V. Measuring, Diagnosing, and Perhaps Understanding Objects

Our purpose when undertaking this venture was not to *explain* data or even to build better instruments. It may not seem like it based on the discussion so far but our goal is say something useful about the objects. Although the person and item parameters have equal status in our symmetrical model, they aren't equally important in our minds. The items are the agents of measurement; the person is the object.

5.1 Diagnosis <u>with</u> the Model

Assume we have an object (e.g., person) who behaved appropriately when interacting with an agent (e.g., set of test items) in a way that conforms to the requirements of the Rasch model. (It actually happens quite frequently even when we are talking about high school students and standardized tests.) What can we say about the object based on the measure and the definition of the scale? Fortunately, the model has given us something to measure with.

If we are dealing with a physical lump of material (*Table II.1*) with a score of, say, *150 Shore* units, we know with almost complete certainty that it is harder than quartz and softer than topaz. If we are dealing with an ambient temperature of, say, 45°F, we can say, with less certainty, that Minnesotans drive with the windows down but don't sunbathe in Duluth. This measure might be labelled *Cool* in Minnesota, but it could be called *Warm* in the Antarctic and *Cold* in Florida.

If we are dealing with an elementary school student with a math scale score (*Figure II.3*) of, say, 250, the student almost certainly can add two 3-digit numbers but probably can't solve word problems involving fractions. We aren't sure if the student can interpret calculator displays, identify correct numerical expressions, or solve problems with liquid measures or not. We do know the measure is just above the line separating the *Proficient* Performance Level from the *Basic* but we don't know who defined those levels for whom or what purpose: are we talking *Warm* in Antarctica, Duluth, or Miami?

Diagnosis with the model means marking the measure on the continuum defined by the agents and performance level and noting what is above, below, or near the object's measure. All of these illustrations simply refer the *quantitative measure* label to the *qualitative descriptor* labels contained in the tables and spread along the rulers. Full reporting of results should include both *quantitative* and *qualitative* labels. The quantitative give the precision for doing science; the qualitative give the meaning for communicating and understanding.

5.2 Diagnosis <u>against</u> the Model

Assume we have an object (e.g., person) who behaved inappropriately when interacting with an agent (e.g., set of test items) so that some responses do not conform to the requirements of the Rasch model. (It happens often enough when we are talking about students and standardized tests.) While we may not have a useful ruler to measure with in this case, the model has given us something to compare against.

A strong model gives a frame-of-reference, with some theoretical basis and prior experience, that tells us when to be surprised or when to be alarmed. When quartz scratches diamond, we know something isn't what we thought it was and we need to find additional evidence to determine if it is the quartz, the diamond, or our notion of hardness[1]. When a Duluth denizen has the car windows up while driving to the beach to sunbathe, she may be responding to something other

[1] It's probably not a real diamond.

than cool air; perhaps wind, perhaps hot air. The temperature aspect we have posited is not producing the consequences we imagined or that we observed during the calibration / control phases of development.

If our 250-math student were to miss the simple addition items but pass the word problems with fractions, we might guess, rather unimaginatively, that the misses were due to carelessness and the passes due to luck or other malfeasance and discount both in the student's measure. Or we might devise a small test to see if the results would replicate, or we might ask the expert in the classroom with the student.

There is something operating other than our one-dimensional construct. We would like to know if the student has found a different path through the curriculum, has acquired competencies outside the classroom, or if the item is tapping something we did not anticipate. Maybe our understanding of math proficiency needs rethinking.
There are smart ways to get items wrong and not-smart ways to get items right; all of which implies the student's mind was not doing what we intended.

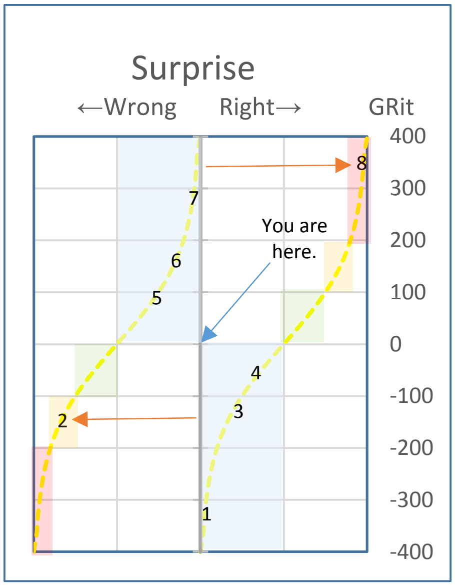
> An item is valid only if the students' minds are doing what we want them to show us they can do. *A. Pollitt*

Against this model, it is straightforward to recognize surprises given a well-calibrated instrument and even easier to generate explanations for it by the carload lot.

