Using Practical Exhibits to Present Selected Measurement Topics with Lertap 5, Excel, and R Packages

Larry R Nelson

Dated: 5 December 2022

I have initiated work on developing a fresh set of practical exhibits, ones which differ very substantially in both scope and content to my class notes from back in 2018. This new work presents entirely self-contained exhibits, designed for self-study without need for an instructor.

The first of these, "Practical Exhibit Test 13", has to do with fundamentals of classical test theory (CTT). It uses Lertap5 only. The actual test items are presented. Practical Exhibit BDI-21 processes a popular scale (a depression inventory); it also has the actual items.

Note: the text below dates back to 8 October 2018. I still use some of the material below but, as will be noticed, by and large the content below are personal notes from me to me, reminding me of issues to address as classes are presented to students.

- **Exhibit 1**: The Lertap Quiz (Laboratory of Educational Research Test Analysis Package)

Background. It's assumed that the audience is not well-versed on the terms used in item and test analysis; this is introductory content getting into important basic terms, and introducing Lertap 5. It might be called a "gateway" exhibit. It's based on "CTT", classical test theory.

References. The Lertap 5 user manual will be an adequate reference for this exhibit. The Lertap Quiz itself will be the basis for the discussion; here's the link.

Suggested reading. Chapter 2 of the Lertap 5 manual.

Resources required: Lertap5.xlsm (the main Lertap 5 workbook). Handout mentioned below.

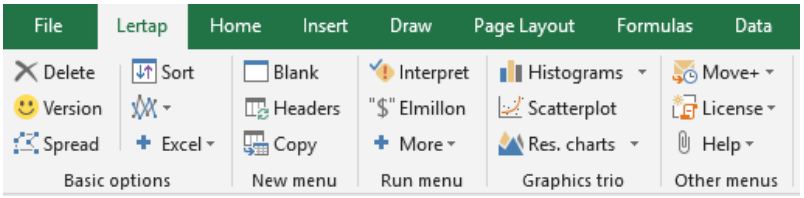
Description. The quiz has 25 multiple-choice cognitive items, 10 Likert-style affective items, and two numeric short-answer "supply" items. Data were collected from 60 respondents. The instrument may be seen here. Respondents circled and wrote their responses directly on the "test paper". Responses were then processed using a keypunch machine and a verifier. Lertap 2, running on a Burroughs B6700 mainframe computer, was used to produce printed results on a 132-column line printer.

Demonstration

> Handout. A copy of the instrument, or perhaps a partial copy with just some selected items.
>
> Things to note and mention. The cognitive items use letters as "response codes". The number of options for the items differ (*this is not typical*). The affective items use digits as their response codes. Some of the affective items are "positive statements", whereas others are negative. Likert-style items are super common.
>
> *Discuss how these items will be scored. This will probably involve reversing some of the affective items. Make the distinction between item response and item score.*
>
> *Introduce these terms: distractors, dichotomous, binary, polytomous , partial-credit*
>
> *What's to be done if a person doesn't answer an item?*

Setting: the instrument was used at the end of a workshop for secondary and tertiary teachers held in Dunedin, New Zealand. Workshop objective: how to process test and survey results using Lertap 2 on a mainframe computer. The cognitive questions probe understanding of how to use Lertap 2; the affective questions gather feedback on opinions about the workshop and Lertap itself; the two short-answer questions regard respondent backgrounds.

How to get a copy of the data: (1) download a copy of the Lertap 5 system for Windows and Macintosh computers from here. The Data worksheet in the Lertap5.xlsm workbook has item responses; the CCs worksheet has the syntax lines set up to create two "subtests", one cognitive and one affective. Subtest short titles are, respectively, "Knwldge" and "Comfort". (2) create a copy of the Data and CCs sheets by using the "Copy" option on the Lertap 5 tab, shown below:



The Copy option will get Excel to create a new workbook having Data and CCs worksheets; the name of the workbook will be automatically assigned by Excel; it may be "Book2" or, more generally, "BookX", where X will be an integer usually no greater than 9.

This new workbook will be a typical, official, Lertap 5 workbook (see definition).

Lertap workbooks are not usually set up in this manner. There are a variety of ways more commonly used to prepare Excel workbooks so that they meet the requirements of a true "Lertap workbook". Refer to "Lelp", the online help system.

NOTE: consider having the class follow along by doing these steps on their computers.

Look at the Data and CCs worksheets

There's much to discuss. This topic may be useful at this point.

The Data worksheet has "Record" in the first column, then the item responses from the students, then "YrsComp", then "YrsTest".

Describe YrsComp and YrsTest.

In the CCs sheet, subtest information starts with a *col line. This gives the location of the item responses in the Data worksheet. When there is a second subtest (as in this example), the second *col line specifies where its item responses are located.

In this case, the cognitive subtest is unusual as its items use a different number of response options. The *alt line has been used to indicate the last response code used by each item. For example, the first item, found in column 3 of the Data sheet (c3), uses just {A,B,C}; the second item uses {A,B,C,D,E} and is found in column 4 of the Data sheet.

*alt lines not usually needed (see example C7 at this page)

A *pol line has been used to indicate how the affective items are to be scored.

Wt=0 declarations have been used so the Lertap will not add the two subtest scores together to create a total score.

Get "Freqs"

Use the Interpret option.

What does Freqs tell us? Are the responses what we expected, they're aren't any weird characters found, such as a small B where there should have been a large one?

Notice the little blue h? (It leads to help.)

We started out with just two worksheets, Data and CCs, now we have three.

Run "Elmillon" (using the Elmillon option, of course)

Now we have seven new worksheets: Scores, Stats1f, Stats1b, Stats1ul, csem1, Stats2f, Stats2b.

Let's look at Scores

Get Histograms

The z-scores are very important. Talk about the normal curve, and the -3 to +3 range commonly used. Z-scores outside this range are sometimes called outliers.

Get a Scatterplot

Maybe get a regression line to display, with the value of R^2.

There are so many other worksheets to look at. What was it I'd like to know? Maybe I don't need to look at each and every of these worksheets! Here are my main questions at the moment:

Did the students reach a "mastery" level on the cognitive items, or, are there areas where I need to direct more teaching? I can use just Stats1b for this, but note: there's a little blue h on every sheet – I can find out more by using it no matter what sheet I'm looking at.

Talk about item difficulty. Again mention distractors.

What do their affective responses tell me? I can use Stats2b to get answers. Maybe talk about the cor. values, they're similar to the Discrimination values found on Stats1b.

Use Res. charts with Stats2b to show off this capability. Discuss. Is it easier to detect trends using graphics?

Comments on Exhibit 1

This has been a straightforward demonstration without needing to consider the statistics found on Stats1f, Stats1ul, csem1, and Stats2f. Lertap produces more information than is often needed; I'm free to pick and choose as I wish.

(At this point, I have elected to leave a discussion of reliability and standard error of measurement for another exhibit.)

It would be possible to dig deeper. For example, we could investigate the relationship between subtest scores and the last two items on the instrument, the ones asking respondents to indicate the number of years they'd been using this and that. To do this, it would be necessary to have first run Elmillon so that the Scores worksheet is established. Then the two experience variables can be copied to Scores using this option. The correlation matrix at the base of Scores will be updated automatically.

- **Exhibit 2**: MathsQuiz

Background. It's assumed that the audience has been through a discussion of "the basics", as presented in Exhibit 1. The main purpose of this second exhibit is to elaborate on item "discrimination". Quintile plots will feature prominently (a gateway to IRT's ICC plots to come in later exhibits). There's not all that much discussion here – this is a "concise" exhibit with a limited topic focus.

Suggested reading. Chapter 7 of the Lertap 5 manual.

References. The Lertap 5 user manual will be a basic reference for this exhibit; particularly relevant is Chapter 7, where the Stats1ul worksheet is discussed. Some webpages will be useful too: an updated Stats1ul discussion is here; quintile plots are here; "binary plots" are here. The MathsQuiz sample dataset website gives access to valuable resources, including a *useful* discussion of the Stats1ul "report", and quintile plots.

Resources required: Lertap5.xlsm (the main Lertap 5 workbook). The "BinaryItems" macro should be available via the Macs Menu. The dataset itself, an Excel workbook, may be reached at this link. A version of the dataset with binary item scores in its Data worksheet is on one of my laptops, "Lensito", where it is located under SampleDatasets / MathsQuiz / RMCS-2018, saved as "MathsQuizRMCS-Binaries.xlsx" (I used the steps in this webpage to create this xlsx workbook).

Description. This dataset involves 15 cognitive items, all multiple-choice. Data were collected from 999 respondents.

A sample (n=100) of this dataset is included when Lertap51099.zip is downloaded; it comes as "MQuiz100.xlsx" once the zip's files have been unzipped (extracted). The ReadMeMacWin.pdf might be useful here as it references MQuiz100.xlsx, and has a suggested exercise based on it.

