One Student per Group

The discussion thus far begs the question of what the right number of ability groups is and there is no one right answer. The proper number is somewhat arbitrary but the choice can affect what the data reveal. Perhaps there should be as many as your data and software will tolerate. The examples above have used eight, which is a common choice. Had we used only four, the curious and rather messy *GOK* pattern in Figure *1f* would have been completely concealed. We have no way of knowing how the pattern and our interpretation might change if 12 groups were used instead. With a fixed form, we could go all the way to one group for each raw score, but that runs into problems with minimum group size for the *goodness-of-fit* purists.

This leads one to ponder what if we go to the absolute limit and use groups of size one[1]. This is more or less where we started with:

55. $y_{vi} = x_{vi} - p_{vi}$, $x_{vi} = 1$ or 0; and $0 < p_{vi} < 1$, and

56. $$z_{vi} = \frac{x_{vi} - p_{vi}}{\sqrt{Var(x_{vi})}} = \frac{y_{vi}}{\sqrt{p_{vi}(1-p_{vi})}}.$$

We can build any number of things out of these basic elements starting with a couple rather global indices of the item's behavior. Historically we started by squaring and averaging the z_{vi} over the examinees:[2]

57. $$T_i = \frac{1}{N}\sum_{v=1}^{N} z_{vi}^2 = \frac{1}{N}\sum_{v=1}^{N} \frac{y_{vi}^2}{p_{vi}(1-p_{vi})}.$$

As suggested earlier, this is the average <u>*Odds against the Observed*</u> outcome because *y* is either -*p* or *1-p*. If the result is "near enough" to one, we have, in language only a statistician could love, *no reason not to believe there is no problem with the item*[3]. In other words, we are willing grudgingly to admit the item into our bank tentatively. If the result is not "near enough" to one, there are some unlikely events in the data. At the atomic level, there are only two kinds of unlikely events:

- able person misses an easy item, or
- unable person passes a hard item.

If either of these happen, the observed *x* is far from the expected *p* and z^2 becomes large quickly (because $x - p$ is large and $p(1 - p)$ is small.) Given the nature of items and examinees, this often isn't enough to make us want to discard the item but it should get our attention.

Wright proposed a *weighted* average as an alternative to muffle some alarm bells.

---

[1] This is something of an end run around the sample size issue but we didn't really believe the chi-square rationale anyway.

[2] To justify my notation, *T is for total*; in contrast, *G is for group*.

[3] *Being a statistician means never having to say you're certain*. I credit this to Robert Lissitz but *everything of importance has been said before by someone who didn't discover it* (Alfred Lord Whitehead.)

58. $$T_i^* = \frac{\sum_{v=1}^{N} w_{vi} z_{vi}^2}{\sum_{v=1}^{N} w_{vi}} = \frac{\sum_{v=1}^{N} y_{vi}^2}{\sum_{v=1}^{N} w_{vi}} = \frac{\sum_{v=1}^{N} (x_{vi} - p_{vi})^2}{\sum_{v=1}^{N} p_{vi}(1-p_{vi})}, \quad \text{where } w_{vi} = p_{vi}(1-p_{vi}).$$

This effectively returns to the *y*-metric, which is less volatile than the *z*-metric, but has been likened to shooting the messenger. There will be fewer false alarms but an increased risk of overlooking a legitimate issue. And it loses its natural interpretation as odds. The distinction between the two diminishes with well targeted tests. Smith and Smith (2003) refer to *T** as the "*weighted mean square*" and, by contrast, to *T* as "*unweighted*." Wright, with his penchant for the well-turned phrase, uses "*Infit"* for *T** and *"Outfit"* for *T* (Linacre, 2014).

Both *T* and *T** are very global indicators of item acceptability. They can sound the alarm but don't help much with the explanation and may miss some interesting details. Understanding what happened means understanding what the item really requires (not just what the item writers thought it requires) and which examinees surprisingly missed or passed it. This brings us full circle to wanting to form groups of people and clusters of items.

