Measurement Classes with Lertap5, TAM, jMetrik, SAS University, ConQuest5, BILOG-MG, SPSS, Xcalibre, +

Larry R Nelson
Curtin University / Burapha University / Prince of Songkla University

Document date: 10 February 2024

Check for an updated copy of this paper by using this link.

What's been updated?

Work in the latter half of 2023, carrying into 2024, has focused largely on the practical use of "Jamovi", a free data analysis system quite a bit like JASP.

I first used JASP back in 2021 when I worked on this paper having to do with McDonald's omega reliability coefficient.

What attracted me to Jamovi was its ability to read Excel workbooks directly, something JASP could not do when last I looked at it mid-2023.

Thus far two papers have resulted: TestAnalysisWithJamovi, and UsingRStudioAndSnowIRT. The first of these has to do with CTT, classical test theory, while the second looks at the application of the Rasch IRT model.

Both JASP and Jamovi are free-standing cross-platform apps.

In response to a Burapha University colleague, I added a link here to a dated but still possibly useful introductory document introducing basic measurement topics to medical practitioners. Has mention of relevant apps, a solid overview of multiple-choice items, and a real-life example of item and test analysis with a comparison of results from two countries.

This fairly recent document exemplifies some personal approaches to measurement classes. It includes work initiated September 2022, based on practical real-life cases.

Note: the text below has not changed since the end of 2019.

Towards the end of the last edition, I suggested that a CRAN R package called "TAM" showed promise. I have now confirmed that, and have added TAM, a free R package, to the discussion below. Like the other programs mentioned here, TAM has pluses and minuses – as you'll see, I recommend it.

Speaking of R, some quality IRT teaching material has been prepared by staff at the University of Western Ontario, UWO. Much of the R code in their dichotomous IRT document has been incorporated, and extended, in this Rmd script, "IRTmoduleUWO-1.Rmd"; download it from the link at the bottom of this page.

Patrick Meyer has added a useful IRT tool to his website. While I continue to suggest that, by and large, his jMetrik program is too cumbersome to use, his new IRT tool could be of considerable use in measurement classes. The tool is called "IRT Illustrator".

Lertap5 now works on Apple Macintosh computers; read this for more information if you please. It also now has its own Rasch routine (new November 2019).

I have inadvertently omitted to reference the excellent software sourcing support provided by NCME, the National Council on Measurement in Education. It was here up until July 2018 (let's hope it returns as it was a comprehensive resource – in November 2019 it was still missing – still missing 16 December 2020.)

16 December 20: do the programs include the calculation of omega?

Background

This document has largely been motivated by an annual need to update the resources used in my university classes on tests and measurement. I often make substantial changes in the software I use, especially if there have been some promising recent developments, or if someone has sent in a note regarding a program they recommend and it's one that I'm not familiar with.

The classes cover CTT, classical test theory, and IRT, item response theory, usually in just five three-hour sessions. The first half of each class features a lecture / demonstration mode, then a tea break, then a workshop with hands-on exercises.

Starting with CTT, I discuss typical steps in the construction of a cognitive test: given a subject, say, for example, Earth Science, and a year level, say junior high school, my classes begin with the matter of developing an item matrix, with subtopics along the top (*earthquakes, minerals, water, ….*), and cognitive taxonomy levels down the side (*remembering, understanding, applying, ….*), ending up with the concept of an "item pool".

Then I briefly discuss item writing, with an almost exclusive focus on multiple-choice items. Following this we get into a discussion of creating a test with a given length by item sampling from the pool, going from there to test administration, collecting item responses, assessing test reliability and measurement error.

We look at the relationship between CTT item statistics, difficulty and discrimination, reliability, and measurement error – this gets into distractor analysis and the likely need to revise some of the test items. I include the derivation and use of conditional standard errors of measurement based on the binomial error model. If time permits, I cover the limitations of coefficient alpha and introduce alternative reliability measures, such as McDonald's omega.

Having presented traditional basic CTT concepts, I delve into "mastery" testing and the statistics associated with the use of cut scores. This is a good chance to introduce DIF, differential item functioning, but DIF is a topic I'll bring in only if the classes are moving along well and there appears to be sufficient time.

From here we move into latent trait theory, theta, and IRT. I link CTT item statistics, difficulty and discrimination, to their equivalents in the 2PL IRT model. We get into other dichotomous models, item fit and then, finally, move on to CAT, computerized adaptive testing.

The last three-hour class session discusses affective scales and polytomous IRT, again bringing in CAT.

It's fairly basic stuff, not much more than an introduction to a smattering of core topics. But it sets students up for additional (optional) studies in the development and use of tests.

Software

I have students complete practical exercises in the workshops. I need appropriate software for them to use. I want a single package which covers CTT and IRT and involves a minimum of scripting. It should be easy to define how items are scored. It must be backed up by thorough documentation, and, *critically*, it should be very easy to setup and to get running – the inclusion of sample data will always be welcome. A capability to run with Microsoft's Windows operating system is assumed – an ability to also run under Apple's MacOS is not assumed but seen as desireable.

My own professional interests, as far as psychometrics go, has centered on the use of cognitive instruments. This shows in my comments below, however I will point out that some packages (such as SPSS and JASP), while incapable of processing raw (unscored) cognitive item responses, do have certain utility for supporting the analysis of the affective items found in scales.

Ideally, software should be free for students to use, even if only on a limited term. If can be free and have restrictions as long as it has the capability to handle a "respectable" number of test items (say at least 80) and a "reasonable" number of respondents (say at least a few hundred).

There is such a package. It's my "Lertap 5" system[1]. Lertap has been around since 1973, running on mainframe computers, then desktop computers. For all but the last five years it only supported CTT[2], something I think it has done fairly well – in my somewhat biased opinion it has more than a few strengths, particularly, perhaps, when it comes to distractor analysis, the use of cut scores, and differential item functioning. It creates a variety of graphics to assist in the interpretation of tabled results.