Demonstration

Looking at "MathsQuizRMCS-Binaries.xlsx", see binary item scores in Data; res=(0,1) in CCs; binary "responses" in Freqs; look at Stats1ul in detail – see group statistics at the bottom, then scroll up to results for the first item, and discuss.

Then show "PackedPlots" and discuss. There are some good examples of trace lines in the plots. Mention how "discrimination" is related to the slope of the trace line; the maximum value for "Disc." is 1.00. Item `I11` has negative slope due to being mis-keyed.

- **Exhibit 3**: FIMS

Description. Has 14 cognitive items used in the FIMS study. Mixture of supply and multiple-choice items. Respondents were junior high-school students from Japan (n = 2,051) and Australia (n = 4,320). Dataset includes two demographic variables, Country and Gender.

Background. The utility of this study has to do with item and test development using a matrix of topic objectives, resulting in the creation of an item pool. It also lends itself to a discussion of result breakouts by Country and by Gender, allowing a discussion of "DIF", differential item functioning. Will also be used to discuss "dimensionality" – was there perhaps a "reading" factor as well as the expected maths ability factor?

References. The FIMS sample dataset website.

Resources required: Lertap5.xlsm (the main Lertap 5 workbook; must have the "experimental features" option on, row 18 in the System worksheet). The workbooks to use are in SampleDataSets / FIMS / FIMs-RMCS2018 / FIMS-Original.xlsx, FIMS-MultipleSubs.xlsx, and SampleDataSets / FIMS / GenderDIF.xlsx. Handout: a printout of the actual FIMs items.

Note: delete most of the files in the FIMs-RMCS2018 folder as the Omega1 macro will be used and it will want to write and download new files to this folder.

Computers need to have R and RStudio installed[1], along with these R packages: "psych", "GPA-rotation", "ltm", "TAM", "WrightMap". Students may download the special R script made for installing packages from a selected CRAN site (such as Curtin University in Australia).

Demonstration

Review the actual questions (on the handout).

Use FIMs-Original.xlsx[2]. Make a copy. Use the "Interpret" and "Elmillon" options.

Were there country differences? Use the "Breakout scores by groups" option. Get a boxplot. Get histograms. Yes, the students from Japan did better. Next, use the "Item responses by groups" option to show where the differences were, item by item.

Were there gender differences? Mention that we'd take the same options, "Breakout scores by groups", and "Item responses by groups", and now elect to look for DIF too – see the GenderDIF.xlsx workbook where results are already available. Note that there were no gender differences when using the total score, but there were some on the items – see the IbreaksMH1 worksheet where two items have been graphed, one with DIF, one without. Mention that DIF methods have been compared by RMCS doctoral students.

Were the supply items more difficult than the multiple-choice items? The supply items are Q1, Q5, Q10, and Q11. Use scatterplot at the base of Stats1b.

Use FIMS-MultipleSubs.xlsx. Look at Scores, there are "All", "Supply", and "MC". Stats2f reports that the Supply average was 43.8%, while Stats3f shows that the MC average was higher, 50.9%

Was there possibly a "reading factor" here? Do students need two abilities to answer the Supply items: reading and mathematics?

When we want to know if students may need, or may use, more than one ability to answer test items, we are asking about the dimensionality of the test.

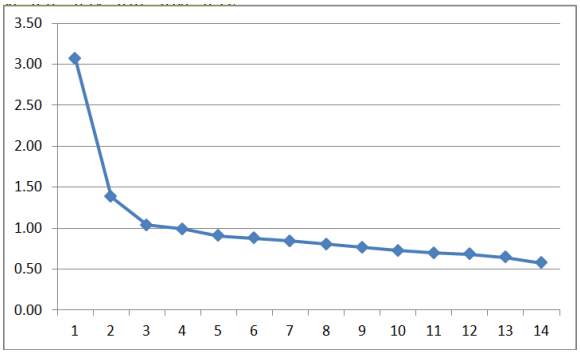
There are numerous methods used to determine the dimensionality of a test or survey.

[1] See this document and its steps for getting R and RStudio ready to use with Lertap.
[2] This is identical to AUS_JPN_2.xlsx as obtained from the website.

One of the most common methods is to find the "eigenvalues" of the inter-item correlation matrix. In Lertap, this is done by using the "Item scores and correlation" option. I will demonstrate the use of this option.

The option adds another worksheet to the workbook, called "IStats". This worksheet has two sections: the first section has the item scores for each student, the second section has the inter-item correlations and associated statistics, including eigenvalues. Wikipedia has a discussion of eigenvalues and eigenvectors – it's here.

We can make a "scree test" from the eigenvalues – watch as I do so – here it is:



Scree tests are very popular (here is a sample reference). The one above indicates that there seems to be at least two factors, or dimensions, to the FIMS data collected from Japanese and Australian students.

What the scree plot is not too good at is suggesting what the factors may be. We want to know if there may be a "reading" factor, and a "mathematics" factor.

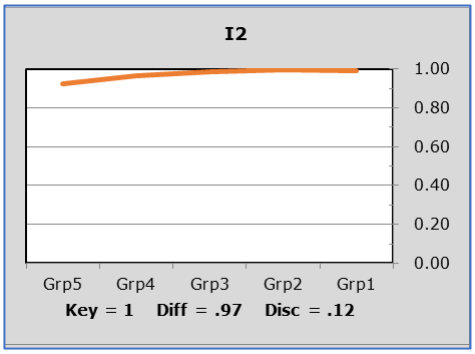
End of Class 1.

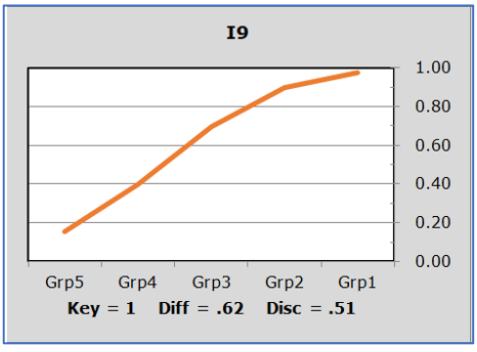
- **Start Class 2**

Last week we looked at three "tests": the Lertap Quiz, the MathsQuiz, and the FIMS test.

I talked about item difficulty and item discrimination, using CTT (classical test theory) terminology, where item difficulty = the proportion of students who got the item right, and item discrimination is related to the slope of the response trace line for the item "key" (the correct answer).

In CTT, difficulty may have a minimum of 0.00, maximum of 1.00. Easy items have high difficulty values (this is not what we would expect; it's a small 'problem' in CTT). Discrimination should not be below 0.00, and cannot be greater than 1.00.

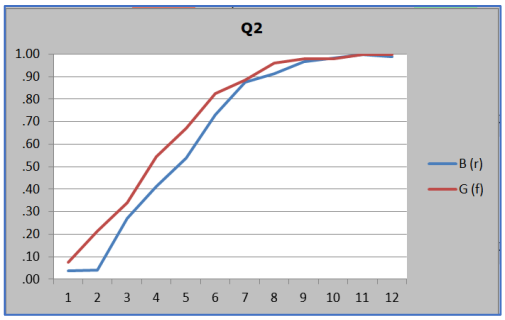


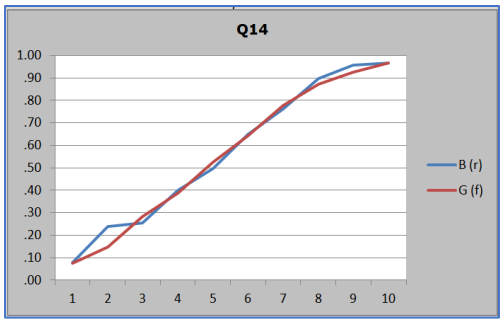


The plots above indicate the proportion of correct responses to two MathsQuiz items, I2 and I9, using five groups of students: Grp5 has the weakest students, Grp1 has the strongest students. I2 was very easy and showed little discrimination. I9's performance was much better if we want to have a discriminating test item, one that lets us identify the best and the worst (weakest) students.

A recommended reference for these statistics is Chapter 7 of the Lertap manual, and also this webpage.