Clusters and Groups: L Items, N People, One Array

Our focus is still on dichotomously scored items, i.e., 1 or 0; our data is still a bunch of people responding to a bunch of items; we have been calling any one of the observations x_{vi} for person v and item i. We also have a model that gives us an expected value for x_{vi}, which we have been calling p_{vi}, defined as the probability that the person will pass the item and may be written as $p_{vi} = B_v / (B_v + \Delta_i)$. We can subtract the later from the former to get a residual y_{vi}. Because the observed data are dichotomous, the simple residuals are either $(1 - p_{vi})$ or $-p_{vi}$, which (unsigned) is the *probability* against the observed result. We might as well throw in the standardized residual z_{vi} squared, which is either (Δ_i / B_v) or (B_v / Δ_i)., which is the *odds* against the observed.

With a total of *N* people in our group taking *L* items[4], the observed data or the expected data or the difference between or all of the above can be arranged in arrays with *N* rows and *L* columns. Because *N* times *L* can be a large number, there will also certainly be by chance large values for some of the y_{vi}, but there should be no discernible patterns in the array.

We can (and probably should) ask the computer to discover any structure by running an exploratory technique like Principle Components Analysis (*PCA*) etc., but that leaves the analyst with the task of naming the components the computer discovered, which most of us are remarkably adept at. The analysis is more powerful and more convincing and more difficult if we identify beforehand where the structure might be. This is asking what characteristics of the person and the item might interact to disrupt the measurement. *PCA* is insurance against our lack of imagination and insight.

Table 4-3 is an attempt to show the *N* by *L* array. It also has an example partitioning with six clusters of items (by columns) and 12 groups of examinees (by rows) in a 2 by 3 by 2 design on three factors. Within each cell, there are as many examinees as fit the classification, each responding to each item in the cluster.

[4] We are assuming for the moment a fixed form of *L* items taken by all *N* examinees, although that is rarely the psychometrically optimal strategy.

Table 4-3: N by L Residual Analysis Array

Group G			Person	Cluster C					
				1	2	3	4	5	6
Gender	Ethnic	Ability	n	Item 1			…		L
Female	A	Low	1						
		High							
	B	Low							
		High							
	C	Low	. . .			y_{vi}	$v \in G_{L,C,F}; i \in C_3$		
		High							
Male	A	Low							
		High	. . .						
	B	Low							
		High							
	C	Low							
		High	N						

Assuming we are dealing with the entire universe (i.e., the *N* examinees were used to estimate the item difficulties and the *L* items were used to estimate the person abilities,) the residual $y_{vi} = x_{vi} - p_{vi}$ will sum to zero when added over any row, any column, and of course the entire array. If Specific Objectivity obtains, every sub-array will sum to zero statistically. "*Sum to zero statistically*" means it is not constrained to be zero, unlike the total array, but should be within statistical limits of zero.

We can split the array in half, as we have earlier, by gender. If we add up all the males and add up all the females, each sum should be statistically zero and we have no evidence of a gender effect on the estimation. But because of the way the estimation is done, one of the sums will be positive and one negative; we just hope not too positive or too negative. If the sums are too large, it doesn't exactly mean that the test is biased against one gender and biased for the other; just that there is a difference. The effect could be due, for example, to confounding with some other factor, observed or not.

We can continue partitioning the array until we exhaust our data or ourselves. For example, we could split the gender groups by ethnicity, and again by ability ranges. The items could be split by difficulty, sequence, content. This is a process like the *Analysis of Variance*. We are sorting through the higher order interactions, lower order interactions, and main effects looking for the simplest (and not one bit simpler) explanation of the results. We don't want to find even main effects in the residuals but if they are there, we want to know about it.

Perhaps, *Analysis of Covariance* is a better analogy because we have adjusted for the model parameters as best we can. Comparisons of the residual sums within sub-arrays should all begin with the phrase, *after for ability and difficulty…* This is another way of saying *controlling for local independence*.

Each sub-array average y_{vi} is the change in p-value for that portion of the universe.