5.3 Number-Freed Reporting: One Person with Surprises

The model, the mathematics, and the logits provide the structure that enables us to make sense of the person's responses. It's not necessary, and perhaps counter-productive, to actually show them to the people who are trying to understand the student. Figure V.1 is one approach to reporting the results visually for one person on one subtest. While the underlying math is essential when constructing the chart, focusing the display on the math doesn't help communicate what is important to take away.

Figure V.1: GRits versus Surprise



The figure is centered on one student on one subtest (the *number sense* scale of an elementary math proficiency test;) the person's proficiency (not shown here) is based on a longer test of which these items are a part. The only numbers shown are in the right column (labeled *GRits*) and are the distance from the person's location (zero on the chart) to each item; the picture and its interpretation would not change if these numbers weren't shown. Item 1 is about 320 *GRits* (relatively easy) below the person and item 8 about 360 *GRits* above (relatively hard.)[2]

The item numbers are plotted vertically by the *GRit* distance from the person to the item, and horizontally by our surprise as measured by $y_{vi} = x_{vi} - p_{vi}$. The correct responses are on the right and incorrect on the left. All must follow one of the yellow dashed lines. Items plotted in the blue boxes are not surprises; the probability against the observed outcome, which is one definition of y_{vi}, is less than 0.5. Items in the green (if there were any) are not worrisome because they are not all that unlikely, probability between 0.5 and 0.75. The orange box (item 2) with probability between than 0.75 and 0.9 begins to get our attention; the red boxes (item 8) definitely get noticed, probability greater than 0.9.

On the total score level, the person behind these data is quite well behaved. Four of the eight items were answer correctly and the person was located, based on performance on the total test, nearly in the center of this subtest. Six of the eight items are in the blue (no worries) but two are interesting: one easy item missed and one hard item passed. But for items 2 and 8, this person would be too good to be true and we would be back in the previous section, diagnosing with the model. Because of the two items, we are diagnosing against the model. If one knew the student and had the items available, it might be possible to say something useful about what was going on in the student's mind.

5.4 Number-Full Reporting

When working with a small group of examinees (classroom?) and a reasonable number of items (fixed form?), building what might be described as a *Rasch-enhanced* Sato *Student-Problem* (S-P) chart is a good way to start the diagnosis process:

a. Construct a *person x item* table with a row for each person and a column for each item.
b. Sort the people into increasing order of Scale Score (or raw score if everyone takes the same form) from top to bottom.
c. Sort the items in increasing order of Difficulty (from the Bank) from left to right.
d. Enter the value of $y_{vi} = x_{vi} - p_{vi}$ for all *NxL person x item* interactions into the body of the table.

We *expect*, in a stochastic sense, the table to contain small values everywhere, approaching ±0.5 near a diagonal running from upper left (low ability people and easy items) to lower right (high ability people and hard items), where the person's ability matches the item's difficulty. The values below the diagonal should be positive because the person exceeds the item; and negative above the diagonal. In theory, we expect this; in practice, we know better. All this assumes all students and items will behave the way they should. The table *V.1* is based on simulated data that conform to the model's requirements and so gives a suggestion of what the chart can look like even when there is no problem.

[2] The chart uses the distance from the person to the item $d_i - b_v$ so that the hard items are at the top. *GRits* are calculated as 91 times the logit, which is 100 / ln 3, so that a difference of 100 means odds of 3^1 to 1 (or a probability of 0.75), a difference of 200 means odds of 3^2 to 1 (or a probability of 0.90), a difference of 300 means odds of 3^3 to 1 (or a probability of 0.96), and a difference of 400 means odds of 3^4 to 1 (or a probability of 0.99.) All we have really done is change from base *e* to base *3*, times 100.

Table V.1: Rasch-Enhanced Sato S-P Chart, using y_{vi} = *probability against result*