In 2015 I added support for both dichotomous and polytomous IRT analyses. I use it to cover basic IRT topics. But there are some "howevers": I didn't write the code used for Lertap IRT -- it's open source, and I have not yet gone into it to polish it as I should; its documentation is pretty weak; it does not output empirical ICCs (item characteristic curves).

In late 2019 a Rasch item analysis capability was added to Lertap5. Although slow when running with the latest versions of Excel (Mac and Windows), it is nonetheless useful with classes.

I feel that graphical summaries of item performance can be invaluable. Not just for me, but also for students. Plots which trace how item options perform are an example, and they're a core feature in Lertap 5.

I am always on the lookout for IRT programs capable of displaying well-formatted item characteristic curves (ICCs) – there are several of these, even the ones in Lertap aren't too bad. But I want these curves to also provide some indication of model fit, and programs that will do that remain quite limited in number (*as far as I know*), especially when I look for a program free for students to use. Lertap 5 is one of the programs which falls short in this area – it's free for students, yes, but its ICCs are limited.

[1] Lertap5 allows for an unlimited number of items, but its free "Mini" version has a limit of 250 respondents.
[2] This isn't strictly true. From its inception, Lertap has had an option to output Lord's 1980 CTT-based approximations to IRT's 2PL parameters. See this topic.

So it is that I frequently renew my search for alternatives to Lertap 5. Something which also covers both CTT and IRT in a single package but has better ICC plots.

jMetrik and SAS University were two possibilities that came immediately to mind when I published the initial version of this paper in July, 2017. They are free systems for academic users. I did know of the TAM package back in July, but had not taken the time to learn more about it. Now I have; it has some good points (discussed below), and it's free.

jMetrik has benefited from substantial enhancements at the end of 2016, when its support for IRT was extended to cover more IRT models. jMetrik's graphics are excellent (*in my opinion*), and they include an option to overlay IRT plots with empirical points.

ConQuest5

The University edition of SAS was released in 2014. A special macro in Lertap 5 makes it possible so that users could take advantage of the strengths found in SAS's IRT procedure. It works well; I demonstrate its use in this paper.

SAS is programmable – in a 2014 text, Cody and Smith show how SAS can be used to get CTT statistics. They've included sample SAS code, and sample test results, to be used with the text. Up until now I had not tried to use the CTT programs. (Actual SAS code modules are found at the end of this document.)

TAM is not exactly a new CRAN package. What has brought it to life, and to my attention, is a TAM Tutorials website. As of this document's date, the website was still under development but nonetheless showing promise.

I set about to check out jMetrik's new version, and to have a look at the CTT capabilities of SAS University, in the previous edition of this paper. I have retained my original comments below, and now have added fresh material on TAM. Might one of these be better than Lertap 5 for my classes?

System requirements

Both SAS University and jMetrik use virtual machines and will run on Linux, Macintosh, and Windows computers. jMetrik is easy to install; its setup program includes a Java virtual machine if user computers don't already have one installed.

jMetrik's virtual environment is smooth and transparent to users. Such is not the case with SAS University where users have to choose their preferred virtual machine, install it and then configure it. SAS University then runs via a web browser. See this page for a complete list of requirements for running the University Edition of SAS. Here are two links to better understand it: what it incudes, and some of its limitations.

TAM is a CRAN package. It requires a version of R; R is free, and there's more than one flavour of it. All flavours will work under Windows, and under MacOS.

Lertap 5 requires a Windows machine and Microsoft Excel. It will also run on a Macintosh providing it has the latest version of Excel installed[3].

Item responses and item scores

I should clarify these terms. On a mid-term exam with five multiple choice question, each question having five options {A, B, C, D, E}, student Sally's five answers were {B, C, C, E, A} – these are the options she selected on each of the five questions – they are her "item responses".

If the correct answers to the five items were {B, D, C, E, B}, and if each correct answer was worth one point, then Sally's item scores would be (1, 0, 1, 1, 0), for a total score of 3 on the exam (60% as a percentage-correct score).

On an end-of-term class survey with five Likert-style items, each item with five options {SD, D, N, A, SA} for {strongly disgaree, disagree, neutral, agree, strongly agree}, Pedro's five answers were {SD, SA, N, A, D}. If the scoring weights on each item were {SD=1, D=2, N=3, A=4, SA=5} then Pedro's item scores would be {1, 5, 3, 4, 2} for a total survey score of 14.

Now, it is not at all uncommon to find that items using a Likert response style will reverse the scoring on some of the items – for example, an item might be negatively worded, making SD a positive response on that item – the scoring weight (or points) for such an item might then be {SD=5, D=4, N=3, A=2, SA=1}.

With this in mind, I can now present an important question: does the software app I want to use expect to have item scores as its input, or will it allow me to enter item responses and then score each of them according to scoring "rules" that I provide? As mentioned below, there are some apps that deal only with item scores – one cannot undertake a full item analysis with such apps, but at times they are nonetheless definitely of use. For those apps that allow for item scoring "rules" to be entered, the ease with which the rules are specified will often be a matter of interest.

Selected datasets

When I set about to work on this paper, my main focus was going to be to experiment with the SAS CTT programs found in the Cody-Smith text (2014). So it was that, at the outset, I selected one of the sample test results files which accompanied the text, a 56-item biomedical multiple-choice test sat by 137 university students. I ended up running item responses through Lertap 5, jMetrik, and SAS University, getting CTT and IRT results for each.

Then I added results from another test, a 48-item multiple-choice test on geology sat by over 4,000 senior high-school students in the U.S. several years ago. This was done to get the "sample size" up to something more respectable for IRT item calibration.

Later I added a third dataset found on the TAM Tutorials website. It's based on test items developed for the FIMS study. I used the items in classes in August 2017, and also in an invited lecture to post-graduate Faculty of Nursing students at Burapha University in September 2017.

[3] The EIRT add-on for Excel will not work on a Mac. Please see this document.