The discussion thus far has been primarily on summing and averaging the *y*-residual within sub-arrays. The *y* metric, while having its limitations, is readily understood and communicated. Each sub-array average is the change in p-value for that portion of the universe. The major limitation is that surprising right answers can cancel out surprising wrong answers. This sends us back to

what statisticians have done since Fisher and Snedecor: square, sum, and divide by the error term.

The analysis of Rasch residuals has one great advantage over the typical Analysis of Variance: we know the within cell error $\sigma_{vi}^2 = p_{vi}(1 - p_{vi})$ in theory. More accurately, we only have an estimate of the within-cell error but it comes from the model, not the pooled within-cell mean square. And there is one great disadvantage or at least inconvenience in our context: the within-cell error is not homogeneous.

Rasch residual analysis lets us practice the four basic operations of arithmetic:

- Subtraction: fill the *person-by-item* array with the differences between the observed and the expected.
- Addition: sum within the appropriate sub-array.
- Multiplication: multiply by itself; i.e., square.
- Division: divide by error term; i.e., square root of the sum of *pq*.

Today, you can do these operations on your phone. We will always start with the subtraction but the order of the other three determines what we get so, if you had something else on your mind in junior high, you might review two fundamental laws of binary operators:

- Associative:
- Commutative:

That's about all the math you need to know here; the rest is just fighting through the notation.

Group Mean Square for Group g, Cluster c:

$$59. \qquad G_{gc} = \frac{\left[\sum_{v \in g}\sum_{i \in c} y_{vi}\right]^2}{\sum_{v \in g}\sum_{i \in c} w_{vi}}, \qquad \begin{array}{ll} w_{vi} = p_{vi}(1 - p_{vi}) & \\ w_{vi} = -y_{vi}(1 + y_{vi}) & \text{if } y_{vi} < 0, \text{ or} \\ w_{vi} = y_{vi}(1 - y_{vi}). & \text{if } y_{vi} > 0. \end{array}$$

This dates back to Wright and Panchapekesan (1969) and is logically identical to G_{gi} of expression 62; only the summations have been changed. It would be zero (nearly) if summed over an entire row, column, or the array. It will be large for a cell if the surprises tend to be in the same direction, either too many right or too many wrong. Surprises in both directions will tend to cancel out. It is sensitive to the interaction of a group of people with a cluster of items; it will be large if the group *g* of people experienced cluster *c* of items differently than everyone else. The fundamental question is, Do the global estimates of parameters correctly predict the observed mean p-value of this cluster of items for this group of people?

Unweighted Mean Square (Outfit) for Group g, Cluster c:

$$60. \qquad T_{gc} = \frac{1}{N_{gc}}\sum_{v \in g}\sum_{i \in c} z_{vi}^2 = \frac{1}{N_{gc}}\sum_{v \in g}\sum_{i \in c} \frac{y_{vi}^2}{w_{vi}}.$$

This would be T_i = *Outfit* for an item if summed over one entire column or for a person if summed over one entire row. It includes a within-cell variance component and will be large

when the cell contains surprises in either or both directions. Surprises in opposite directions will not cancel. This form is very sensitive to able examinees missing easy items or unable examinees passing difficult items, but we should try to avoid administering those items in the first place.

Weighted Mean Square (Infit) for Group g, Cluster c:

$$61. \qquad T_{gc}^{*} = \frac{\sum_{v\in g}\sum_{i\in c} w_{vi} z_{vi}^{2}}{\sum_{v\in g}\sum_{i\in c} w_{vi}} = \frac{\sum_{v\in g}\sum_{i\in c} y_{vi}^{2}}{\sum_{v\in g}\sum_{i\in c} w_{vi}}.$$

This is T_i^{*} = *Infit* for an item if summed over one column or T_v^{*} for a person if summed over one row. It includes the variance within a cell and will be large when the cell contains surprises in either or both directions. Surprises in opposite directions will not cancel. This form is only moderately sensitive to able examinees missing very easy items or unable examinees passing very difficult items, but we should try to avoid administering those items in the first place.

