Examinee Report with Scaffolding and a Few Numbers: an interactive report

The sample report is hardly the be all and end all of examinee reports. It would probably make any graphics designer cry but it does have the important elements: identification, results, details, and discussion. While I have crammed it on to one 8.5x11 page, it highest and best incarnation would be interactive. The first block is the minimum, which would be enough to satisfy some certifying organizations if not the psychometricians. The remaining information could be retrieved *en masse* for the old pros or element by element to avoid overwhelming more timid data analysts. All (almost[1]) the examinee information needed to create the report[2] can be found in the vector of residuals: $\underline{y}_{ni} = \underline{x}_{ni} - \underline{p}_{ni}$.

Comments in the sample are intended to be illustrative of the type of comments that should be provided, more positive than negative, supportive not critical. They should enforce what is shown in the numbers and charts but should provide insights into things not necessarily obvious, but suggestive of what one should be considering. Real educators in the real situation will without doubt have better language and more insights.

There should also be pop-ups for definitions of terms and explanations of charts and calculations, for those who choose to go that route. I have hinted at the nature of those help functions in the endnotes of the report. The complete Help list should not be what I think needs explaining but should come from the questions and problems of real users.

There are many issues that I have not addressed. Most testing agencies will want to put some promotional or branding information on the screen or sheet. That should never include the contact information for the item writers or psychometricians, but can include the name of the governor or state superintendent for education. I have also omitted any discussion of more important issues like how to present longitudinal data, which should become more valuable, or how to deal with multiple content areas. There's a limit to what will go on one page but that should not be a restriction in the 21st Century. Nor should the use of color graphics.

Discussion for the Non-Faint of Heart

This report was generated for a person taking a rather sloppy computer adaptive test, with 50 multiple choice items, and five item clusters, four with 10 or 11 items and one [*E*] with 8 items. One cluster [*E*] was manipulated to disadvantage this candidate. I call it *rather sloppy* because the items cover a four logit range and I doubt if I would have stopped the CAT with the person sitting right on the decision line. (Administering one more item at the 500-GRit level would have a 50% probability of dropping Ronald below the *Competent* line.) Nonetheless, plus and minus two logits is sufficiently narrow to make it unlikely that any individual responses will be very surprising, i.e., it is difficult to detect anomalies. Or maybe it excludes the possibility of anomalies happening. I'll take the later position.

Reporting for a certification test, all you really need is the first block, the one with no numbers. It answers the question the candidate wants answered; in this case, the way the candidate wanted it answered. The psychometrician's guild requires the second block, the one with the numbers, to give some sense of our confidence in the result. Neither of these blocks is very innovative.

[1] We would also need to be provided the person's logit ability.

[2] The non-examinee information includes the performance level criteria and the keyword descriptors. If we have the logit ability, we can deduce the logit difficulties from the residuals, which frees us even more from fixed forms. Obviously, if we know the difficulties and residuals, we can find the ability.

To be at all practicable, the four paragraphs of text need to be generated by the computer but they aren't complicated enough to require much intelligence, artificial or otherwise. The first paragraph, one sentence long, exists in two forms: the *Congratulations* version and the *Sorry, not good enough* version. Then they need to stick in the name, measure, and level variables and it's good to go.

The first paragraph under '***Comments'*** is based on the *Plausible Range* and *Likelihood of Levels* values to determine the message, depending on whether the candidate was nervously just over a line, annoyingly almost to a line, or comfortably in between.

Paragraph two relies on the *total mean square* (either unweighted or weighted, outfit or infit) to decide how much the *GRit* scale can be used to interpret the test result. In this simulated case, the candidate is almost too well behaved (unweighted mean square = 0.79) so it is completely justifiable to define the person's status by what is below and what is above the person's measure. The chart that shows the *GRit* scale, the keyword descriptors, the person's location and *Plausible Range,* and the item residuals has everything this paragraph is trying to say without worrying about any mean squares.

Paragraph three uses the *Between Cluster Mean Squares* to decide if there are variations worth talking about. [In the real world, the topic labels would be more informative than *ABC* and should be explained in a pop-up help box.] In this case, the cluster mean squares (i.e., [the sum of *y*] squared divided by the sum of *pq*, for the items in the cluster) are 2.7 and 1.9 for clusters *E* and *C*, which are on the margin of surprise.

With a little experience, a person could infer all of the comments from the plots of residuals without referring to any numbers; the numbers' primary value is to organize the charts for people who want to understand the examinee and to distill the data for computers that generate the comments. Because the mean squares are functions of the number of items and distributions of ability, I am disinclined to provide any real guidelines to determine what should be flagged and commented on. Late in life, I have become a proponent of quantile regressions to help establish what is alarming rather than large simulation studies that never seem to quite match reality.

A Very Small Simulation Study

With that disclaimer, the sample candidate that we have been dissecting is a somewhat arbitrarily-chosen examinee number four from a simulation study of a grand total of ten examinees. The data were generated using an ability of 0.0 logits (500 GRits), difficulties uniformly distributed between -2 to +2 logits (318 to 682 GRits), and a disturbance of one logit added to cluster *E*. A disturbance of one logit means those eight items were one logit more difficult for the simulated examinees than for the calibrating sample that produced the logit difficulties in our item bank. The table below has the averages for some basic statistics, averaged over the ten replications.

The total mean squares (*Infit and Outfit*) look fine. The *Cluster mean square* (1.60) and the *mean squares by cluster* begin to suggest a problem, particularly for the *E* cluster (2.03.) This is also shown, necessarily, in the change in p-value (0.14 lower for *E*) and the change in logit difficulty (0.79 harder for *E*.) It would be nice if we had gotten back the one logit disturbance that we put in but that isn't the way things work. Because the residual analysis begins with the person's estimated ability, the residuals have to sum to zero in some metrics, which means if one cluster becomes more difficult, the others, on average, will appear easier. Thus even though we know the disturbance is all in cluster *E*, there are weaker effects bouncing around the others. The statistician has no way of knowing what the real effect is, just that are differences.

	Measure	StDev	Infit	Outfit	Cluster M.S.
Observed	-0.24	0.42	0.97	0.99	1.60
Model	0.00	0.32	1.00	1.00	1.00
	A	B	C	D	E
Number of Items	10	10	11	11	8
M.S. by Cluster	1.05	1.07	1.57	0.71	2.03
P-value change	0.10	0.01	0.11	-0.07	-0.14
Logit Change	-0.42	-0.04	-0.45	0.27	0.79

The most disturbing, or perhaps just annoying, number in the table is the observed mean for the measure. This, for the average of 10 people, is -0.24 logits (478 GRits) when it should have been 0.0 (and 500. While we actually observed a measure of 500 for the examinee we used in the sample report, that didn't happen in general.) We might want to consider leaving Cluster *E* out of the measure to get a better estimate of the person's true location, or we might want to identify the people for whom it is problem and correct the problem rather than avoid it. For a certifying exam, we probably wouldn't consider dropping the cluster, unless it is affecting an identifiable protected class, if we think the content is important and not addressable in another way.

And, as Einstein famously said, "Everything should be explained as simply as possible, and not one bit simpler." However, that is not necessarily my policy.