Applying Lertap 5

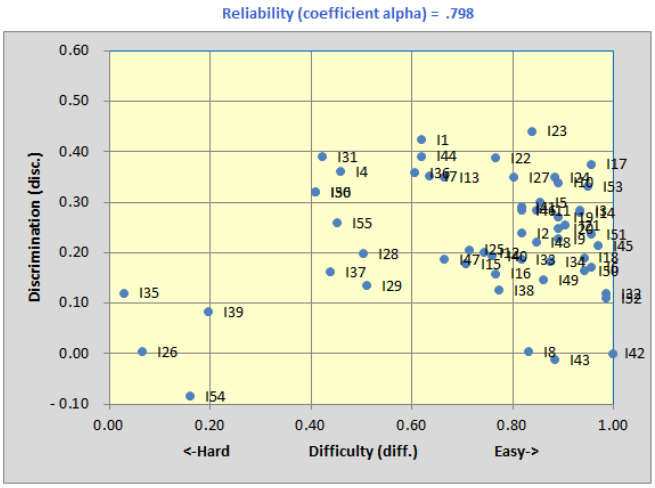
Lertap 5 produces multiple "reports" for a test. Some of these reports are brief, using a single line in a table to summarise the performance of each item. Other reports are usually quite longer, with some using a page of output, or nearly a page, for each item.

Here is an example of Lertap 5's brief report, "Stats1b", for the 56-item biomedical test:

Lertap5 brief item stats for "Test1", created: 7/04/2017.

Options->	A	B	C	D	E	other	Difficulty	Discrimination	?
I1	1%	1%		62%	36%		0.62	0.42	C
I2	1%	5%	82%	1%	10%	1%	0.82	0.24	
I3		5%	1%		93%		0.93	0.28	AD
I4	17%	46%	18%	12%	7%	1%	0.46	0.36	
I5	6%	7%	85%	1%	1%		0.85	0.30	D
I6	1%	96%	3%				0.96	0.17	DE
I7	64%	21%	6%	1%	8%		0.64	0.35	D
I8	7%	83%		4%	5%	1%	0.83	0.00	AC
I9	9%		89%	1%	1%		0.89	0.23	B
I10		3%	4%	4%	89%		0.89	0.34	A

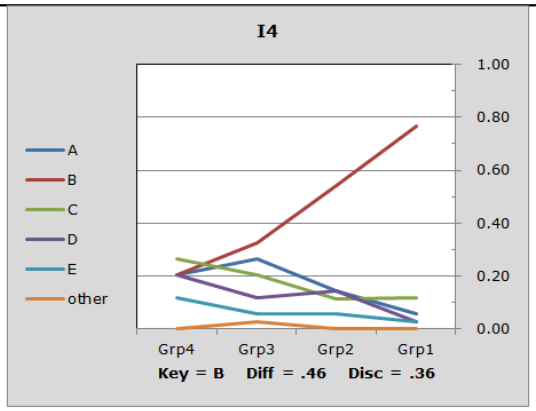
The bottom of the Stats1b report includes a scatterplot which looks like this:



The scatterplot rapidly summarizes how test items have performed. We see, for example, that many items were easy, having a "difficulty" of 0.80 or higher. A few items, I26, I54, I8, I43, and I42 had discrimination figures at or below 0.00, always an unwanted outcome. (To read a bit more about this plot, click here.)

Another Lertap 5 report has option trace line plots and related data tables for each item, as seen below for item 4[4]:

[4] See this paper for more about Lertap 5's plots.



I4 (c5)					*alpha with & without:*		*0.798*	*0.791*
option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	0.00	23	0.17	-0.18	-0.27	38.43	-0.41	
B	1.00	63	0.46	0.36	0.45	43.70	0.47	
C	0.00	24	0.18	-0.17	-0.25	38.63	-0.37	
D	0.00	17	0.12	-0.14	-0.22	38.71	-0.36	
E	0.00	9	0.07	-0.14	-0.26	37.78	-0.51	
other	0.00	1	0.01	-0.04	-0.17	38.00	-0.48	

The statistics which follow result when Lertap 5 users set up an item analysis using a cut score[5]:

Summary group statistics

	n	n(%)	avg.	avg%	s.d.	min.	mdn.	max.
masters	1,204	25.55%	38.6	80%	2.9	35	38	48
others	3,508	74.45%	24.6	51%	6.7	4	26	34
everyone	4,712		28.1	59%	8.5	4	29	48

This was an upper-lower analysis based on a mastery cutoff percentage of 72.91666% (cut score = 35).

Variance components

	df	SS	MS
Persons	4711	7028.19	1.49
Items	48	6790.98	141.48
Residual	226128	42624.49	0.19

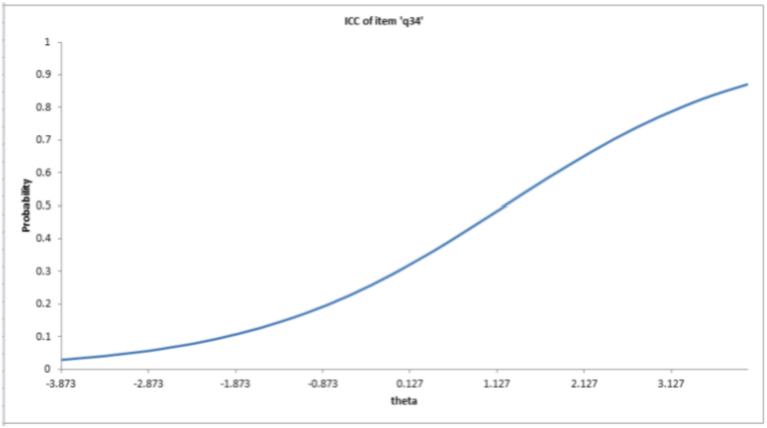
Hoyt's reliability coefficient:	0.874	
CSEM at the cut score:	2.915	
Livingston's coefficient:	0.929	
Index of dependability:	0.912	
Estimated error variance:	0.004	
For 68% conf. intrvl. use:	0.067	
Prop. consistent placings:	0.878	(Estimated number of incorrect classifications: 573)
Prop. beyond chance:	0.655	

An interesting class exercise has students varying the cut-score so that its effect on the proportion of consistent places can be investigated (as in this paper).