Model Control Displays

Any and all of the three indicators G_{gc}, T^{*}_{gc}, and T_{gc} could be displayed as the summary statistic in its group-by-cluster cell of Table 4-3. When presenting the information to educators and feeling especially hospitable, I prefer to show:

$$62. \qquad P_{gc}^{'} = \frac{\sum_{v\in g}\sum_{i\in c} y_{vi}}{N_{gc}}$$

This retains the plus or minus sign and can be interpreted as the difference in the *p-value* that was observed for this group on this cluster compared to the value expected.

If I am feeling less hospitable or presenting to psychometricians, I would show:

$$63. \qquad D_{gc}^{'} = -\frac{\sum_{v\in g}\sum_{i\in c} y_{vi}}{\sum_{v\in g}\sum_{i\in c} w_{vi}}.$$

While P'_{gc} is interpreted as the change in *p-value* for the cell, D'_{gc} is interpreted as the change in *logit item difficulty*. This also retains a plus or minus sign although reversed from P_{gc}. While the interpretation is easy enough, the explanation of the interpretation is more abstruse. The numerator, $\Sigma\Sigma y_{vi} = \Sigma\Sigma x_{vi} - \Sigma\Sigma p_{vi}$, set equal to zero, is the marginal maximum likelihood estimation equation for item difficulty; the denominator is the second derivative of the log likelihood function. Ergo, D'_{gc} is the first correction in the estimation process, using Newton's method, for the logit difficulties needed to "fit" the observed data in the cell.

The minus sign in front comes with the second derivative, implying we have a maximum, not a minimum. More functionally, it makes the change in logit difficulty go in the opposite direction as the p-value. If the p-value goes up, the item cluster is easier than expected and the logit difficulty should go down.

Regardless of which summary indicator is displayed in the cell, any or all could be used to set flags, with colors, fonts, stars, bells, pop-ups, or text messages, to submit the cell for our

consideration. Our work understanding what happened for a cell begins once the computer has our attention.

Moving to a Higher Level

The partitioning of the *N* by *L* examinee-item array used as our example in Table 4-3 gives a total of 72 cells. This is a manageable number to inspect manually, if intelligent, automated flags ensure we don't miss anything interesting. When you wish to focus on particular factors, you can readily sum over the other factors for any of the three indicators. For example, to obtain an indicator for an item cluster across all groups of examinees, compute:

64. $$T_c^* = \sum_g T_{gc}^* .$$

It is probably defensible to treat this as a chi-square statistic with degrees of freedom equal to the number of terms in the summation over *g*. T_{gc}*, T_{gc}, and G_{gc} for any cell should each have about one degree of freedom.

Rather than sum the unwanted factors away (expression 75), it may be preferable, and about as easy, to redefine the partitioning to include the item clusters and a single, all-inclusive examinee group. This reverts to reduced versions of expressions 70, 71, and 72. In most cases with approximately equal cell sizes, the two strategies will lead to the same conclusions if not the same result. If the cells differ dramatically in the numbers of examinees, the small cells will be overly weighted in expression 75.

Build to Break

What experiment could disprove your hypothesis? John Platt

All models I will consider here require a single one-dimensional construct[5], for which all items are, presumed, equally valid and reliable. The obvious, simple, logical strategy to meet this requirement is to construct instruments with very homogeneous items and administer them to carefully screened examinees. Like most obvious, simple, logical solutions, this isn't the answer.

I was once a small part of a study[6] using an instrument for spatial reasoning. The instrument consisted of number series completion problems, for example, 1, 1, 2, 3, 5, … What number is next? Almost everyone in the group agreed that the best strategy is to keep taking successive differences until the rule became obvious and then apply it to find the next in the series. A small group of examinees, who were the highest scoring, thought that was ridiculous and that you did it by counting down the series and listening for the right number to come around. This high scoring block was the drummers and music teachers in the group. Because the homogeneous items were so strictly one-dimensional, they could not discriminate between spatial reasoning proficiency and rhythm.