I used FIMS to talk about "DIF", differential item functioning. Lertap will make a DIF plot, and I showed an example of two items, one with DIF, and one without:



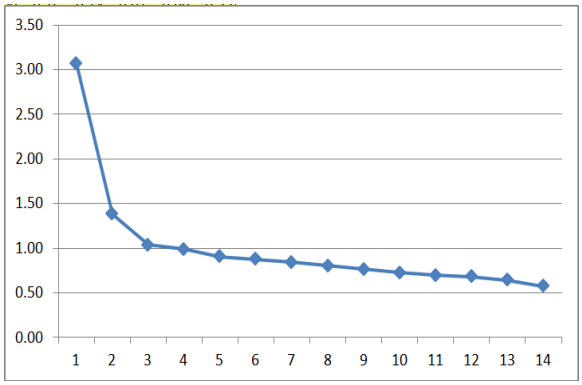


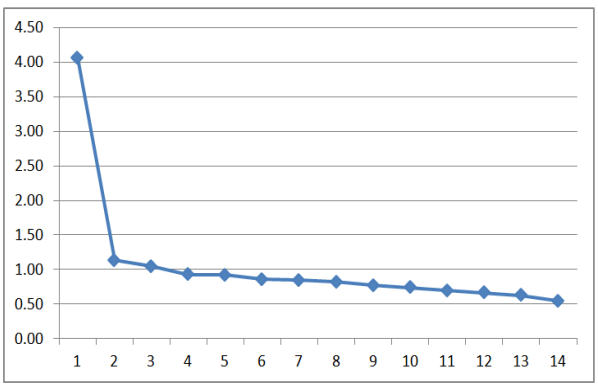
I mentioned that "DIF" has been a popular research topic among RMCS students. This is because there are now numerous ways to determine if an item has DIF. Lertap's DIF method is one of the oldest – read about it here.

I think an interesting research topic would be to see how often DIF studies have included graphical results of DIF. My impression is that some of the DIF methods are probably based only on a statistical test of some sort, and may not involve a graph – this would be unfortunate, in my opinion, because if there is DIF then a graph should be used as well as the statistics found in DIF analyses. *("A picture is worth a thousand words.")*

I also used FIMS to introduce factor analysis. This was because of my experience in Venezuela where we found that a test of mathematics was, for some students, also a test of reading ability.

If this is true for the FIMS results, then a factor analysis should show it. I used Lertap results to make a "scree plot" for FIMS, and also for MathsQuiz[3]:





[3] The eigenvalues plotted are from Pearson product-moment correlations (not tetrachorics).

Above, the plot on the left is FIMS; on the right is MathsQuiz. The FIMS scree plot does suggest that there may be two factors; this is because the "scree" appears to start at the third eigenvalue.

The "scree" for MathsTest appears to start at the second eigenvalue.

Scree plots are, after many years, still a very popular way to see if there might be more than one factor behind test results. When there is only a single factor, a scree plot will look more like the one made from MathsQuiz results (the plot on the right).

There are some problems with scree plots. The main problem is one of interpretation – when I look at a scree plot and say "I see two factors", someone else may say "I see only one factor". In other words, interpreting scree plots is a "subjective process".

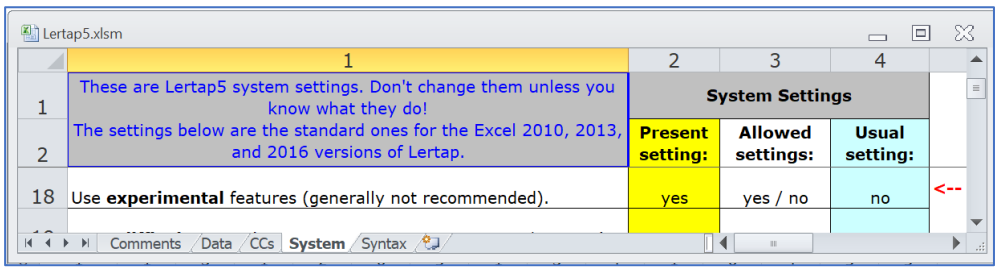
When there may be more than one factor indicated in a scree plot, the next problem will be finding out what the factors are.

I will use one of the special macros in Lertap, "Omega1", to get a complete factor analysis, one which uses some of the latest statistical methods seen in the literature.

If there is a reading factor in the FIMS results, then I should expect to see those FIMS items that require students to read some text will group together to form a factor. (We will see that this did not happen. But the results are interesting – Q7 and Q8, for example, seemed to be unrelated to the other questions – why may this have happened?)

Demonstration

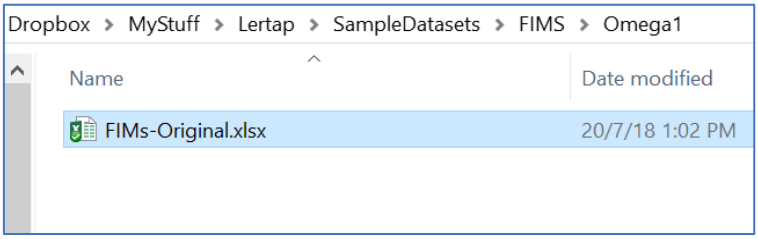
I will show how to use the Omega1 macro with FIMS data. Before I do, I will modify the settings in my Lertap5.xlsx workbook so that it will use "experimental features"[4]. The way to do this is to change row 18 in the System Worksheet, putting "yes" in column 2:



Once I have made this change, and saved Lertap5.xlsm, Lertap will add IRT statistics when it makes its "Stats1b" report. We will see these after I have run the "Interpret" and "Elmillon" options.

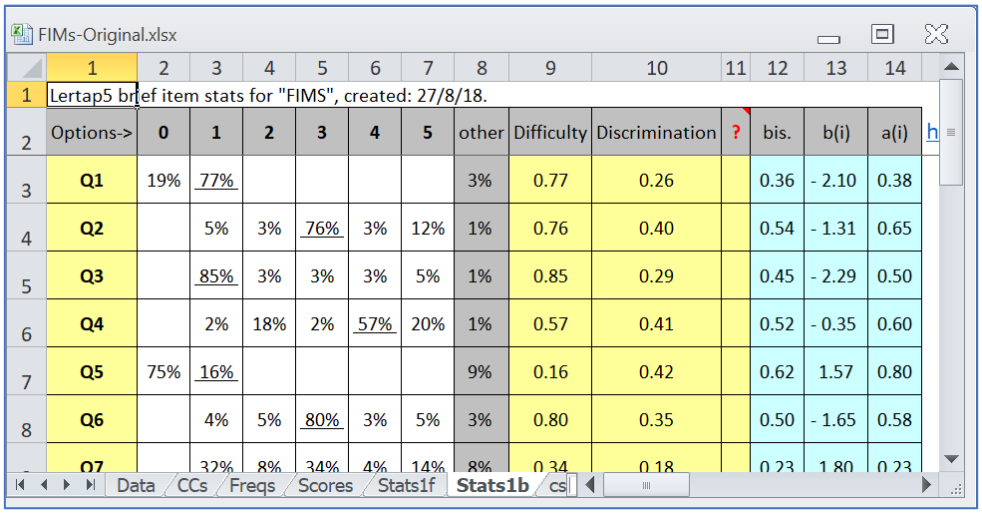
[4] It is **not** really necessary to make this change. I have done it only as a demonstration. It is **not** important.

I will be working with the "FIMs-Original.xlsx" workbook found in this folder on my laptop (notice that there is just one file in the folder as I start):



Okay, ready! I open Excel. I open Lertap5.xlsm and then FIMs-Original.xlsx. I use the Interpret and Elmillon options.

Look at Stats1b now, it has three blue columns with some new statistics[5]:


FIMs-Original.xlsx

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Lertap5 brief item stats for "FIMS", created: 27/8/18.													
2	Options->	0	1	2	3	4	5	other	Difficulty	Discrimination	?	bis.	b(i)	a(i)
3	Q1	19%	77%					3%	0.77	0.26		0.36	- 2.10	0.38
4	Q2		5%	3%	76%	3%	12%	1%	0.76	0.40		0.54	- 1.31	0.65
5	Q3		85%	3%	3%	3%	5%	1%	0.85	0.29		0.45	- 2.29	0.50
6	Q4		2%	18%	2%	57%	20%	1%	0.57	0.41		0.52	- 0.35	0.60
7	Q5	75%	16%					9%	0.16	0.42		0.62	1.57	0.80
8	Q6		4%	5%	80%	3%	5%	3%	0.80	0.35		0.50	- 1.65	0.58
	Q7		32%	8%	34%	4%	14%	8%	0.34	0.18		0.23	1.80	0.23

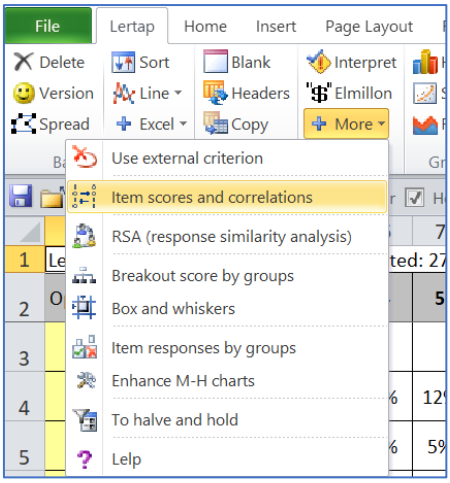
Data | CCs | Freqs | Scores | Stats1f | **Stats1b** | cs

The b(i) and a(i) statistics are the IRT b and a figures for item (i). They are mentioned in this webpage. (Remember that Professor Baker's book is our recommended reading for IRT. That book describes IRT parameters and models, and is a must read if you want to understand IRT.)