IRT in Lertap 5: an Excel add-in, "EIRT", is called by Lertap 5 for IRT model calibration, and for corresponding plots of item characteristic curves. The following is a typical example:

[5] A 48-item test sat by 4,712 students was used in this analysis. The cut-score was set at 35.

EIRT's plot of the ICC for q34 is shown below. No empirical results are plotted.



Data preparation in Lertap 5: data may be keyed directly into Lertap 5; all that's involved is typing in an Excel worksheet, an activity many students are familiar with. A shortcut is available to ease the work required.

Data may also be imported from csv and text files. Please refer to these two topics: here for csv, and here for text. Users with data in the old Iteman3 format may make use of a special macro to import Iteman3 files into Lertap 5. Ready-to-use datasets for classes are available at this website; a standard Lertap 5 option makes it possible to draw a random set of records from a dataset.

Scoring a test with Lertap 5 is accomplished by writing two or three lines of syntax in a "CCs" worksheet; examples are seen here and here.

I have already mentioned that Lertap 5 has been my "go-to" program of choice for measurement classes. Its support for CTT is comprehensive and, with its use of numerous graphics, colourful[6]. It also has extensive documentation, and ready-to-use datasets. Item scoring is very simple.

Lertap 5's main weakness, as far as my classes go, has to do with IRT – its ICC plots are just a bit short of adequate, and there is no possibility of overlaying them with empirical results.

Applying jMetrik

The two tables below were obtained by using jMetrik Version 4.0.5, released December 20, 2016. I used the same dataset introduced above, the 56-item multiple-choice test taken by 137 university students.

In the item analysis report below I have copied results for just the first four items in order to save space.

[6] See reviews, and documentation.

ITEM ANALYSIS
codysmith56.BIOMEDTEST
May 2, 2017 09:15:45

Item	Option (Score)	Difficulty	Std. Dev.	Discrimin.
i1	Overall	0.6204	0.4871	0.4233
	A(0.0)	0.0073	0.0854	-0.2252
	B(0.0)	0.0146	0.1204	-0.0578
	C(0.0)	0.0000	0.0000	NaN
	D(1.0)	0.6204	0.4871	0.4233
	E(0.0)	0.3577	0.4811	-0.5075
i2	Overall	0.8175	0.3877	0.2377
	A(0.0)	0.0146	0.1204	-0.2186
	B(0.0)	0.0511	0.2210	-0.0811
	C(1.0)	0.8175	0.3877	0.2377
	D(0.0)	0.0073	0.0854	-0.0835
	E(0.0)	0.1022	0.3040	-0.2407
i3	Overall	0.9343	0.2487	0.2847
	A(0.0)	0.0000	0.0000	NaN
	B(0.0)	0.0511	0.2210	-0.2444
	C(0.0)	0.0146	0.1204	-0.2986
	D(0.0)	0.0000	0.0000	NaN
	E(1.0)	0.9343	0.2487	0.2847
i4	Overall	0.4599	0.5002	0.3615
	A(0.0)	0.1679	0.3751	-0.2409
	B(1.0)	0.4599	0.5002	0.3615
	C(0.0)	0.1752	0.3815	-0.2324
	D(0.0)	0.1241	0.3309	-0.1887
	E(0.0)	0.0657	0.2487	-0.1764

RELIABILIY ANALYSIS

Method	Estimate	95% Conf. Int.	SEM
Guttman's L2	0.8107	(0.7625, 0.8533)	2.6209
Coefficient Alpha	0.7981	(0.7467, 0.8435)	2.7069
Feldt-Gilmer	0.8042	(0.7545, 0.8483)	2.6651
Feldt-Brennan	0.8033	(0.7533, 0.8476)	2.6712
Raju's Beta	0.7981	(0.7467, 0.8435)	2.7069

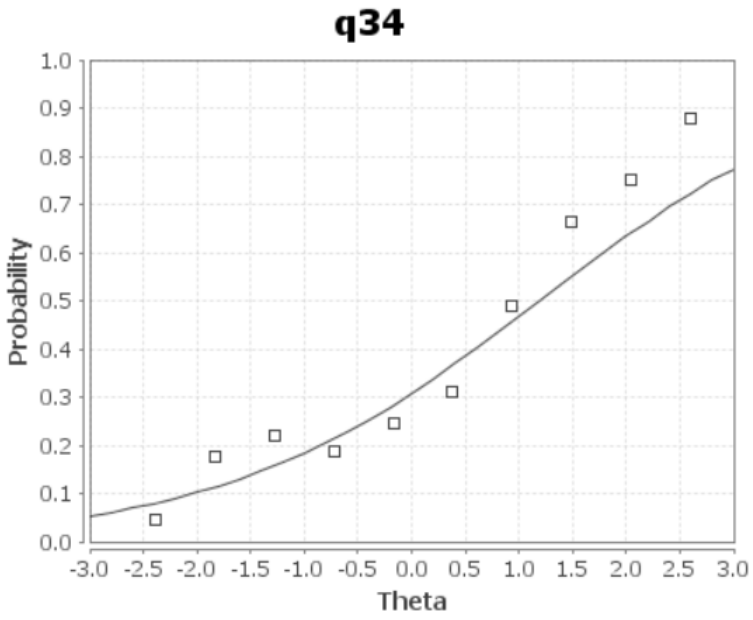
Getting jMetrik to use a cut score is straightforward. I had jMetrik undertake an item analysis of the 48-item test used to get Lertap 5's cut score analysis seen above, again using a cut score of 35. jMetrik's results are shown in the little table below[7]:

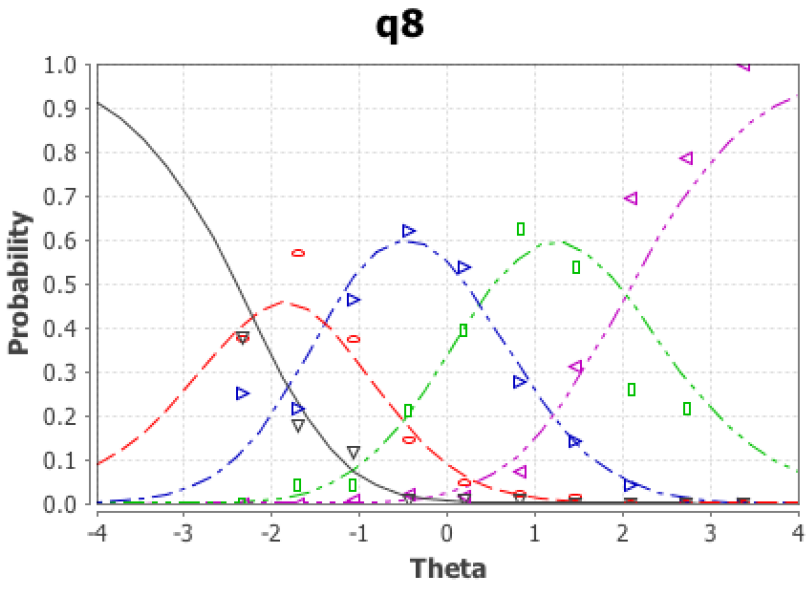
DECISION CONSISTENCY

```
Huynh's Raw Agreement Index = 0.86
     Standard Error of Agreement: 0.00
     95% Conf. Int. of Agreement: (0.86, 0.87)
Huynh's Kappa = 0.64
     Standard Error of Kappa: 0.00
     95% Conf. Int. of Kappa: (0.63, 0.65)
KR-21: 0.86
Beta-binomial alpha: 4.63
Beta-binomial beta: 3.27
```

IRT support in jMetrik is *comprehensive*, undoubtedly a strong feature. I particularly like it as it can output well-formatted empirical ICCs. I show two examples below. The first is from the same 48-item cognitive test mentioned above; the second is an example taken while fitting a five-option Likert item using the partial credit IRT model[8].

[7] Lertap 5 estimates the two Huynh statistics using methods mentioned in this paper.
[8] The second plot (Q8) is best seen in colour.





Data preparation in jMetrik: item response data may be in a csv file, or a text file. Data may not be directly entered in jMetrik itself.

Scoring test items in jMetrik involves using one of two options, "Basic Item Scoring" or "Advanced Item Scoring". These are described in Meyer (2014), and, in May 2017, were also described at this page. I have another paper which discusses item scoring in a bit more detail.

The item analysis output in jMetrik is quite basic, rather "flat" and void of colours. The help documentation which comes with the system is also very basic and, at that, does not cover all of the program's options – to get complete help, thorough explanations of jMetrik output and options, a recommendation is to purchase the book (Meyer, 2014)[9].

[9] The book is very good, but it does not cover new IRT options introduced in the late-2016 version.

Another rather basic characteristic of jMetrik relates to how errors are reported when program options are activated. Messages are written to a log file and can be very cryptic. Users are directed to write to the jMetrik support desk for assistance.

Despite these limitations (*as I see them*), in terms of my classes, jMetrik's IRT support is appealing. Other positive factors: jMetrik is free and easy to install. It runs on Macintosh as well as Windows.

To be noted: I have been unable to find a change log for jMetrik. After I had completed an earlier version of this paper, and later returned to use the program again, a notice was displayed to tell me that a new version was available for download from the jMetrik website. However there wasn't any information about the changes to be found in the updated version – the website does not appear to provide access to a summary of program changes over time.

Applying SAS University

The tables below were obtained by using SAS Program 5.11 and Program 7.6 as described in Cody-Smith (2014). I have again used the 56-item biomedical test first mentioned above.

Here I have included results for just the first 15 items in the test; the alpha value of 0.79805, however, was computed using all 56 items. (All labels below are as made by SAS.)

Item Statistics

	Choices							Quartile			
	A	B	C	D	E			1	2	3	4
	%	%	%	%	%	Diff.	Corr.	Prop. Correct	Prop. Correct	Prop. Correct	Prop. Correct
# Key											
1 D	1	1	.	62	36	62%	0.49	22.6%	68.6%	67.5%	87.1%
2 C	1	5	82	1	10	82%	0.30	63.3%	71.4%	97.5%	93.5%
3 E	.	5	1	.	93	93%	0.32	80.6%	94.3%	97.5%	100%
4 B	17	46	18	13	7	46%	0.43	16.1%	35.3%	52.5%	80.6%
5 C	6	7	85	1	1	85%	0.35	64.5%	91.4%	90.0%	93.5%
6 B	1	96	3	.	.	96%	0.20	87.1%	97.1%	100%	96.8%
7 A	64	21	6	1	8	64%	0.42	32.3%	62.9%	65.0%	93.5%
8 B	7	84	.	4	5	84%	0.07	77.4%	82.9%	82.5%	93.3%
9 C	9	.	89	1	1	89%	0.28	80.6%	80.0%	95.0%	100%
10 E	.	3	4	4	89	89%	0.38	71.0%	91.4%	97.5%	93.5%
11 A	85	5	4	1	6	85%	0.34	67.7%	85.7%	87.5%	96.8%
12 B	20	75	5	.	.	75%	0.27	60.0%	74.3%	77.5%	87.1%
13 D	10	6	4	66	13	66%	0.42	35.5%	62.9%	82.5%	80.6%
14 D	5	.	.	95	.	95%	0.32	86.2%	97.1%	95.0%	100%
15 B	7	71	13	7	2	71%	0.25	51.6%	68.6%	77.5%	83.9%

Listing of Chronbach

TYPE	Alpha
RAWALPHA	0.79805

The first column above, "# Key", gives the item number and the correct answer for the item. The "Choices" columns indicate the percentage of test takers who selected each item option. The

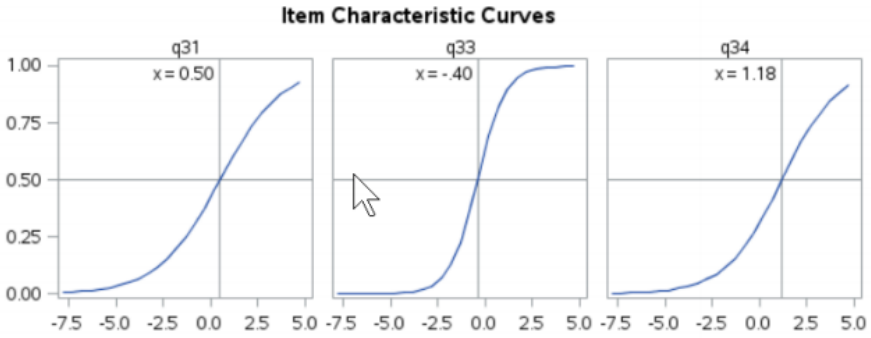
correct answer for the first item was D; 62% of the test takers selected it, this carries over to the column with the "Diff." figure for the item.

