A question can only be valid if the students' minds are doing the things we want them to show us they can do. Alastair Pollitt

Able people should pass easy items; unable people should fail difficult ones. Everything else is up for grabs.

One can liken progress along a latent trait to navigating a river; we can treat it as a straight line but the pilot had best remember sandbars and meanders.

However one validates the items, with a plethora of sliced and diced matrices, between group analyses based on gender, ethnicity, ses, age, instruction, etc., followed by enough editing, tweaking, revising, and discarding to ensure a perfectly functioning item bank and to placate any Technical Advisory Committee, there is no guarantee that the next kid to sit down in front of the computer won't bring something completely unanticipated to the process. After the items have all been "validated," we still must validate the measure for every new examinee.

The residual analyses that we are working our way toward is a natural approach to validating any item and any person. But we should know what we are looking for before we worry about the arithmetic. First, we need to make sure we haven't done something wrong, like score the responses against the wrong key. This is no different than checking for miskeyed items; the examinee would have both surprising misses and surprising passes in the response string. Having gotten past that issue, we can then check for differences by item type, content, sequence to just note the easy ones. Then depending on what we discover, we proceed with doing the science either with the results of the measurement process or with the anomalies from the measurement process.

Model Control ala Panchapekesan

The exposition that follows, which should seem more familiar to some than the previous section, is based on the estimation method suggested by Nargis Panchapakesan (Wright & Panchapakesan, 1969) when computers were new, slow, and expensive. The method is now generally known as *marginal maximum likelihood*, although Panchapekesan (from the Chicago school) referred to it as *unconditional maximum likelihood* (aka, *UCON*) to contrast it with the philosophically more satisfying, mathematically more elegant, and computationally more demanding fully conditional and likelihood ratio methods of the European school of Fischer and Andersen.

One Student, One Item

At the most basic level, measurement begins with one examinee responding to one item and creating a response x_{vi} that can be represented by *1* or *0*. The model provides a probability p_{vi} that the process will produce a *1*. This gives us an observation x_{vi} and an expectation p_{vi} for that observation. The observation and its expectation will never be equal: x_{vi} is always either 1 or 0; p_{vi} is always between 0 and 1, never equal to either.

It's a small step to ask just how different they are.

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

If x_{vi} is one (i.e., the response is right), y_{vi} will be positive, above expectation, and equal to:

56. $y_{vi} = 1 - p_{vi}$, $x_{vi} = 1$.

It will be close to *0* if the person passed an easy item and close to *+1* if the person passed a difficult item. It will be *0.5* if the person's ability matches the item difficulty exactly; in which case, the psychometrician really doesn't care if the response is right or wrong although teachers and parents may feel differently.

If x_{vi} is zero (i.e., the response is wrong), y_{vi} will be negative, below expectation, and equal to:

57. $y_{vi} = -p_{vi}$, $x_{vi} = 0$.

The |*y*-residual| is the *probability* against the outcome that we observed.

It will be close to *0* if the person missed a difficult item and close to *–1* if the person missed an easy item. It will be *–0.5* if the person ability matches the item difficulty.

For any person-item interaction, there are only two possible values for *y*; either $-p_{vi}$ or $(1-p_{vi})$. Stripped of its sign, the *y*-residual is the probability against the outcome that we observed. It is large (i.e., approaching ±1) when the outcome is a surprise; small when the outcome is pretty much what we expected.

We expect high ability people to pass low difficulty items and low ability people to fail high difficulty items. Easy and hard are relative to the person. The magnitude of the residual is strictly a function of the distance between the person and item. The range of possibilities is shown as the solid blue lines in Figure 4-1. The line above zero represents correct responses; the one below represents incorrect responses. If the person is well below the item (on the left,) we are surprised by right responses; if well above (on the right,) we are surprised by wrong responses.

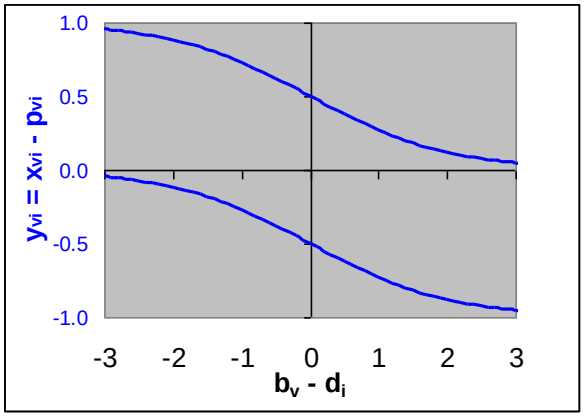


Figure 4-1*: Probability Metric Residuals determined by (Ability – Difficulty)*

The *p*-metric *y*-residual allows immediate probability statements about each response with no additional assumptions. A residual of, for example, +0.9 means the student had 90% probability of missing an item that was passed. A residual of –0.9 means the student had a 90% probability of passing an item that was missed. While this outcome may seem unusual, it should happen 10% of the time. This is a very natural, easily understood metric, especially for small group analyses where the person doing the understanding knows the examinees and items.