Person	Item 1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0.2	-0.7	0.5	-0.4	-0.4	-0.3	-0.2	-0.1	-0.1	-0.1	-0.1	**1.0**	0.0	0.0
B	0.2	0.3	0.5	0.5	-0.4	-0.3	-0.2	**0.8**	-0.1	-0.1	-0.1	0.0	0.0	0.0
C	0.1	0.2	0.4	0.4	0.5	0.6	-0.3	-0.2	**0.8**	-0.1	-0.1	-0.1	0.0	0.0
D	0.1	0.2	0.4	0.4	-0.5	-0.4	-0.3	-0.2	-0.2	**0.9**	-0.1	-0.1	0.0	0.0
E	**-0.9**	0.2	0.4	-0.6	-0.5	-0.4	0.7	-0.2	-0.2	-0.1	-0.1	-0.1	0.0	0.0
F	0.1	**-0.8**	-0.7	0.4	-0.5	0.5	-0.3	-0.2	-0.2	-0.1	-0.1	-0.1	0.0	0.0
G	0.1	0.2	-0.7	0.4	0.5	-0.5	0.6	-0.3	-0.2	-0.1	-0.1	**0.9**	0.0	0.0
H	0.1	0.2	0.3	-0.6	0.4	-0.5	0.6	-0.3	-0.2	-0.1	-0.1	-0.1	-0.1	0.0
I	0.1	0.2	0.3	-0.6	0.4	-0.5	-0.4	-0.3	-0.2	-0.1	-0.1	-0.1	-0.1	0.0
J	0.1	0.2	-0.7	0.3	0.4	-0.5	-0.4	-0.3	-0.2	-0.2	-0.1	**0.9**	-0.1	0.0
K	0.1	**-0.9**	0.3	-0.7	-0.6	0.4	-0.4	0.7	0.7	-0.2	-0.1	**0.9**	-0.1	0.0
L	0.1	0.1	-0.7	-0.7	0.4	0.4	-0.4	-0.3	-0.3	-0.2	-0.1	-0.1	-0.1	0.0
M	0.1	0.1	0.2	0.3	0.3	-0.6	-0.5	-0.4	0.7	-0.2	-0.1	-0.1	-0.1	0.0
N	0.1	0.1	0.2	0.3	-0.7	0.4	0.5	0.6	-0.3	-0.2	-0.2	-0.1	-0.1	0.0
O	0.1	0.1	**-0.8**	0.3	0.3	-0.6	0.5	-0.4	-0.3	-0.2	**0.8**	-0.1	-0.1	0.0
P	0.1	0.1	**-0.8**	0.3	-0.7	-0.6	0.5	0.6	-0.3	-0.2	-0.2	-0.1	-0.1	0.0
Q	0.1	0.1	**-0.8**	0.2	0.3	0.4	-0.5	-0.4	0.7	-0.2	-0.2	-0.1	-0.1	-0.1
R	0.1	0.1	0.2	0.2	0.3	0.3	0.5	0.6	0.6	**0.8**	-0.2	-0.1	-0.1	-0.1
S	0.1	0.1	0.2	**-0.8**	-0.7	0.3	-0.5	0.6	-0.4	-0.2	-0.2	-0.1	-0.1	-0.1
T	0.1	0.1	0.2	0.2	0.3	0.3	0.5	0.6	-0.4	-0.2	-0.2	-0.1	-0.1	-0.1
U	0.1	0.1	0.2	0.2	-0.7	-0.7	-0.6	0.6	-0.4	-0.2	**0.8**	-0.1	-0.1	-0.1
V	0.1	0.1	0.2	0.2	0.3	0.3	0.4	-0.4	-0.4	-0.3	-0.2	-0.1	-0.1	-0.1
W	0.0	**-0.9**	0.2	0.2	0.3	-0.7	0.4	0.5	-0.4	-0.3	-0.2	**0.9**	-0.1	**0.9**
X	0.0	0.1	0.2	**-0.8**	0.2	0.3	0.4	-0.5	0.6	0.7	-0.2	-0.2	-0.1	-0.1
Y	0.0	0.1	0.2	**-0.8**	0.2	-0.7	-0.6	0.5	-0.4	-0.3	-0.2	**0.8**	-0.1	-0.1
Z	0.0	0.1	0.1	0.2	0.2	0.3	-0.6	-0.5	0.6	-0.3	-0.2	-0.2	-0.1	-0.1
AA	0.0	0.1	0.1	0.2	**-0.8**	0.3	-0.6	-0.5	-0.4	-0.3	-0.2	-0.2	-0.1	-0.1
BB	0.0	**-0.9**	0.1	0.2	**-0.8**	-0.7	0.4	0.5	0.6	-0.3	-0.2	-0.2	-0.1	-0.1
CC	0.0	0.1	0.1	0.2	**-0.8**	0.3	-0.6	-0.5	-0.5	0.7	0.7	-0.2	-0.1	-0.1
DD	0.0	0.1	0.1	0.2	0.2	0.2	-0.7	0.5	-0.5	0.7	-0.3	-0.2	-0.1	-0.1
EE	0.0	0.1	0.1	**-0.8**	0.2	0.2	0.3	0.5	-0.5	0.7	-0.3	-0.2	-0.1	-0.1
FF	0.0	0.1	0.1	0.1	0.2	0.2	-0.7	-0.6	-0.5	0.7	-0.3	-0.2	-0.2	-0.1
GG	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.4	0.5	-0.3	0.7	**0.8**	-0.2	-0.1
HH	0.0	0.1	0.1	0.1	**-0.8**	**-0.8**	-0.7	0.4	0.5	0.6	0.7	**0.8**	-0.2	**0.9**
II	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.4	-0.5	0.6	-0.3	-0.2	-0.2	-0.1
JJ	0.0	0.0	0.1	0.1	0.1	**-0.8**	0.3	0.4	-0.6	-0.4	-0.3	-0.3	**0.8**	-0.1
KK	0.0	**-1.0**	0.1	0.1	0.1	0.2	0.3	-0.6	-0.6	-0.4	0.6	0.7	-0.2	-0.1
LL	0.0	0.0	0.1	0.1	**-0.9**	0.2	0.2	-0.7	-0.6	0.5	-0.4	-0.3	-0.2	-0.1
MM	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	-0.6	0.5	-0.4	-0.3	**0.8**	-0.1
NN	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.3	0.4	-0.5	-0.4	-0.3	-0.3	-0.2
OO	0.0	0.0	0.1	0.1	0.1	0.1	**-0.8**	0.3	0.3	-0.5	0.6	0.7	0.7	-0.2
PP	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.3	0.3	0.4	0.5	-0.4	-0.3	-0.2
QQ	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.4	-0.5	-0.4	-0.3	-0.2
RR	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	-0.7	0.4	0.5	-0.4	0.7	-0.2
SS	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	**-0.8**	0.4	0.4	-0.5	-0.4	-0.2
TT	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	-0.6	0.5	0.6	-0.2
UU	0.0	0.0	0.0	0.0	0.1	0.1	**-0.9**	0.2	0.2	-0.7	-0.6	0.5	0.6	-0.3
VV	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	-0.6	0.5	-0.5	0.7
WW	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.3	-0.7	0.4	-0.5	-0.3