The "Corr." column is the correlation between an item and the total test score, generally used in CTT as an indicator of item discrimination. Item 8 has the lowest correlation, 0.07[10].

In this example, the SAS program created four quantiles, or "quartiles", from the total test scores – the first quartile refers to those test takers whose test scores were in the bottom 25% of the distribution of test scores, while the fourth quartile refers to the top 25%. For the first item, the percentage of students able to identify the correct answer climbs from 22.6% in the lowest quartile to 87.1% in the top group[11].

As seen above, SAS has also computed the value of alpha for the 56-item test, 0.79805.

As to the IRT capabilities in SAS, an article by Choi (2017) provides a comprehensive and favourable review of SAS PROC IRT. There's a macro in Lertap 5 which creates data and SAS code for use with this PROC. It works fine, but PROC IRT does not yet have empirical ICCs, and its IRT plots are, in my opinion, a bit on the plain side. An example is shown here:



Data preparation in SAS University: data may be entered directly into a SAS table, or imported using one of a variety of methods, including fetching data directly from an Excel workbook.

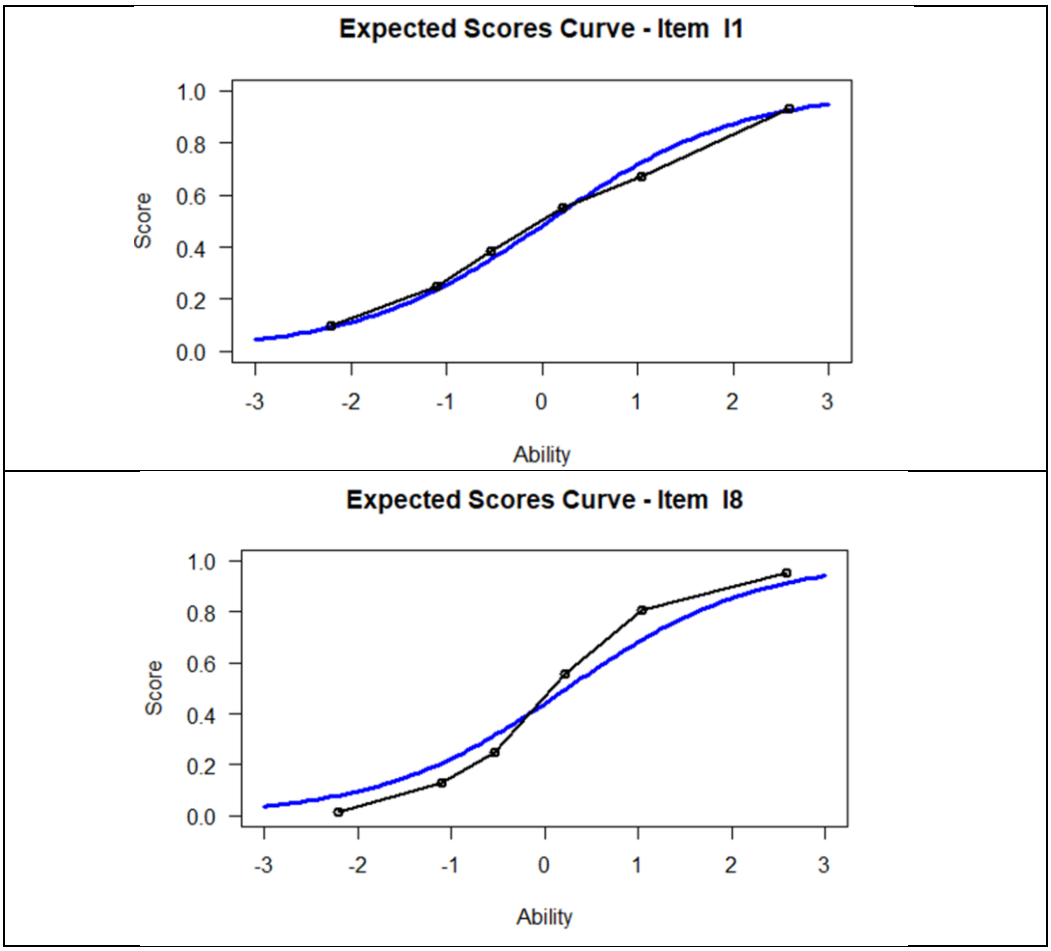
Scoring test items in SAS is exemplified in Chapters 2 and 11 of Cody-Smith (2014); one of the two SAS code samples found below, at the end of the present document, has a macro showing how scoring may be accomplished.

Applying TAM

TAM is an R package. I would not use it, nor any other R package I currently know of, for CTT. If students are using a Windows computer, or running Windows on a Macintosh using a VPM such as Parallels, and if they have Excel, then why wouldn't I, and they, use Lertap for CTT? I have tried a few R packages capable of CTT analyses and found their output to be similar to that shown above for jMetrik: basic, not easy to read, colourless, and well below what Lertap will deliver.

[10] Unfortunately, these Corr. figures are not adjusted for part-whole inflation.
[11] Lertap 5 and Iteman will also routinely output quantile percentages or proportions. See this page.