The best protection against extraneous traits in the examinees is, don't put all your item eggs in one basket. Spatial reasoning is surely more than just completing number sequences. Find another consequence of being high or low on the trait and devise items that reveal behaviors determined by that consequence. Instruments should be constructed with multiple approaches to revealing the aspect so that we can be certain that high scores truly reflect high ability on the

[5] Not everyone inhabits the same world I do. See Wilson et al (2003).
[6] Everyone in this activity was a graduate student in measurement.

construct and not proficiency as a percussionist. Or maybe you don't have a spatial reasoning test but an admissions test for Pete Best's School for Drummers.

There is a corollary for the examinees that is almost too obvious to mention. When developing items for use with a general population, don't limit your try-outs to first-born, 11-year-old, sixth-grade males of European ancestry. *Person-freed* item calibration doesn't mean we are free to design bad studies.

Cleansing items, censoring people, getting to the bottom of things

At this point, our objective is to identify problematic items, fix them if we can, dispose of them if we can't. Understanding people is secondary; that comes later. It is tempting as innocents abroad in measurement to compute a global version of an unweighted mean square, like

$$65. \quad T = \frac{1}{NL}\sum_{v=1}^{N}\sum_{i=1}^{L} z_{vi}^2 ,$$

use the *F*-distribution, with degrees of freedom of ($N \times L$) and the largest number your software will allow, to establish a reasonable significance level, and, if not significant enough, blithely proceed on our way believing everything is fine.

That's going a little too global a little too fast for my taste. Let's begin at the level of the items with the item version of the unweighted mean square (a.k.a., *Outfit*) to establish the significance level for each item.

$$66. \quad T_i = \frac{1}{N}\sum_{v=1}^{N} z_{vi}^2 , \quad \text{with degrees of freedom } (N\text{-}1, \infty)^7$$

We are now at the stage of sorting the items into four buckets: those we like, those we will live with, those we can salvage, and those we are embarrassed we ever considered in the first place. The criteria for defining the buckets are arbitrary, analogous to measuring with a piece of elastic and cutting with a sharp knife. They usually depend more on resources, economics, schedules, and politics than science and statistics.

I tend to use the unweighted mean square to set the significance level[8], but want to see everything before deciding what the buckets mean, which items are salvageable and which are embarrassing. *Everything* in this sentence means unweighted T_i, weighted T^*_i, several group G_i mean squares, and the changes in p-value P'_i and logit difficulty D'_i. by group and cluster. Most are highly, but not perfectly correlated and perhaps not linearly[9]. An item's global statistics (those with just the *i* subscript) are useful for flagging items, but any diagnosis and understanding looking at more finer detail, i.e., things that require two or more subscripts.

---

[7] The degrees of freedom for the denominator of the F-ratio are infinitely large because the model gave the error mean square, i.e., $s^2 = \Sigma p(1\text{-}p)$. Typically, the numbers are so big it doesn't matter if we use *N* or *N-1* and ∞, 10^{10}, 10^5 or something in between. Whether this is the exact distribution or not, it is place to start.

[8] This may not be all that surprising; considering experiences in my formative years (Mead, 1976), although I didn't say *unweighted* or *Outfit* because the *weighted mean square* and *Infit* weren't even a gleam in Ben's eye at that time.

[9] The point biserial correlation can be thrown into this mix if you are that old school. I also include *foil analyses*, with the count, point biserial, and mean logit difficulty for each foil. It is always useful to know which distracters were popular and with whom.

Beginning with the items with the most disturbing values, examine them from every angle and keep reviewing, revising, and rejecting until you are chasing noise, the item pool is depleted, or the client runs out of money.

The choice of a favorite strategy may be based on unpleasant experiences from your youth or it may be based on deeply-held philosophical principles. My deeply-held philosophical principles are:

- *Some notion of what might go wrong is more important than how you do the arithmetic*.
- *Contemplation of raw observations with an empty mind, even when it is possible, is often hardly more beneficial than not studying them at all.* (Martin Wilk)
- *If your data have something to tell you, your statistics won't stop them.* (G. Box)
- *No single fit statistic is either necessary or sufficient.* (David Andrich).