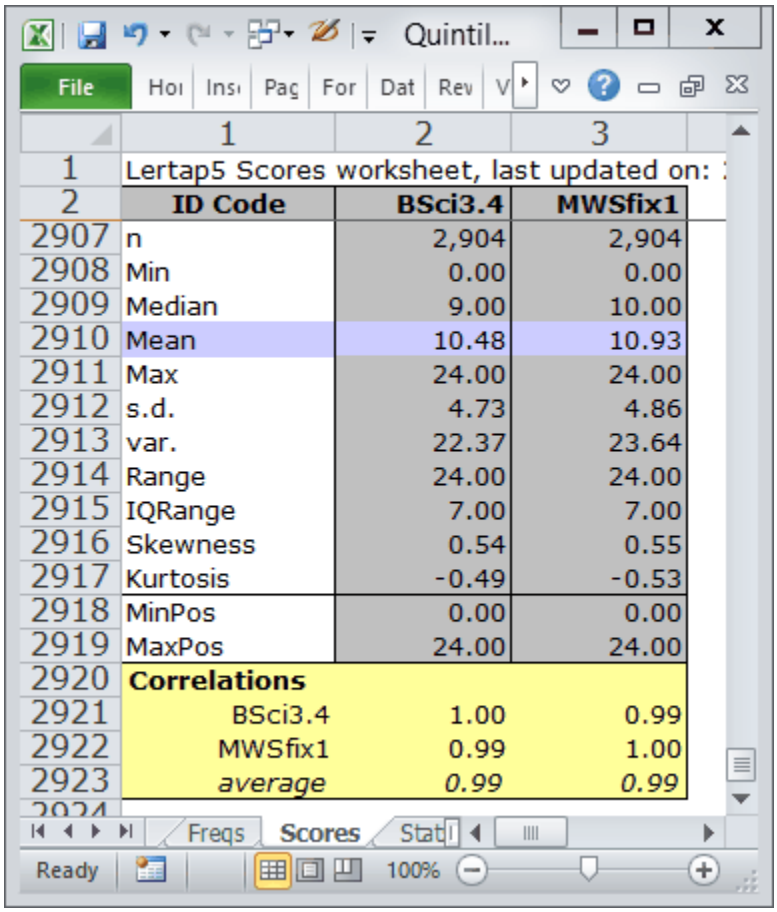
Now I've added another "subtest" with a title of MWSfix1. This subtest uses the same items as the first subtest: the *col lines are the same for each subtest. And the *key lines are the same for each subtest.

The big change is seen in rows 7 and 8, where I've added two *mws lines.

"mws" means multiple-weights specification. For the item found in c12 (column 12) of the Data worksheet, students will get a point if they select either the first or the fourth option. For the item found in c19 (column 19) of the Data worksheet, students will get a point if they select either the second or the third option. (A reference for *mws is here.)

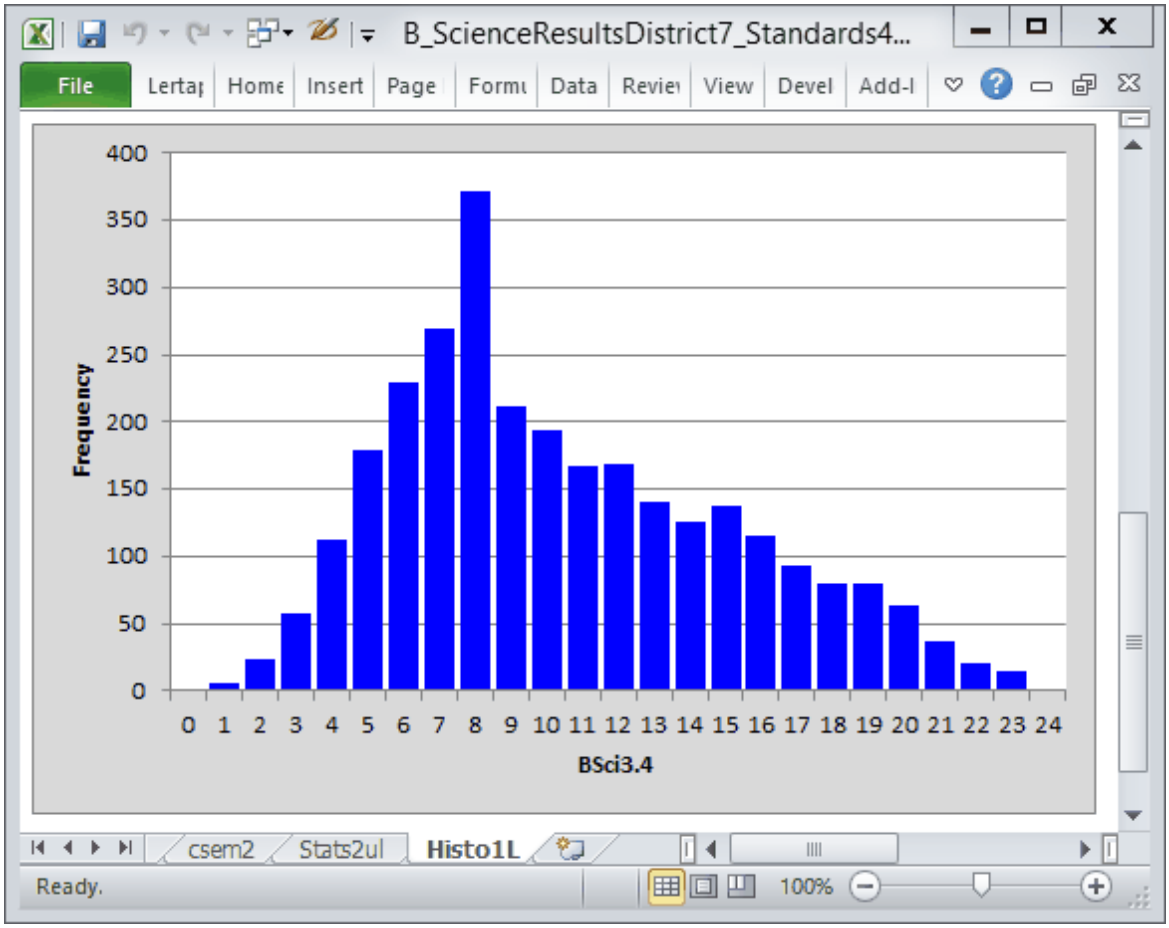
Lertap will now create two scores for each student, a BSci3.4 score, and an MWSfix1 score. The little wt=0 entries at the end of the two *sub lines keep Lertap from adding these two scores together to make a "total score". These wt=0 things are not crucial at all; with them the Scores worksheet will have just two results columns, one for BSci3.4 and one for MWSfix1; without them the Scores worksheet will have three scores, BSci3.4, MWSfix1, and "Total", the sum of BSci3.4 and MWSfix1. In this case adding the two scores doesn't make sense as they're based on the same 24 items. (An example where a total score was wanted is here.)