What appeals to me in TAM, thinking of my classes, are the ICC plots with empirical overlays. They're easy to get, and, in my opinion, they look quite smart:



I have now added a capability to Lertap 5 which makes it a fairly simple matter to invoke TAM and get IRT results, including ICCs like those above. What students will have to do is install R and RStudio. Once these free systems are installed, Lertap 5 is used to pass item responses and suitable code modules to R. TAM results are then but a single click away.

This paper sets out the steps needed to install R, RStudio, and TAM. This Lertap help topic describes how to get Lertap 5 to pass things over to R.

Lertap 5, TAM, jMetrik, or SAS University?

My objective has been to find a free program that will cover both CTT and IRT, and cater to the particular topics I focus on in classes.

For leading students through classical test theory, as I cover it, Lertap 5 is still a winner. Its output is much more extensive than that found in the other programs; it includes numerous graphical summaries of results (even for DIF, differential item functioning, should I have time to get into it). It's strong when it comes to the used of cut scores. It uses flags to highlight weak items, can get all of its quantile plots together in a gestalt display, and has an outstandingly simple way to accomplish

item scoring. Its Rasch routine, added November 2019, may help as an IRT lead-in. Documentation is another Lertap 5 strength, and a big one at that; see this webpage.

Lertap 5 falls short in these areas: it requires Excel; its IRT plots cannot have empirical points superimposed; its coverage of reliability estimates is not as complete as that found in jMetrik. (Lertap 5 is capable of finding split-half and odd-even reliabilities, has a Spearman-Brown calculator, and even a way to get coefficient omega – but all of these require a bit of special effort to get.)

jMetrik's strengths relate to its IRT support, a very big plus as I see it. It runs on any computer. However, from my point of view, except for its ability to output a variety of reliability measures, jMetrik has weak CTT support. The program is very poorly documented at the moment – it's not clear how to use quite a number of jMetrik's options, experimentation is required. Yes, there is a book for it should users care to purchase it, and it is quite a good one (Meyer, 2014) – hopefully it will soon be updated so as to cover the new IRT features added at the end of 2016. I note that jMetrik was updated shortly after I used it for this paper, but I have been unable to find a change log for the program. All I know is that it moved from version 4.0.5 to 4.0.6 – there's nothing I can find that would suggest why I should update my installation of 4.0.5.

Item scoring in jMetrik is rather cumbersome, time consuming, and, in my opinion, on the error-prone side. A work-around is to rely on Lertap 5 for CTT analyses and then, for IRT, port Lertap item scores over to jMetrik where they will be *very* easy to import and "transform".

With regard to SAS University: it's free; look at some of the things it can do:

Statistical methods (SAS/STAT®)

- Extensive statistical capabilities in over 80 procedures:
  - Analysis of Variance
  - Bayesian Analysis
  - Categorical Data Analysis
  - Cluster Analysis
  - Descriptive Statistics
  - Discriminant Analysis
  - Distribution Analysis
  - Exact Inference
  - Finite Mixture Models
  - Group Sequential Design and Analysis
  - Longitudinal Data Analysis
  - Market Research
  - Missing Data Analysis
  - Mixed Models
  - Model Selection
  - Multivariate Analysis
  - Nonlinear Regression
  - Nonparametric Analysis
  - Nonparametric Regression
  - Post Processing
  - Power and Sample Size
  - Predictive Modeling
  - Psychometric Analysis
  - Quantile Regression
  - Regression
  - Robust Regression
  - Spatial Analysis
  - Standardization
  - Structural Equations Models
  - Survey Sampling and Analysis
  - Survival Analysis

This will be tempting for some. But not for me. Installing and configuring SAS University are hurdles many of my students cannot do without help. Running the Cody-Smith programs? The same: students will need real assistance.

Now that Lertap 5 has the ability to prepare data and R code modules, TAM is the package I am very likely to use this year. TAM documentation is not for beginners; even the sample code modules found at the TAM Tutorial site will not be easy for students to follow. But I can get TAM to do what I need, IRT-wise, for classes. I prefer the graphics produced by jMetrik, but TAM is *much* easier to use given the R code modules now available from Lertap 5.

The winner?

Lertap 5, TAM, jMetrik, SAS University? There isn't a single clear winner for me.

Of the four programs, Lertap 5 is way ahead of the others when it comes to the CTT coverage I need. It's weak on the IRT side, but nonetheless is quite usable for my classes – it is good enough to get basic IRT concepts across and is very easy to use – the new Rasch routine, introduced November 2019, will be useful. Students can readily experiment with fitting multiple IRT models with EIRT. The real limitation of Lertap is that I'd like to have better ICC plots, and the ability to have them include empirical results – I have recently found that TAM will do the job for me, and Lertap 5 has now been equipped to pass data and suitable code modules over to TAM. If students have R and RStudio installed, getting TAM results requires just one little mouse click.

Xcalibre? BILOG-MG?

Something I have done in the past, when my classes ran from six to eight weeks instead of just five, was introduce students to two top of the line, off the shelf IRT programs: Xcalibre and BILOG.

There is quite a handy option in Lertap 5 that will create the two files Xcalibre requires as input: a data file, and the item scoring information referred to as the "ICF", the item control file. This option is adequately documented and has been found to be easy enough for students to use; in my experience, three hours is easily sufficient time for students to take a Lertap 5 dataset, apply that "handy option" in Lertap and get Xcalibre to output its comprehensive, colourful report, a report that will superimpose empirical results on the ICCs made for dichotomous items.