There will always be some exceptions, which are the most interesting elements in the table. A high scorer can over-complicate, get careless, or apply the wrong rule and a low scorer can get lucky and know things we can't explain. The first type of anomaly might be labeled the *Chesterton*[3] effect; the latter the *Slumdog Millionaire*[4] effect.

All we are looking for in the table are the:

a. Big numbers implying surprises, i.e.:
    a) smart ways to get the item wrong, or
    b) not-smart ways to get it right.

b. Rows with lots of big numbers: people who surprised across items.

c. Columns with lots of big numbers: items that surprised across people.

It's up to you to decide what "big" and "lots" mean.

Table *V.1* was filled with the y_{vi}, which in absolute value is the probability against the outcome we observed. A plus or minus one means the result was very unlikely; anything near zero means the result was very likely; anything near ±0.5 means neither right nor wrong could surprise us. This is a natural and comfortable metric that is relatively simple to explain and straightforward to set threshold values qualifying our surprise from *yawn* to *curious* to *alarm*. We have been considering y_{vi}, but we can put any form of the response we choose into the table: x_{vi}, z_{vi}, or z_{vi}^2. All have their uses and limitations: The one-zero responses, x_{vi}, are probably the least useful; either form of the standardized metric, z_{vi} or z_{vi}^2, is trickier to set threshold values but both are useful for building mean squares; the column average of the y_{vi} is the discrepancy between the item's *p-value* for this group and the *p-value* for the calibrating group, and weighted z_{vi} can be aggregated to show how far off the bank difficulties and logit abilities are for the situation.

Because we have sorted the table, in both directions, by logits, we don't really need to show the numbers at all. The big surprises, when they happen, will always be in the upper right or lower left. They are a function of $(b_v – d_i)$ and that distance will always be greatest in those corners. Simply marking out the regions with isobars for varying levels of surprise, based on the distance between person and item, leaves us free to display more useful information, for example, the actual response the person gave and whether it is right or wrong.

We are not restricted to just looking at rows and columns nor to arranging them in simple logit order. If we want to call out the performance of first-born males for whom Danish is spoken at home on non-calculator items, we can do that. Only our imagination, sample size, resources, and data privacy laws limit how we group people and cluster items. Although policy makers are sometimes disinclined to ask questions they don't want to hear the answers to.

[3] Chesterton, G. K., 1922, *The Man Who Knew Too Much*.

[4] Oscar winning film of 2008 about a teen from the slums of Mumbai who is investigated by the police for unexpected right answers on a television quiz show. Probably the best psychometric movie ever made.