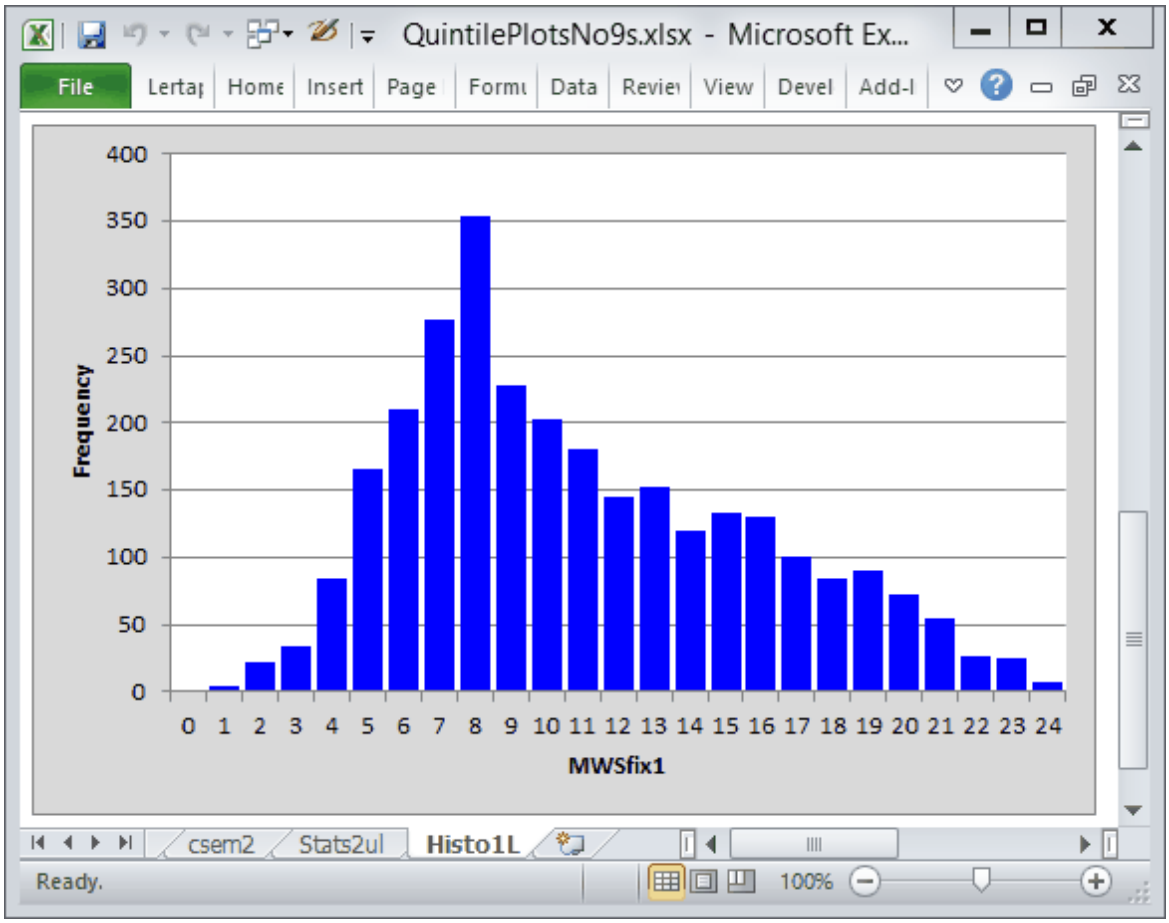
After I've used the Interpret and Elmillon options, the bottom of the Scores worksheet looks like this:



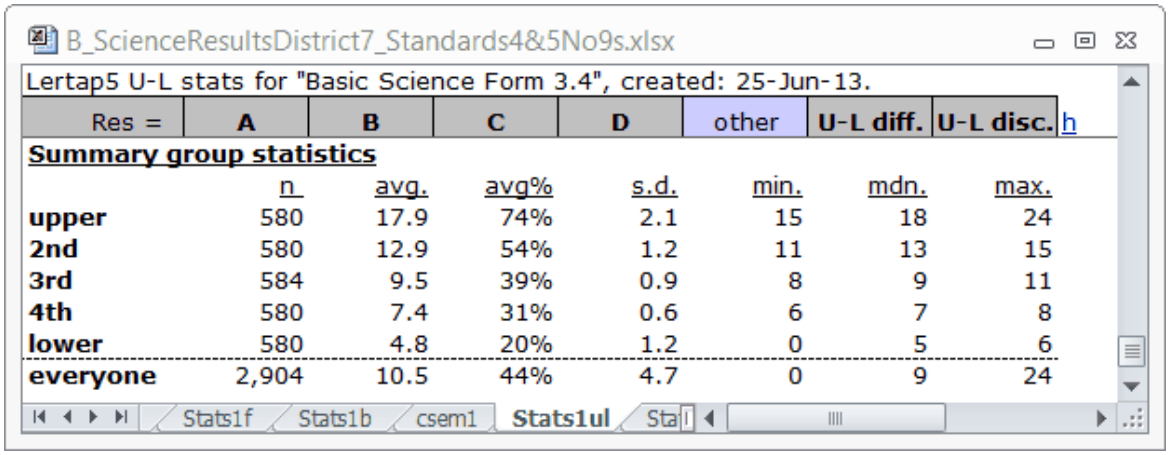
	1	2	3
1	Lertap5 Scores worksheet, last updated on:		
2	**ID Code**	**BSci3.4**	**MWSfix1**
2907	n	2,904	2,904
2908	Min	0.00	0.00
2909	Median	9.00	10.00
2910	Mean	10.48	10.93
2911	Max	24.00	24.00
2912	s.d.	4.73	4.86
2913	var.	22.37	23.64
2914	Range	24.00	24.00
2915	IQRange	7.00	7.00
2916	Skewness	0.54	0.55
2917	Kurtosis	-0.49	-0.53
2918	MinPos	0.00	0.00
2919	MaxPos	24.00	24.00
2920	**Correlations**		
2921	BSci3.4	1.00	0.99
2922	MWSfix1	0.99	1.00
2923	*average*	*0.99*	*0.99*

After double-keying I8 and I15, the Median test score increases by one point; it's now 10.00. The Mean (or average) test score goes up too, from 10.48 to 10.93.





Look carefully, clean your glasses, and you'll spot some differences in the two score histograms. The top one is "before"; the bottom one is "after". There are more high scores now; compare the height of the bars starting at about a score of 13.



B_ScienceResultsDistrict7_Standards4&5No9s.xlsx

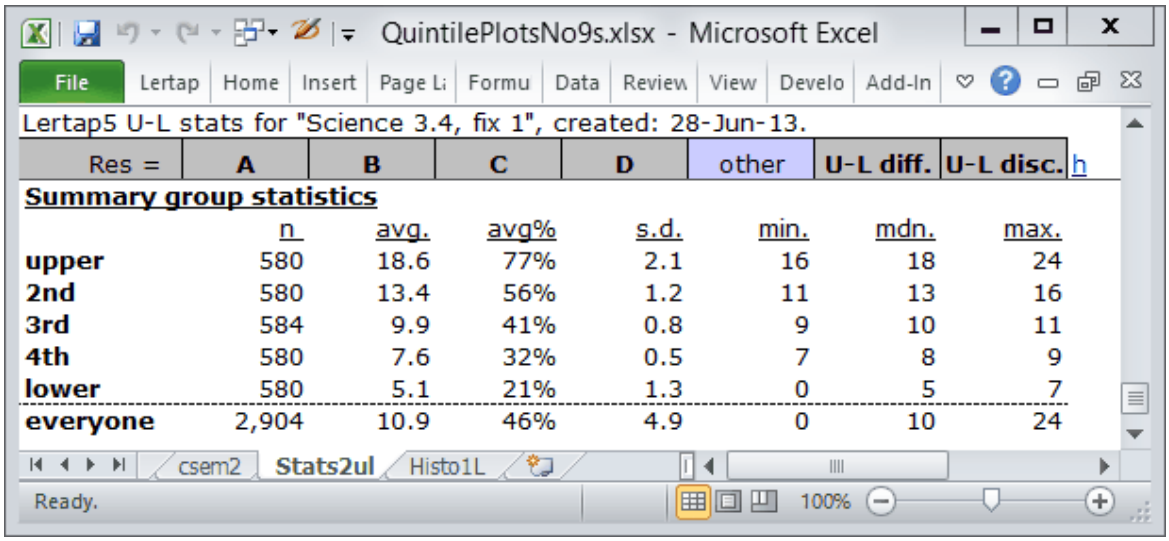
Lertap5 U-L stats for "Basic Science Form 3.4", created: 25-Jun-13.