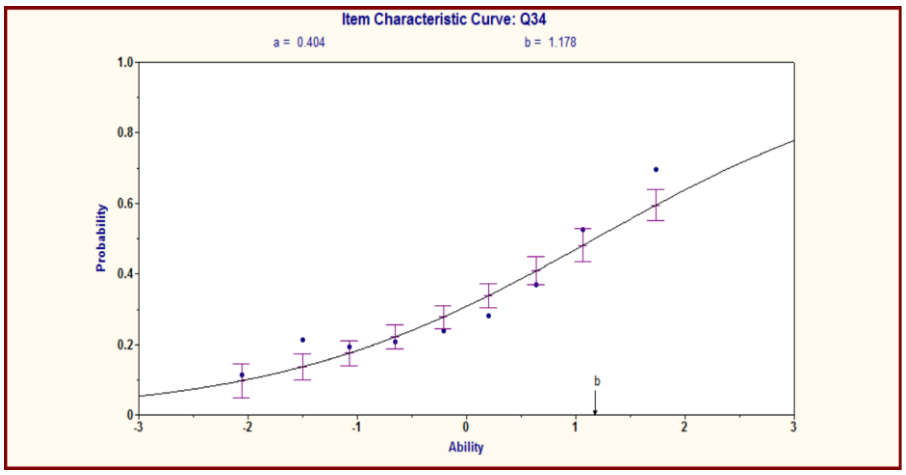
Xcalibre is not free, but its limited "DEMO" version is certainly adequate to get the basics across.

UPDATE July 2019: it is now possible for students to get a free copy of Xcalibre, limited to 50 items and 50 data records, and also for its CTT stablemate, Iteman, limited to 100 items and 100 data records. See the references below. This development might bring Iteman into focus as another possible freebie for my classes were it not for comments found in this paper.

I have also tried BILOG-MG with classes. The advantage of BILOG is that there's a free fully-capable trail version. The trial version remains operational for 15 days – that's the good news. On the downside: it's not a package for the majority of beginning students, even if they're at the postgraduate level and have good computer skills.

BILOG is still an IRT standard after many, many years. Some would likely say it remains the standard. But BILOG's age shows in its interface. It's complex. BILOG is one of those packages which can take over a class – a great deal of time will be devoted to explaining how to use it. In my opinion, it might be easier to get students to use R packages than having them come to grips with using BILOG-MG.

To be noted are the plots produced by the IRT GRAPHICS routines used with BILOG. They're best-of-class in my opinion. Here's an example of a plot easily created after calibrating dichotomous items with BILOG-MG:



What about ConQuest?

The TAM program discussed above comes (largely) from the same colleagues who developed the initial version of the ConQuest program about a decade ago. I used it in a paper written in 2008. In that paper I made use of just the Rasch capabilities in ConQuest. It can do much more than Rasch, and students can get a fully-functional 30-day version for free. Documentation is extensive. It might be worth looking at were it not for the fact that TAM is much faster, free, and can do most of what ConQuest does, plus a bit more (see this webpage).

Examples of SAS programs for IRT and CTT

Two examples of SAS code are found in the boxes below. The first was created by a special Lertap 5 macro, as described here; the second is from Cody-Smith (2014).

```
proc import datafile = '/folders/myfolders/SAS-IScores.xlsx'
  OUT = LertapIScores DBMS = XLSX Replace;
run;
proc print data = LertapIScores (obs=10);
 Title 'Item scores from Lertap 5 (first 10 records)';
run;
* IRT with SAS;
ods graphics on;
 Title 'SAS IRT 2PL model with Lertap IScores';
proc irt data=LertapIScores plots=(scree(unpack) icc) itemfit resfunc=twop;
  var q01-q49;
run;
```

```
*This macro and following program are from Chapter 11 of Cody-Smith 2014;

*Macro Score_Text
 Purpose: To read data from a text file and score the test;
%macro Score_Text(File=, /*Name of the file containing the answer
                           key and the student answers        */
            Dsn=,       /* Name of SAS data set to create    */
            Length_ID=, /*Number of bytes in the ID           */
            Start=,     /*Starting column of student answers */
            Nitems=     /*Number of items on the test        */);
   data &Dsn;
      infile "&File" pad end=Last;
      array Ans[&Nitems] $ 1 Ans1-Ans&Nitems;     ***student Answers;
      array Key[&Nitems] $ 1 Key1-Key&Nitems;     ***Answer Key;
      array Score[&Nitems] 3 Score1-Score&Nitems; ***score array
                                                  1=right,0=wrong;
      retain Key1-Key&Nitems;
      if _n_= 1 then input @&Start (Key1-Key&Nitems)($upcase1.);
      input @1  ID $&Length_ID..
            @&start (Ans1-Ans&Nitems)($upcase1.);
      do Item = 1 to &Nitems;
         Score[Item] = Key[Item] eq Ans[Item];
      end;
      Raw=sum (of Score1-Score&Nitems);
      Percent=100*Raw / &Nitems;
      keep Ans1-Ans&Nitems Key1-Key&Nitems Score1-Score&Nitems ID Raw Percent;
      label ID      = 'Student ID'
            Raw     = 'Raw score'
            Percent = 'Percent score';
   run;

   proc sort data=&Dsn;
      by ID;
   run;

%mend score_text;

*This is the code to use with SAS University Edition; it calls the macro above.

%score_text(file='/folders/myfolders/test scoring/stat_test.txt',
       dsn=score_stat,
       length_id=9,
       Start=11,
       Nitems=56)

*This is the original code as found in Cody-Smith 2014 text;

/* Test the macro
%score_text(file='c:\books\test scoring\stat_test.txt',
       dsn=score_stat,
       length_id=9,
       Start=11,
       Nitems=56)
*/
```

References

Assessment Systems Incorporated: see their links to Iteman and Xcalibre.

Choi, J. (2017). A review of PROC IRT in SAS. *Journal of Educational and Behavioral Statistics (Vol. 42, 2, pp. 195-205).*

Cody, Ron (2015). *An introduction to SAS University Edition.* Cary, NC: SAS Institute Inc.

Cody, R. & **Smith,** J.K. (2014). *Test scoring and analysis using SAS.* Cary, NC: SAS Institute Inc.

Meyer, J.P. (2014). *Applied measurement with jMetrik.* New York, NY: Routledge.

Nelson, L.R. (2000). *Item analysis for tests and surveys using Lertap 5.* Perth, Western Australia: Curtin University of Technology. (Click here to select chapters, and here for the main website.)

Scientific Software International: see their links to IRT programs.

Note

Reader comments are most welcome, as are suggestions for expanding this little paper so as to include another package or two suitable for use in measurement classes. Just drop me an email: l.nelson@curtin.edu.au