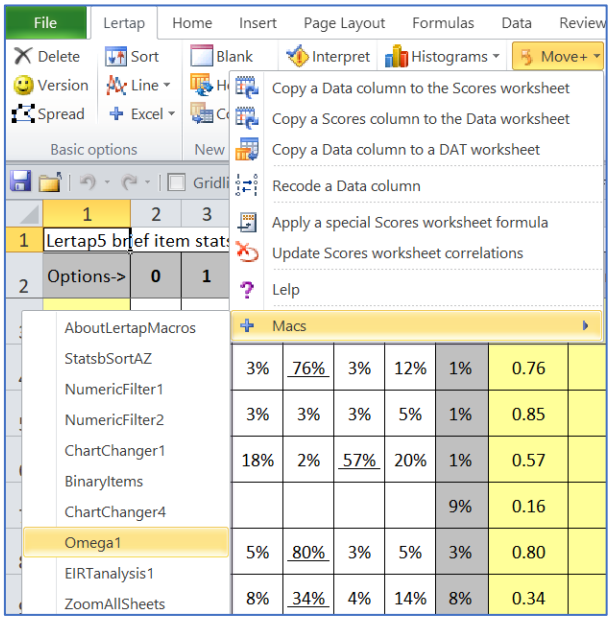
I can use the free "IRT Illustrator" program with the b(i) and a(i) values above to see how the ICCs (item characteristic curves) look. I'll do so with data from Q1 and Q5.

Now, one of my main objectives for this class is to use the Omega1 macro. Before I can do so, I need to get Lertap to make an "IStats" worksheet. I will use the "Item scores and correlations option" in Lertap:

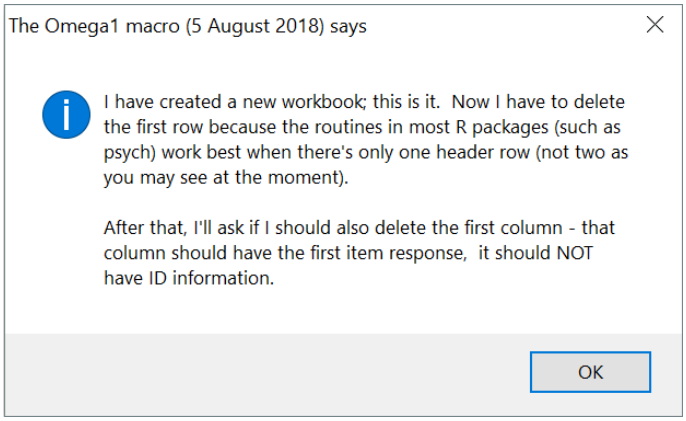
[5] The method Lertap uses to get the IRT statistics is only appropriate when all the test items have reasonable performance. In FIMS, item 12 was a problem: diff = .23, and disc = 0.04. It was a poorly-performing item.

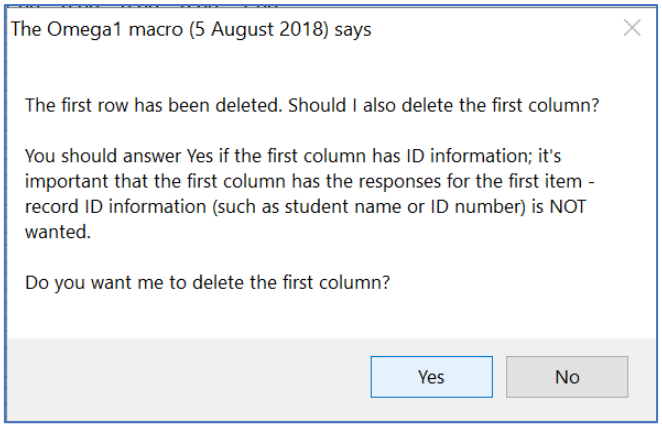


Now I will use the [Omega1 macro](#).

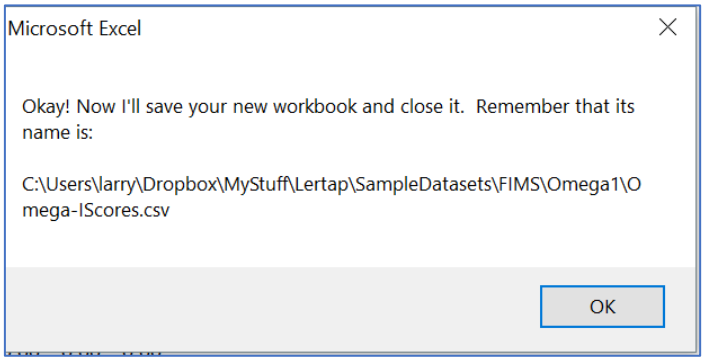


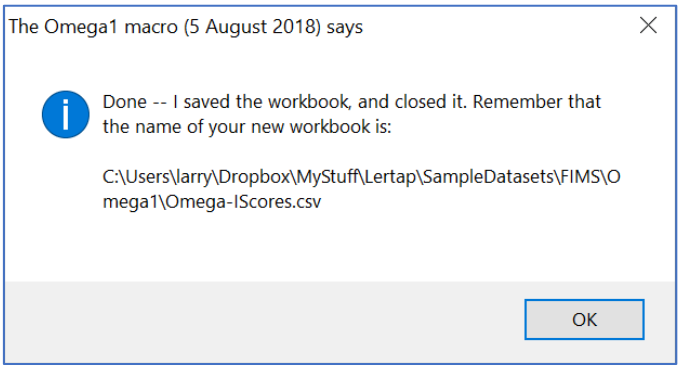
The macro creates a new workbook, and displays some messages:



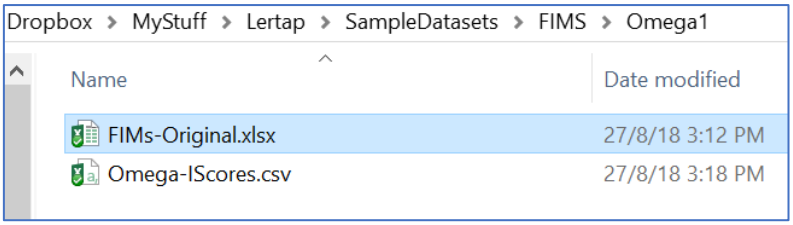


The answer to this question is almost always “Yes”.

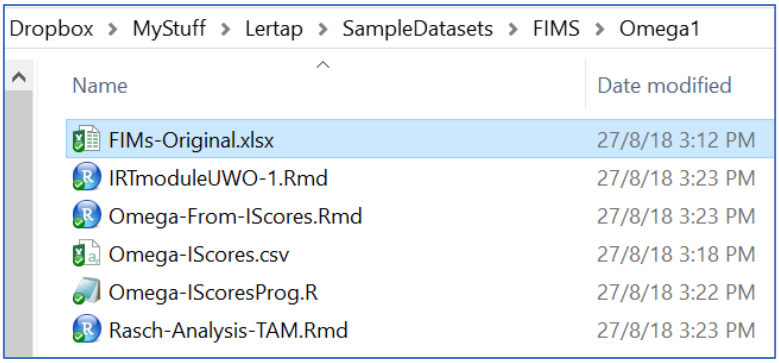




I can now see that the macro has added the csv file to my folder:



The csv file will be used by some Rmd files I’ve made to use with RStudio. The Omega1 macro will download these files, and save them in my folder:



Now I am finally ready to get results. I double-click on "Omega-From-IScores.Rmd" and we will see results during our class. I may also demonstrate what happens when I double-click on the other two Rmd files, "IRTmodulesUWO-1.Rmd", and "Rasch-Analysis-TAM.Rmd".

It would certainly be possible for you to repeat this exercise, and I suggest that you try it. You can get a copy of the FIMS data from here (it will come as a workbook called "AUS_JPN_2.xlsx").

Possible research topic: there are statistics in the omega factor analysis report that could be investigated further. One of these is the "explained common variance of the general factor"; I wrote a paper related to the omega factor analysis report which could be improved and expanded. (I should have used tetrachoric correlations before getting the eigenvalues, and I should also have included the "explained common variance of the general factor".)

Item banks and item pools were things I mentioned when I discussed the FIMS project. I will use this link to remind us of the design of the FIMS item pool.

The FIMS test I have been using has only 14 items. They were probably selected from the item pool to meet the objectives of part of the FIMS study – in this case we might assume that the objective was to create a short test to see how students would perform on some of the "performance objectives" used in the creation of test items for the pool.

There were more than 100 items in the FIMS pool (in many other subject areas item pools will usually have many more items, perhaps as many as 400). The little FIMS test I have been using is really just a small sample of items from the pool. I could, of course, draw another sample of 14 items from the pool and create another test. In other words, my 14-item FIMS test is just one of many 14-item samples I could draw from the pool.