Res =	A	B	C	D	other	U-L diff.	U-L disc.	h

Summary group statistics

	n	avg.	avg%	s.d.	min.	mdn.	max.
upper	580	17.9	74%	2.1	15	18	24
2nd	580	12.9	54%	1.2	11	13	15
3rd	584	9.5	39%	0.9	8	9	11
4th	580	7.4	31%	0.6	6	7	8
lower	580	4.8	20%	1.2	0	5	6
everyone	2,904	10.5	44%	4.7	0	9	24

Stats1f | Stats1b | csem1 | **Stats1ul** | Sta



QuintilePlotsNo9s.xlsx - Microsoft Excel

File | Lertap | Home | Insert | Page Li | Formu | Data | Review | View | Develo | Add-In

Lertap5 U-L stats for "Science 3.4, fix 1", created: 28-Jun-13.

Res =	A	B	C	D	other	U-L diff.	U-L disc.	h

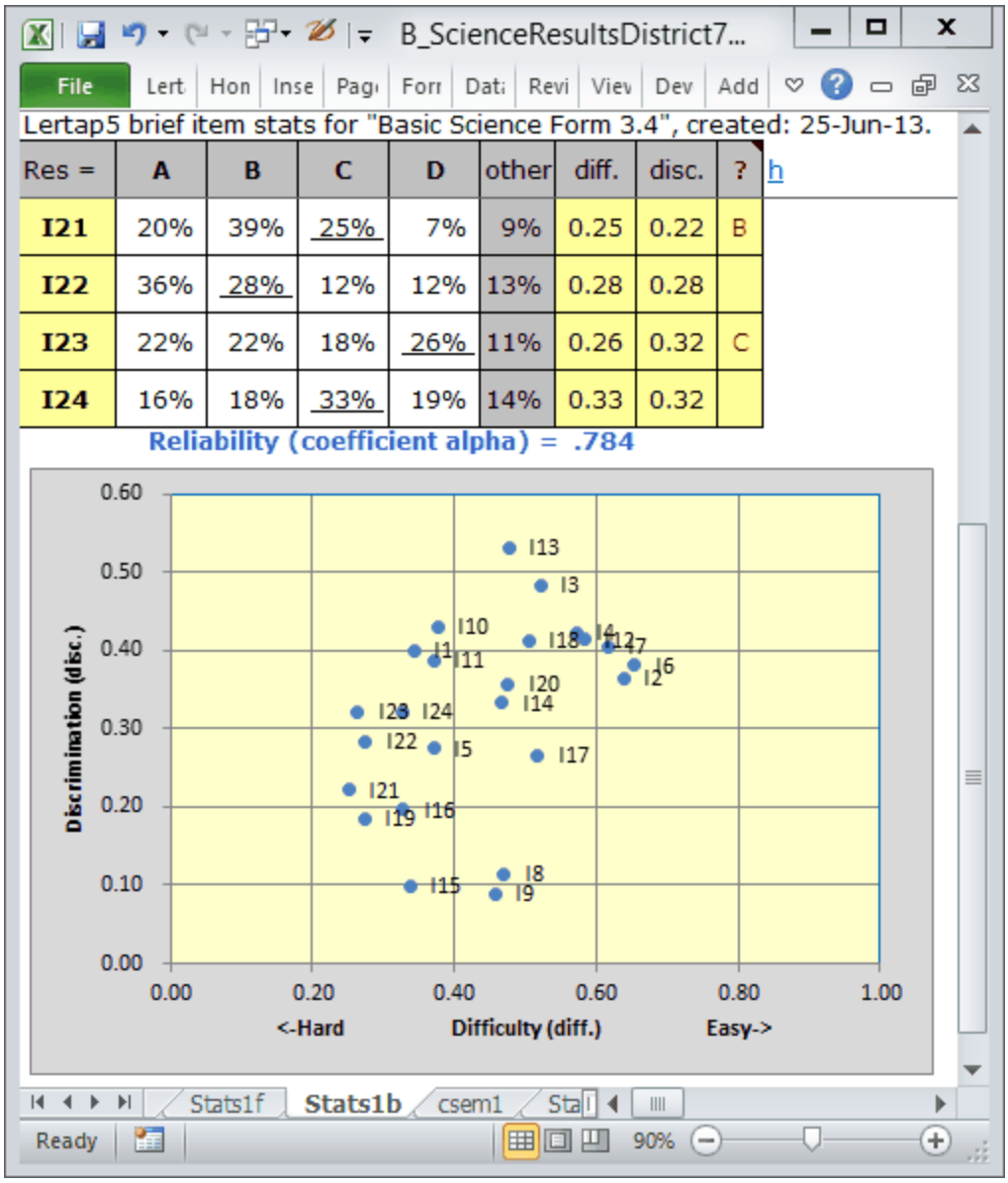
Summary group statistics

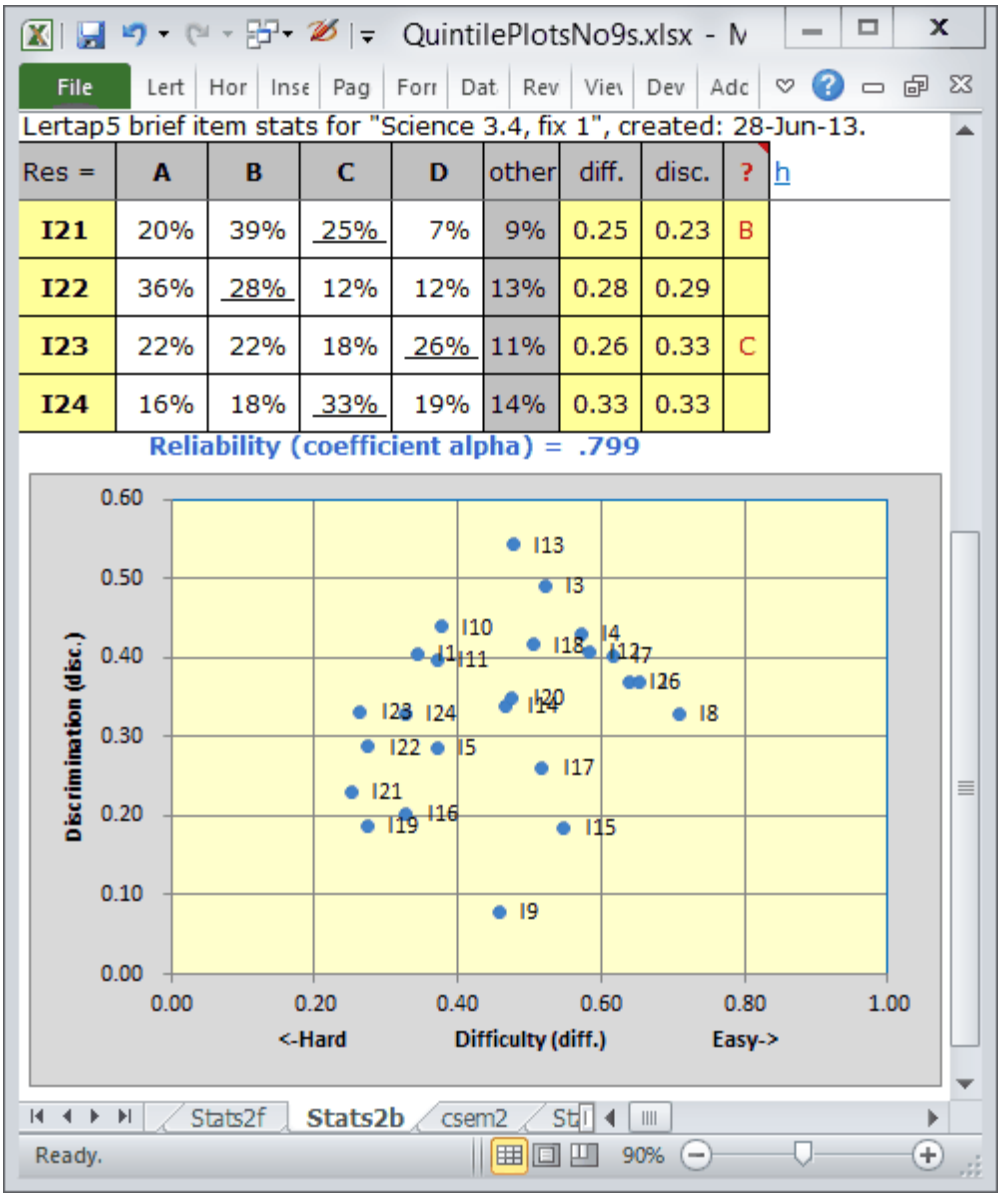
	n	avg.	avg%	s.d.	min.	mdn.	max.
upper	580	18.6	77%	2.1	16	18	24
2nd	580	13.4	56%	1.2	11	13	16
3rd	584	9.9	41%	0.8	9	10	11
4th	580	7.6	32%	0.5	7	8	9
lower	580	5.1	21%	1.3	0	5	7
everyone	2,904	10.9	46%	4.9	0	10	24

csem2 | **Stats2ul** | Histo1L

Ready. 100%

The "Summary group statistics" (above) have changed in an expected manner. Comparing the two "avg." columns indicates that all five groups have benefited from the new scoring.





Here (above) there are some interesting changes. (The top scatterplot is "before"; the bottom is "after".)

The reliability has increased, moving from .784 to .799. Note that the two items which have been double-keyed have shifted in the scatterplots. I8 and I15 were at the bottom before, next to I9. Now they've improved, with I8 becoming particularly attractive with its discrimination now above the 0.30 line.

I8 (c12)

option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	1.00	1,366	0.47	0.11	0.14	11.57	0.23	
B	0.00	493	0.17	-0.24	-0.36	7.92	-0.54	
C	0.00	220	0.08	-0.23	-0.42	6.71	-0.80	
D	0.00	689	0.24	0.16	0.22	11.85	0.29	<-aa
other	0.00	136	0.05	-0.12	-0.25	8.01	-0.52	

I8 (c12)

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	1.00	1,366	0.47	0.07	0.09	11.79	0.18
B	0.00	493	0.17	-0.26	-0.39	8.11	-0.58
C	0.00	220	0.08	-0.23	-0.43	6.95	-0.82
D	1.00	689	0.24	0.15	0.21	13.07	0.44
other	0.00	136	0.05	-0.13	-0.27	8.18	-0.57

The bottom table above shows I8 after double-keying. Note how the "avg." and the "z" values have dramatically increased for the 689 students who selected option D? I15's changes show a similar pattern, but are not as dramatic:

I15 (c19)

option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	0.00	945	0.33	-0.12	-0.16	9.64	-0.18	
B	1.00	982	0.34	0.10	0.13	11.79	0.28	
C	0.00	610	0.21	0.07	0.10	11.13	0.14	<-aa
D	0.00	262	0.09	-0.15	-0.26	8.24	-0.47	
other	0.00	105	0.04	-0.11	-0.27	7.71	-0.59	

I15 (c19)

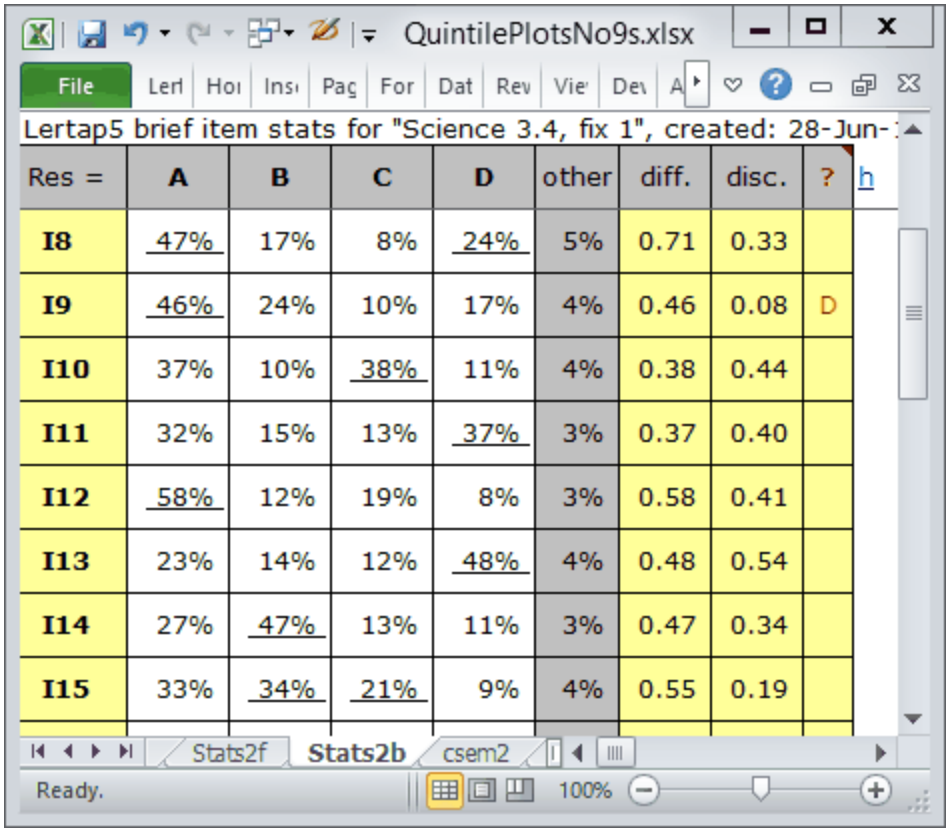
option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	945	0.33	-0.16	-0.20	9.83	-0.23
B	1.00	982	0.34	0.06	0.08	12.06	0.23
C	1.00	610	0.21	0.05	0.07	12.38	0.30
D	0.00	262	0.09	-0.16	-0.28	8.52	-0.50
other	0.00	105	0.04	-0.12	-0.29	7.84	-0.64

As a teacher I usually love it when students ask questions, providing I know the answer. So it is that I am now pleased when Andrés Ricardo, sitting in the back as always, comes forth with a really good question:

> "*Why have the scatterplots shown I8 moving to a discrimination above 0.30 when its pb(r) values are nowhere near that?*"

Well, the pb(r) values are correct, but, when an item is multiply-keyed they're not the appropriate measure of item discrimination. We'd want the correlation of the item with the criterion score, not the correlations of the options with the criterion. Such correlations are given in the Statsb reports, under its "**disc.**" column.

No-one asked, but I was ready to make a similar comment about item difficulty. When an item has just one right answer, the corresponding p value in Statsf will equal "**diff.**" in Statsb. Otherwise, when items have more than one keyed answer, their difficulty index is shown in Statsb under the "diff." column. (For more about item difficulty calculations, click here.)