I will now use a bag of marbles and the binomial theorem to talk about the effects of item sampling on test scores.

● **Topic insertion 1**: the binomial

Description. To this point there hasn't been a focus on test scores per se; we have been looking at how items have performed, DIF, and factor analysis (dimensionality).

Now we'll set up to talk about true scores by using a marbles binomial example. Lertap's csem1 report uses the binomial to estimate conditional standard errors of measurement.

Let's look at FIMS results again, focusing on the "csem1" report produced by Lertap[6].

Conditional standard errors of measur

Score	CSEM 1	CSEM 2	SEM
0	.00	.00	1.51
1	1.00	.85	1.51
2	1.36	1.16	1.51
3	1.59	1.36	1.51
4	1.75	1.49	1.51
5	1.86	1.59	1.51
6	1.92	1.64	1.51
7	1.94	1.65	1.51
8	1.92	1.64	1.51
9	1.86	1.59	1.51
10	1.75	1.49	1.51
11	1.59	1.36	1.51
12	1.36	1.16	1.51
13	1.00	.85	1.51
14	.00	.00	1.51

Suppose we consider a student whose observed test score was 7.

The CSEM2 column in the table can be used to form a 68% confidence interval which I will call "CI-68". The low score will be (7-1.65 = 5.35), the high score will be (7+1.65 = 8.65), so CI-68 = 5.35 to 8.65.

I will convert these raw observed scores to percentage correct scores: 7/14 = 50%, 5.35/14 = 38%, and 8.65/14 = 62%

I can then say, with 68% confidence, that the student's true score (percentage score) lies in the range 38% to 62%. This is a big range. I could make it smaller by using more than 14 items in the test.

What does this mean? Test scores are not "true" scores unless we use all of the items in the pool. A test score is an observed score and is not expected to equal the true score due to sampling error.

X=T+e is the fundamental equation of CTT: an observed test score equals the true score plus measurement error (e can be either a positive or a negative number).

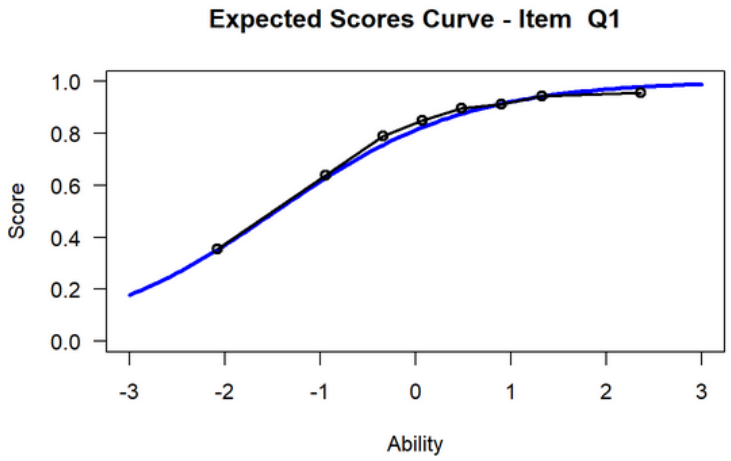
Note that CTT does not have a model that can be tested. X=T+e is just an equation, not a model. One of the advantages of IRT is that it is based on models that can be tested for goodness.

Demonstration

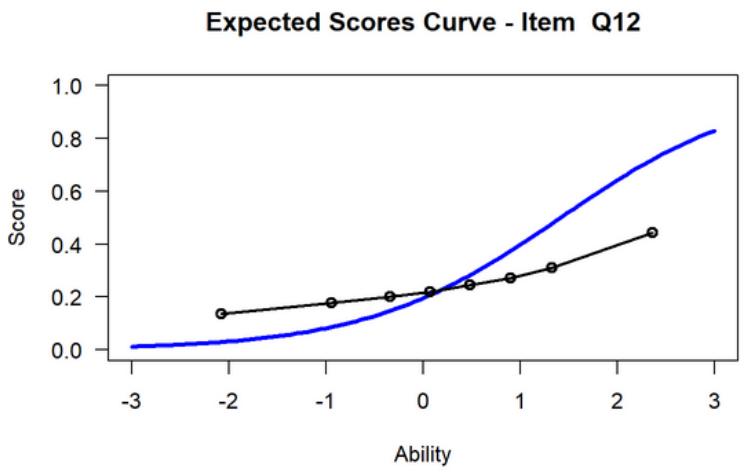
There are several ways we can test IRT results to investigate the goodness of model fit.

One way is seen when we run the Rasch-Analysis-TAM.Rmd program. I will display the output produced for the FIMS data (note to myself: it's in Dropbox\ MyStuff\ Lertap\ SampleDatasets\ FIMS\Omega1 – Copy \ Rasch-Analysis-TAM.html).

The TAM.Rmd program I wrote does not output actual fit statistics; instead it gives a good display of plotting empirical results on item ICC plots. Two examples are below:



[6] Refer to pages 124-126 of Crocker & Algina's text (1986) for a good reference.



The plots above suggest that the Rasch IRT model was a good fit for Item Q1, but not for Item Q12. This paper has examples of the empirical plots made by other programs.

The IRTmodulesUWO-1.Rmd program has another way, a very different way, of assessing model fit. The little table below indicates that the two-parameter logistic IRT model fits the FIMS data better than the Rasch model:

Is Model3, 2PL, better than Model2, the unconstrained Rasch model? That is, does Model3 appear to fit the data better than Model2? If the p-value below is significant, less than 0.050, then yes, it would seem so.

```
anova(fit2, fit3)
```

```
 Likelihood Ratio Table
         AIC      BIC   log.Lik     LRT df p.value
fit2 94269.78 94371.17 -47119.89
fit3 92174.91 92364.18 -46059.46 2120.87 13  <0.001
```

Possible research topic: investigate and compare the various ways IRT fit statistics are calculated and plotted. It seems to me that the methods used in the IRTmodulesUWO-1.Rmd module may be some of the most advanced at the present time.

Xcalibre and BilogMG are two commercial IRT programs, very popular ones – we have used them frequently at RMCS. They have other ways of assessing model fit. The IRToys program is a free R package which can also be used for IRT-fit research – it can work in conjunction with BilogMG. RMCS has licenses for Xcalibre and BilogMG. This paper has examples of the output from Xcalibre and BilogMG.

One of the classic ways of assessing model fit is the one mentioned in Professor Baker's book (see page 49). This method is used in the EIRT program. I will show you how easy it is to use EIRT with (for example) the FIMS data. (Note to myself: use the Omega1 folder, FIMs-Original.xlsx, and be sure to use Excel 2016, not Excel 2010.)

This paper has examples of the empirical plots made by other programs.

- **Topic insertion 2**: Reliability and Validity

Note: these two terms have not received any real focus thus far, but "reliability" will be coming up in the next exhibit.

If students do not seem to have much background in classical measurement topics, and if time permits, the "Beck Depression Inventory" may well be an interesting dataset, and an entertaining discussion.

- **Exhibit 4**: StuIQ

Description. This is an authentic study from New Zealand where the Department of Education undertook the development of an aptitude test for high school students (it was called "TOSCA", the test of scholastic abilities). They developed an item pool, and created two parallel forms, FormA and FormB, by balanced item sampling from the pool.

Background. This study will be used to discuss parallel forms, reliability, and the standard error of measurement (SEM) using CTT (classical test theory) terminology.

References. A discussion of the study, and a description of response coding, is here. Chapter 7 of the Lertap 5 manual is an appropriate reference; "reliability" is discussed on page 103 of the printed manual, with SEM featured on page 104.

Suggested reading. Chapter 7 of the Lertap 5 manual.

Resources required: Lertap5.xlsm (the main Lertap 5 workbook). A calculator. The dataset itself, an Excel workbook, may be reached at this link. A version of the dataset is on one of my laptops, "Lensito", where it is located under SampleDatasets / StuIQ / STUIQ-RMCS-2018b.xlsx. Note that this version of StuIQ has "Sex recoded" in column 146.

Description. See the link in References above, repeated here.

Demonstration

Use STUIQ-RMCS-2018b.xlsx. Make a copy. Use the "Interpret" and "Elmillon" options. Look for the correlation of FA-Tot and FB-Tot. Get a scatterplot, and possibly the regression line too.

Display the SEM equation, page 104 of Chapter 7. Substitute the FA-Tot/FB-Tot correlation 0f 0.84 for alpha, get the average FA score (29.17), and compute the SEM using the FA standard deviation (10.51); with SQRT(1.00 - 0.84) = 0.40, SEM = 4.20.

The 68% confidence interval (C.I.) for the average FormA score would be (29.17 - 4.20) to (29.17 + 4.20), or about 25 to 33. The 95% C.I. would be about (29.17 – 8.40) to (29.17 + 8.40), or about 21 to 38.

Discuss the practical interpretation of the confidence intervals, referring again to page 104 of the Lertap 5 manual.