Res =	A	B	C	D	other	diff.	disc.	?
I8	47%	17%	8%	24%	5%	0.71	0.33	
I9	46%	24%	10%	17%	4%	0.46	0.08	D
I10	37%	10%	38%	11%	4%	0.38	0.44	
I11	32%	15%	13%	37%	3%	0.37	0.40	
I12	58%	12%	19%	8%	3%	0.58	0.41	
I13	23%	14%	12%	48%	4%	0.48	0.54	
I14	27%	47%	13%	11%	3%	0.47	0.34	
I15	33%	34%	21%	9%	4%	0.55	0.19	

In this table (above) we see that I8 has a discrimination ("disc.") of 0.33. I15's disc. value is 0.19.

"*Why does I9 have a D in the ? column. What is the ? column?*"

Flags! I have forgotten to mention flags. Look here, please:

I9 (c13)

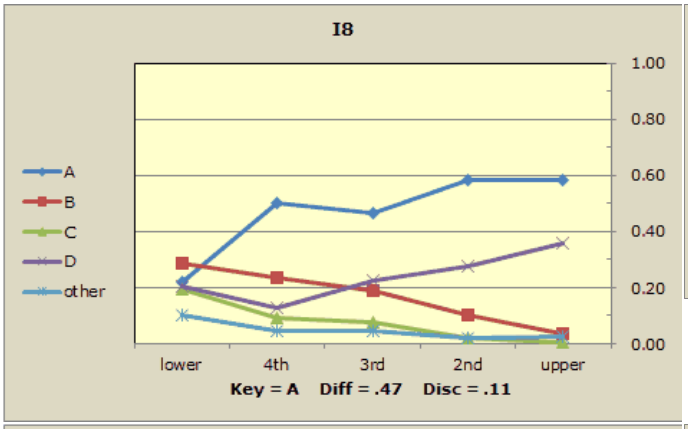
option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	1.00	1,331	0.46	0.08	0.10	11.88	0.20	
B	0.00	685	0.24	-0.10	-0.13	10.09	-0.17	
C	0.00	279	0.10	-0.10	-0.18	9.42	-0.31	
D	0.00	482	0.17	0.00	0.01	10.97	0.01	<-aa
other	0.00	127	0.04	-0.10	-0.22	8.65	-0.47	

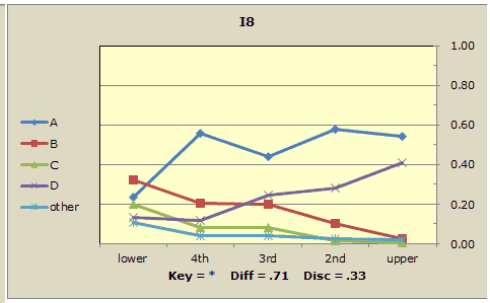
See the "<-aa" indicator, with a little red triangle? It's a flag, something to draw attention to an item option which may have a problem. In the Statsb reports, the "flags" appear in the ? column, but they're not triangles.

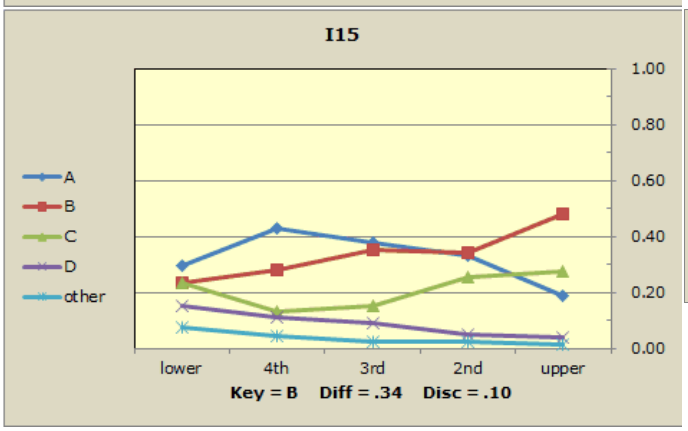
"aa" means "above-average"; the "avg." score of the 482 students who chose option D on I9 was above the overall test score average. Lertap has had its flag waver out to signal this. Why? *Oh come on.* You gotta know by now: an item meant to discriminate should have what? One option with a high "avg." value, just one, and only one "z" with a positive value. So out comes the flagger on I9, saying "*look here, possible rule violation, potential penalty, for I9*". (Read more about flags here.)

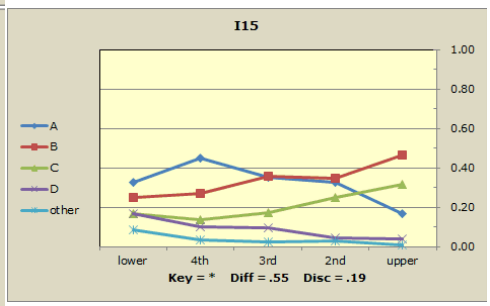
I have a question for you now. How do you think the quintile plots will have changed for I8 and I15?

Get your hopes down. We won't notice a great deal. Here are before (on the left) and after (on the right) pictures:









There is a need to combine the two right answers for these items so that we end up with a single line combining the proportions in each group. Imagine I did this for the upper group in I8. The correct answers are A and D. It looks like A's proportion is about 0.55 in the upper group, with D's about 0.40. Combining these gives 0.95 for

what would be our new trace line for the correct answers combined. This would then make a nice quintile: the trace line for the right answer would flair up on the right, almost reaching 1.00 (or 100%); the other lines would not change.

The effect of combining the two right answers for I15 would not be quite as impressive. It looks like for the upper group I'd be combining about 0.45 (for option B), and maybe 0.31 (for option C), giving 0.76. Nice, an improvement, but not as compelling as the change for I8. (The big problem for I15 is that distractor A has not fallen to zero, or near zero, in the upper group. Distractor D is near zero, but still hanging in a bit.)

What then may we conclude about the effects of double-keying two items, I8 and I15? Well, the real effect has gone unmentioned, hasn't it? We have made our test scoring fairer. After acknowledging ambiguities in these two items, and double-keying them, we've changed the marks for the 689 students who took option D on I8, giving them an extra point. In a similar manner, the 610 students who selected option C on I15 also now have an extra point.

Summary. In this topic I've discussed how item and test performance have changed after two items were double-keyed. The discrimination value for both items has gone up, as has test reliability. But it certainly bears repeating, let me say it once more: these items need to be fixed before they are used again.

5 Wrapping up

Extra credit to you for making it all the way through this document (!). I hope it has been useful.

You will almost certainly know that Lertap is not the only tool, or "app" as we tend to say these days, capable of supporting work of the sort I've demonstrated herein.

When I started out as part of a data collection and processing center at the University of Wisconsin in 1967, we used tools known as "Statjob" and "BMD" for data snooping and cleaning, and Frank Baker's "FORTAP" program for item analysis. If these systems didn't do exactly what we wanted, we at times wrote our own special-purpose routines. When I was fortunate to have the chance to work in Venezuela in the early 1970s, the team I was part of wrote a special suite of data analysis programs in order to have output in Spanish (Lertap grew out of some of that work).

Fast forward to the year 2013. **SPSS** (www.spss.com) is but one example of a "modern-day" system which could do quite a bit of the work I've presented in this document. SPSS is capable of reading Excel worksheets, making it easy to take advantage of Excel's very powerful data snooping support before delving into the extensive data analysis tools offered by SPSS. I've put a document up here with a practical demonstration of moving data from Excel to SPSS.

I have always regarded SPSS as weak when it comes to analyzing the performance of cognitive test items. But it seems I may be wrong -- I trotted out my favorite search "engine", gave it "using SPSS for item analysis", and got some very interesting hits, including a quality "white paper" on item analysis from SPSS itself.

Intrigued by the results of my SPSS item analysis search, I also tried "using Excel for item analysis" and once again got a list of relevant sites. Some of these looked very good (especially, of course, those which referred to the use of Lertap, but there were *many* others).

Searching the internet can snowball quickly. After asking about SPSS and Excel as they might relate to items analysis, I then gave "software for item analysis" to the search tool. There were a few million hits, more than enough to delay my lunch hour.

So there you go. Can you do the things I've wowed you with using software other than Excel and Lertap? Certainly. But allow me to enter a caveat, if I may: don't trust datasets until they've been given a real snoop. It's very easy to take the output from an optical scanner and deliver it immediately to systems which will rapidly produce all manner of statistics. We can get pages of tables and graphs without ever really seeing our data, leaving us very susceptible to the old "GIGO" effect: "garbage in, garbage out". Snoop the data first.

Of course I'm not saying that scanners introduce garbage; it's simply that it can be difficult to see the data after they've been prepared / processed by a scanner. We very often end up with "ASCII" text files, and they do not lend themselves to snooping.

(Fortunately, most scanners can, I believe, be configured to create Excel-compatible files, such as "CSV" files. Nonetheless, there remain numerous programs which have been designed at the outset to work with ASCII files -- I urge caution with such programs, not because the programs themselves are deficient, but because ASCII files (generally having extensions of "TXT" or "DAT") are, as I've said, difficult to snoop. Use of what could be called ASCII-friendly data analysis programs makes it more likely that people will not snoop their data, leading to what can well be mistaken faith in data integrity, possibly increasing the chance of "GIGO". Note: Excel is capable of importing ASCII files; it takes a bit of work, but the process is entirely straightforward. You might read about it here, where ASCII files are referred to as "text" files.)

While data snooping has been a central theme of this paper, there is another, one which hasn't received quite as much air time: adjusting test scores for the effects of item ambiguity. A quality item analysis program ought to have the ability to accept more than one answer as "correct", or, perhaps, partially correct. In Lertap this is done via the use of *mws cards; in case you've already forgotten, click here to jet back to the "double-keying" topic. When I say "partially correct", I mean, for example, perhaps giving half a point to an answer. Examples of how to do this may

be seen here, starting at "Example C12". (In the literature, scoring item responses in this manner sometimes falls under the rubric of "partial credit".)

Larry Nelson (snooperman?)
School of Education
Curtin University
Western Australia

larry@lertap.com

Index

- * -

- ? -

- 3 -

- A -

- B -

- C -

- D -

- E -

- F -

- G -

- H -

- I -

- K -

- L -

- M -

- N -

- O -

- P -

- Q -

- R -

- S -

- T -

- U -

- V -

- W -

- Z -