Possible Research Questions

There is much that could be done. Were there gender (sex) differences on FormA? On FormB? Were the short-answer items (referred to as "supply items" in many other studies, including FIMS) more difficult than the multiple-choice items? What were the correlations between the scores on the short-answer items and the scores on the multiple-choice items?

- **Start Class 3**

Last week we did a lot.

We have now talked quite a bit about CTT, and we have also started to discuss IRT.

In CTT, fundamental terms are: item difficulty, item discrimination, test reliability, the standard error of measurement, and test dimensionality. X=T+e is the basic equation in CTT.

Last week I used the "StuIQ" project to introduce one of the ways test reliability is estimated: we find the correlation between two parallel forms, that is, two tests of equal length carefully selected from the item pool so that each subject topic and each performance objective is sampled to the same extent in each test form.

Demonstration: I will return to the StuIQ results briefly in order to look at FormA – FormB correlations. (Note: make sure Lertap's experimental features are on.)

I mentioned that it's not always easy to create parallel forms. Because of this, CTT users have employed another strategy: they divide a single test into two parts, and then find the correlation between the scores on both parts. This has problems too – how to split the test into the two parts?

Long ago, Professor Lee Cronbach proved that a statistic now known as "coefficient alpha", also called "Cronbach's alpha", was the best method to use to estimate the reliability of a test when we do not have two parallel forms. It's not necessary to split the test into parts in order to calculate alpha.

Another important point: the reliability of a test will increase if we add more items to the test (this is like drawing more marbles from the bag in order to get a better estimate of the true proportion of red marbles).

SEM, the standard error of measurement, is related to reliability by an equation:

$SEM = s.d.\sqrt{1 - alpha}$	s.d. is the standard deviation of the observed test scores

SEM is used when we want to derive a confidence interval which we believe may include a student's true score. Examples are found in Chapter 7 of the Lertap manual.

Summary of measurement basics to this point, especially as seen in CTT, classical test theory:

X = T + e <-- observed score = true score + error

e, the error, is affected by the number of items drawn from the item pool

longer tests have less measurement error

item discrimination greatly affects measurement error – we want all items to have good discrimination (at least 0.20)

e can be affected by the "dimensionality" of a test; tests with only one underlying factor may have less measurement error

reliability

is a measure of test consistency – if we are able to administer a test more than once, will the correlation of test scores be high? If so, the test is said to be reliable

parallel forms reliability has been regarded as one of the best ways to measure reliability – it is simply the correlation between two carefully-constructed tests created by stratified sampling from the item pool

coefficient alpha is a reliability estimate made popular by Professor Cronbach. Basically, it is based on dividing a single test into two half tests, finding the correlation between the scores on the two halves, and then correcting the correlation by applying the Spearman-Brown formula (or a formula like it).

coefficient alpha is without doubt the most common way of measuring test reliability (note: Lertap makes it possible to also get omega reliability)

SEM, the standard error of measurement

is a figure used to form a confidence interval, "C.I.", around X, a student's observed test score

the C.I. is an interval which we believe captures a student's true score with either 68% or 95% confidence

Concerning IRT, we have thus far looked at three models for dichotomous items. One of them, the "2PL" model, estimates two parameters: item difficulty and item discrimination.

As we know, CTT also has item difficulty and discrimination statistics. (*See Stats1b*)

How does CTT difficulty compare with IRT difficulty?

For any test, the two statistics correlate very highly, usually well over r = -0.95. The correlation is negative because high "diff" values in CTT correspond to easy items; in IRT, high b(i) values indicate high difficulty.

How does CTT discrimination compare to IRT discrimination?

The two statistics will usually correlate above r = 0.85 for any test.

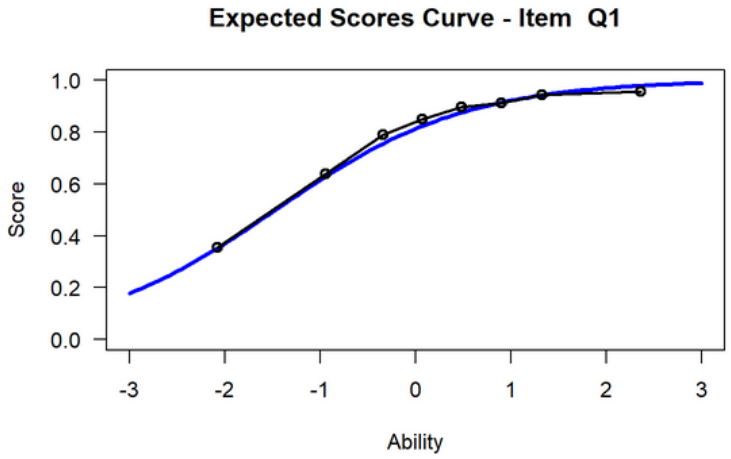
If CTT and IRT results correlate so highly, why is IRT sometimes regarded as a better method?

There are several reasons. One of the biggest is that the IRT difficulty statistic, b(i), is on the theta scale (-3.00 to +3.00), the same scale that is used for test scores when calculated by an IRT program. Another is that IRT is based on models that can be tested for goodness.

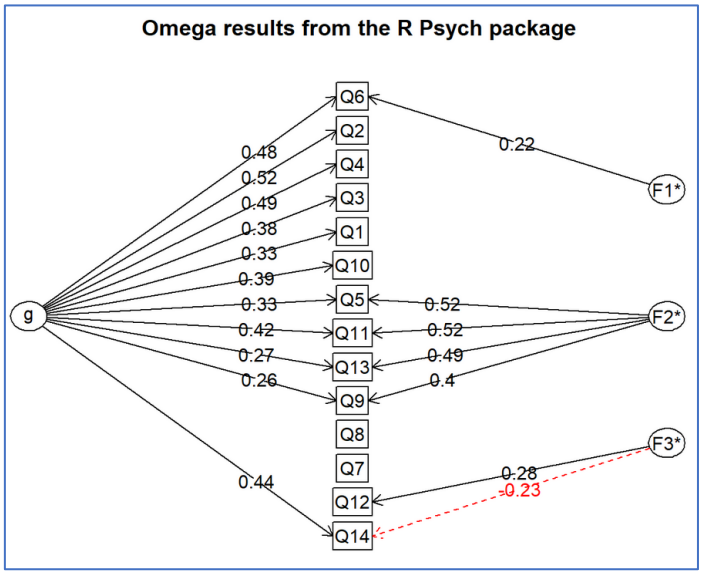
The "1PL" IRT model says that all items will have the same discrimination (usually 1.00), but different difficulties. This model is often called the "Rasch" model.

The "3PL" model adds a guessing parameter to the 2PL model.

We looked at model fit in IRT using Rmd scripts I have written. One way of assessing model fit is by plotting empirical results on the ICC, as exemplified here:



An interesting question came up last week when we were looking at the following output from one of the Rmd scripts: why was the line for Q14 in red as it emerged from F3?

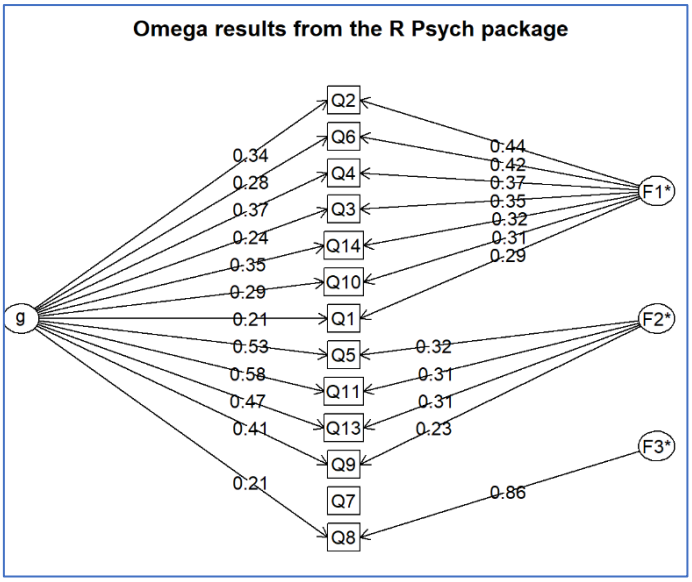


We can see that student responses on Q14 related quite strongly (0.44) to the "g" factor. The F3 factor affected responses to Q12 and Q14, but, in the case of Q14, in a "reversed" way. This may be due to the negative correlation between Q12 and Q14, as seen in Lertap's IStats report:

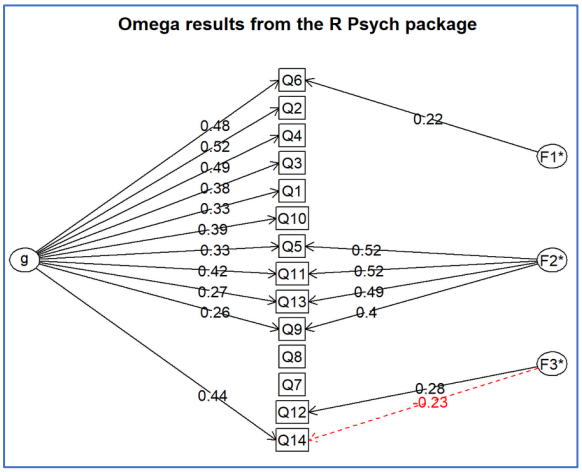
Record No.	Q11	Q12	Q13	Q14
Q9	0.30	0.03	0.27	0.19
Q10	0.21	-0.01	0.12	0.21
Q11	1.00	0.05	0.37	0.24
Q12	0.05	1.00	0.09	-0.04
Q13	0.37	0.09	1.00	0.15
Q14	0.24	-0.04	0.15	1.00
average	*0.22*	*0.02*	*0.16*	*0.17*
SMC	*0.30*	*0.02*	*0.20*	*0.17*

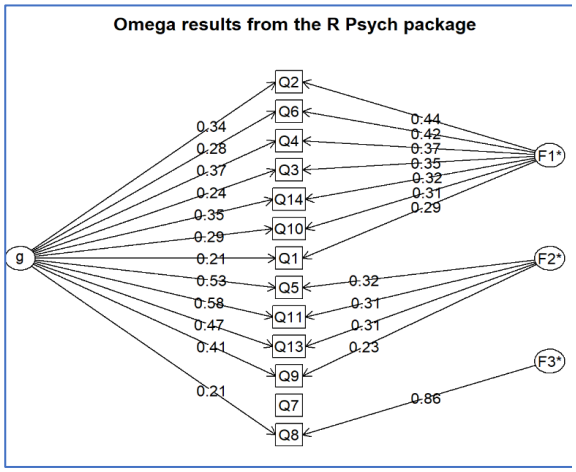
Item Q12 was <u>not</u> a good item. Its average correlation with the other items was only 0.02. This is very low – if we want to have a test where all items are measuring the same factor, then we want average correlations to be 0.20 or more.

I decided to delete Q12 from the Lertap analysis and then see how the omega output looks – after doing that, the results were as seen below – the changes were greater than I had anticipated – now all but one item, Q7, are affected by the "g" factor, and a new factor, "F1", has appeared with some substantial loadings.



I have copied both plots below. The original is on the left; without Q12 is on the right.





Reminder: I started to look at a factor analysis of the FIMS results to see if there might be an unexpected reading factor in addition to an expected mathematics factor.

I do not see a reading factor in these results. In fact, I find it difficult to interpret the output, and have some comments: (1) it would help if I could talk to the FIMS team members, mathematics teachers, to see if they can explain the results, and (2), I might try processing the results again, but do so by countries – at the moment, both Japanese and Australian students have been included – how would the results change if I analysed the two countries separately? (If you might be interested in this research question, visit the FIMS sample dataset website where there are Excel workbooks for each country: AUS.xlsx for Australia, and JPN.xlsx for Japan.)

- Review FIMS and the Rasch results from last week (reminder: Rasch is the 1PL IRT model)

  Rasch-Analysis-TAM.Rmd (<-- the program that made the output)

  Rasch-Analysis-TAM.html (<-- html output: look at this in class, see the WrightMap)

  TAM_Mod1_xsi.csv (<-- csv output: look at this in class, compare diff and b(i))

  *** *A **BIG** advantage of IRT is that item difficulty, b(i), is on the theta scale* ***.

Let's move forward. I have new topics to discuss.

- **Topic insertion 3**: EIRT (see this link for an introduction and more links)

EIRT is a good little program, and has been integrated into Lertap. It works well.

Last year I developed a handout on how to use EIRT. I have copies printed for today's class, but here is the link to it in case I forget to bring the copies with me.

Demonstration: use MathsQuiz, or FIMs, with EIRT to see which model fits better using chi-square tests of fit (as described in Professor Baker's IRT text, page 49). Remember to get error estimates of person scores (similar to SEM, the standard error of measurement in CTT); discuss.

- **Exhibit 5**: ZMed

Description. Has 100 cognitive items. The data are from a prominent medical school in Europe. The test is used to screen applicants and involves the use of a cut-score. Can be used to discuss classification consistency, and the effect of reducing the number of items on reliability, classification consistency, and the standard error of measurement. Can also introduce IRT here with the objective being to locate individuals on the theta scale.

Draw two boxes representing two testings, and how students will go from pass to fail, and fail to pass.

Notes at 8 September 2018

SampleDatasets\ZmedData\RMCS 2018\ZmedDataRMCS2018_50.xlsx

Available from this website.

Start discussion using this workbook. It has two subtests, one without a Mastery= setting, and one with (see the CCs worksheet)

Compare Stats1ul and Stats2ul

Focus on the Stats2ul sheet and classification consistency. Show how little-h help calls in explanations and links to relevant documents.

SampleDatasets\ZmedData\Zmed_3_Testlets_Increasing_Nits.xlsx (on my personal computer)

The Statsf reports show the effects of increasing the number of items

SampleDatasets\ZmedData\RMCS 2018\ ZMed-ManySubsSpecialSummaries-3Aug16.xlsx (on my personal computer)

Has a specially-formatted report showing the effect of test length on a number of statistics; here's a screen capture of the report:

Test Length	Test Score Mean	Test Score% Mean	Test Score Standard Deviation	Test Score% Standard Deviation	Reliability (alpha)	Standard error of measurement (SEM)	SEM as a %	95% confidence interval at a score of 50%	Estimated Number of Classification Errors	Estimated % Classification Errors
10 items	5.74	57.40%	2.30	23.00%	0.7108	1.24	12.40%	27% to 74%	526	21%
20 items	10.30	51.50%	4.70	23.50%	0.8548	1.79	8.90%	33% to 67%	422	17%
30 items	14.96	49.90%	6.67	22.20%	0.8867	2.25	7.50%	35% to 65%	376	15%
40 items	19.42	48.50%	8.86	22.10%	0.9145	2.59	6.50%	37% to 63%	327	13%
50 items	23.91	47.80%	10.89	21.80%	0.9268	2.95	5.90%	38% to 62%	302	12%
60 items	30.38	50.60%	12.51	20.90%	0.9338	3.22	5.40%	39% to 61%	286	12%
70 items	36.89	52.70%	14.24	20.30%	0.9398	3.49	5.00%	40% to 60%	270	11%
80 items	42.79	53.50%	15.95	19.90%	0.9447	3.75	4.70%	41% to 59%	257	10%
90 items	48.71	54.10%	17.72	19.70%	0.9497	3.97	4.40%	41% to 59%	243	10%
100 items	53.77	53.77%	19.85	19.80%	0.9552	4.20	4.20%	42% to 58%	230	9%

There were 2,470 applicants for study at the medical school in 2015.

Only 1,400 could be accepted so they used a "mastery" score of 50%.

- **Topic insertion 4**: CAT, computerized adaptive testing

One of the big advantages of IRT is that we can use "CAT" methods to estimate where a person is on the theta scale, sometimes using only a small number of items.

See this CAT example from Professor David Weiss.

Discuss what's required for CAT: a testing centre equipped with computer terminals – see this video from Pearson VUE. Not all countries are presently well equipped with such testing centres. (CAT software is also required, of course, and an item pool. Consider ASC's CATSIM for possible future class demos, and potential research projects.)

- **Exhibit 6**: BFI-25 (a personality test)

Description: refer to the website.

Lertap analysis

Use \SampleDatasets\BFI-2018\RMCS 2018\BFI_Psych_Dataset.xlsx
(this is the same as the xlsx workbook on the website)

Discuss Data and CCs worksheets

Use the Interpret option

Freqs okay?

Use the Elmillon option

Note that there are many new worksheets because there are six subtests.

Compare alpha values

all items = .690
“agree” items = .697
“consc” items = .723
“extra” items = .759
“neuro” items = .811
“open” items = .599

Stats5b: the neuro subtest’s item cor. values were high
Stats6b: the open subtest’s item cor. values were not as high

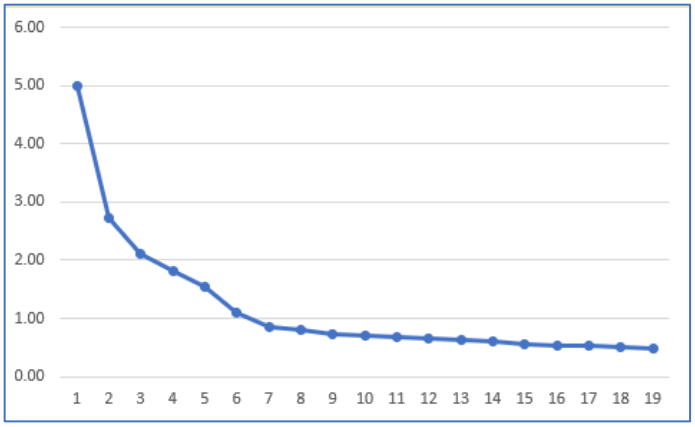
Look at the Scores worksheet

Four subtests have similar means; neuro’s mean is low
The neuro subtest has negative correlations

Dimensionality: IStats scree plot (how many factors?)

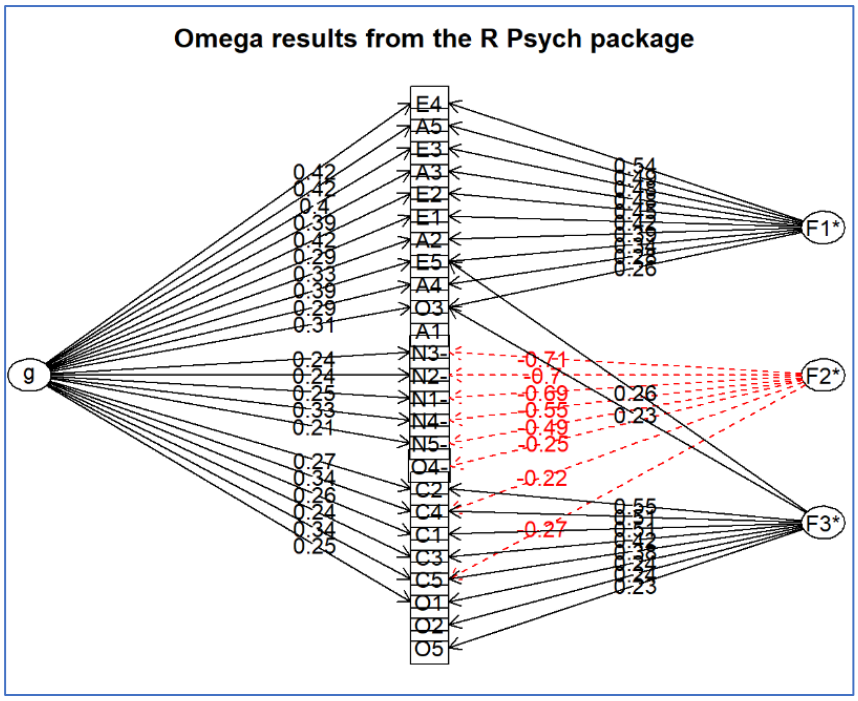
Use Item scores and correlations for the “All” subtest (all 25 items)

Make a scree plot from the eigens in the IStats report:



The scree plot above indicates that there may be 6 factors.

Dimensionality: Omega analysis (how many factors?)



The omega analysis implies that there may be 4 factors. The red arrows indicate reverse loadings – this would be expected for the neuro items as they tended to have negative correlations with the other items (the Lertap results show this).

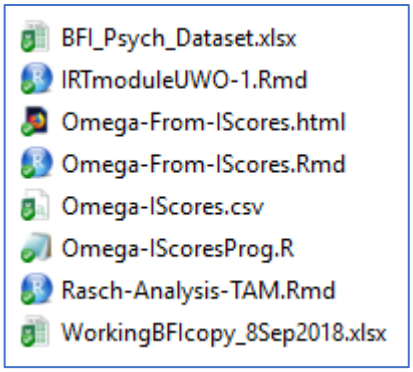
```
Schmid Leiman Factor loadings greater than  0.2
        g   F1*   F2*   F3*   h2   u2   p2
A1                          0.06 0.94 0.42
A2   0.33  0.39             0.27 0.73 0.40
A3   0.39  0.48             0.39 0.61 0.39
A4   0.29  0.28             0.17 0.83 0.49
A5   0.42  0.49             0.42 0.58 0.42
C1   0.26              0.51 0.33 0.67 0.20
C2   0.27              0.55 0.38 0.62 0.19
C3   0.24              0.42 0.24 0.76 0.24
C4   0.34       -0.22  0.51 0.43 0.57 0.26
C5   0.34       -0.27  0.38 0.34 0.66 0.35
E1   0.29  0.42             0.27 0.73 0.32
E2   0.42  0.45             0.42 0.58 0.43
E3   0.40  0.48             0.41 0.59 0.40
E4   0.42  0.54             0.48 0.52 0.37
E5   0.39  0.34        0.26 0.34 0.66 0.45
N1-  0.25       -0.69       0.55 0.45 0.12
N2-  0.24       -0.70       0.55 0.45 0.11
N3-  0.24       -0.71       0.56 0.44 0.10
N4-  0.33       -0.55       0.43 0.57 0.26
N5-  0.21       -0.49       0.28 0.72 0.15
O1   0.25              0.24 0.16 0.84 0.40
O2                     0.24 0.09 0.91 0.20
O3   0.31  0.26        0.23 0.22 0.78 0.42
O4-             -0.25       0.09 0.91 0.00
O5                     0.23 0.08 0.92 0.30

With eigenvalues of:
  g F1* F2* F3*
2.3 1.9 2.3 1.5
```

The neuro (N1 – N5) subtest is the clearest factor (F2). The open subtest (O1 – O5) is ambiguous. The g factor does not affect the open subtest (O1 – O5), and does not affect item A1.

The “With eigenvalues of:” line indicates 4 factors.

IRT analysis

The Omega1 macro left these files in my folder:



There are two Rmd files in the folder which may be used for IRT: “IRTmoduleUWO-1.Rmd” and “Racsh-Analysis-TAM.Rmd”.

However, these two Rmd scripts assume that item responses are dichotomous, like the ones seen in FIMS and MathsQuiz where each item was scored as right or wrong, with item scores of 0 (zero) and 1 (one).

But the items in the BFI study are polytomous. Item scores range from 1 (one) to 6 (six) on each item. We need a new Rmd script, one for polytomous items.

One is available at our class webpage:

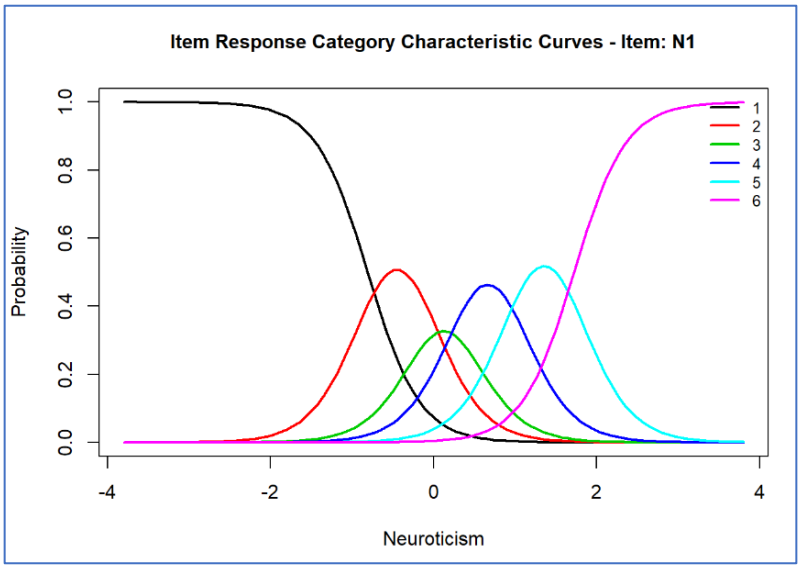
Useful links

- **(1) Lertap 5.10.99 for Excel 2016 (Windows and Macintosh)**
- **(2) How to upgrade Lertap to process more than 100 records.**
- **(3) The "updates summary" for Lertap versions**
- **(4) Website with sample data and exercises**
- **(5) Some tips and tricks to use with Excel & Lertap**
- **(6) YouTube video: Installing Lertap**
- **(7) YouTube video: Using Lertap**
- **(8) RMCS class lecture notes for 2018**
- **(9) RMCS class references for the year 2018**
- **(10) IRT for dichotomous items using R (U of W Ontario)**
- **(11) IRT for polytomous items using R (U of W Ontario)**
- **(12) Polytomous IRT using the BFI-25 study (Rmd file)**
- **(13) R script to install.packages (needed for omega reliability calculations).**
- **(14) The FIMs-Original.xlsx workbook for Lertap**
- **(98) Link to the Lertap 5 e-store**

The Rmd script we want is (12) Polytomous IRT using the BFI-25 study (Rmd file). It leads to an Rmd script called “IRTmoduleUWO-2-August2018.Rmd”.

The script uses the Graded Response Model (“GRM”) to analyse item performance.

The following graph displays GRM model results for the first neuro item, N1:



Unfortunately, Professor Baker’s book does not discuss polytomous IRT models.

A suggested website to serve as a starting reference for polytomous IRT is this one at the University of Kansas. Suitable textbooks are found here – the books by DeMars and Embretson are ones I recommend, and they are available at the BUU Library.

SUMMARY COMMENTS