The *y*-residual plateaus at plus or minus one as the person moves away from the item, reflecting a probability near one for the response we didn't get (or near zero for the one we did get.) This version of the residuals can be very useful for communicating to some audiences but can be

clumsy for other purposes. To get beyond, we do what statisticians have always done: we standardize.

58. $$z_{vi} = \frac{y_{vi}}{\sqrt{Var(y_{vi})}} = \frac{x_{vi} - p_{vi}}{\sqrt{p_{vi}(1-p_{vi})}}, \qquad x_{vi} = 1 \text{ or } 0; \text{ and } 0 < p_{vi} < 1.$$

Although we have labeled the new version as *z*, it has little to do with a standard normal distribution other than its appearance. It starts with a dichotomy and is a straightforward transformation of the probability-based *y*-residual. It may be interpreted as odds (rather the square root of the odds) against the outcome we observed. For the two possible values of x_{vi},

> Switching from the *y*-residual to the z-residual is switching from the *probability* against the response to root *odds* against.

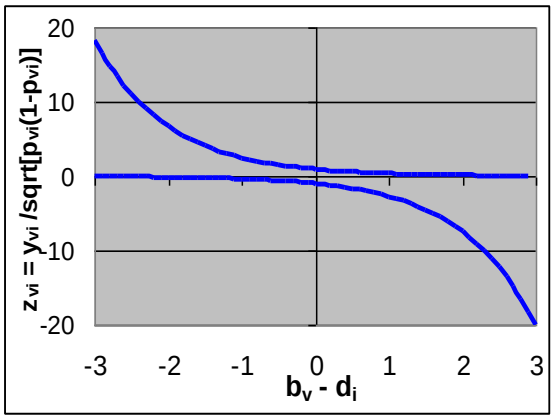
59. $$z_{vi} = \sqrt{\frac{1-p_{vi}}{p_{vi}}} = \sqrt{\frac{D_i}{B_v}}, \qquad x_{vi} = 1;\ D_i = e^{d_i};\ \text{and}\ B_v = e^{b_v}.$$

60. $$z_{vi} = -\sqrt{\frac{p_{vi}}{1-p_{vi}}} = -\sqrt{\frac{B_v}{D_i}}, \qquad x_{vi} = 0.$$

Expression 59 will be large when the item is hard for the person and expression 60 large when the item is easy, compared to the person's ability.

The *z*-residual explodes exponentially (Figure 4-2) in the extremes reflecting very long odds against the response we got, suggesting we should be very worried, or at least a little surprised depending on the stakes.

Figure 4-2: Odds Metric Residuals determined by (Ability – Difficulty)



We can entertain ourselves for a long time studying the individual person-item residuals but unless we talk with the students about what was happening in their minds when confronting the specific item, we aren't apt to learn much. To begin the process of validating an item, we need to look for more generalized patterns. Once again we start from the notion of Specific Objectivity. If the parameter estimates are truly freed from the influence of the abilities of the examinees used in the process, then it shouldn't matter if the estimation is done with high performing examinees or low performing examinees, males or females, third graders or fourth graders, the class of 2020 or the class of 2021, blue states or red states, and on and on.

One Group, One Item

Just as for any individual where we have an observed response x_{vi} and an expected response p_{vi}, for any group g, we have an observed response $O_{gi} = \sum_{v \in g} x_{vi}$ and an expected response $E_{gi} = \sum_{v \in g} p_{vi}$ to the item. Once again it is handy to subtract the expected from the observed, and, even better, subtract the average expected from the average observed:

61. $$\bar{y}_i = \frac{O_{gi}}{N_g} - \frac{E_{gi}}{N_g} = \frac{\sum_{v \in g} x_{vi} - \sum_{v \in g} p_{vi}}{N_g} = \frac{\sum_{v \in g} y_{vi}}{N_g}.$$

This is nothing more nor less than the change in the item's p-value for group g compared to the value predicted for the group based on the total group and total test information. It has a lot in common with the point biserial correlation from true score theory but, with this, we know what to expect, or hope for. We expected a total of $E_{gi} = \sum_{v \in g} p_{vi}$ of the group to get the right and we observed $O_{gi} = \sum_{v \in g} x_{vi}$ correct responses. We hope the difference will be statistically zero, which it will if specific objectivity holds. If the item was more difficult than expected for the group, the difference will be negative; if easier, then positive. Like all fit statistics, and unlike Rasch parameter estimates, this difference is specific to the group tested and can never be *sample-freed* and we wouldn't want it to be. We are trying to validate the model in this context, not calibrate it.

> Tests of Significance are things to do while one tries to think of something sensible. *Martin Wilk*

To give the statisticians something to do while the educators make sense of the discrepancies in *p*-values, we can produce a χ^2*-goodness-of-fit* test out of the same data.

Table 4-1: Two-by-Two Table for Examinee Group g, Item i

Response	Group *g*	
	Observed	Expected
0 = incorrect	$O_{gi0} = \sum_{v \in g} (1 - x_{vi})$	$E_{vi0} = \sum_{v \in g} (1 - p_{vi})$
1 = correct	$O_{gi1} = \sum_{v \in g} x_{vi}$	$E_{vi1} = \sum_{v \in g} p_{vi}$

If we add these up in the usual manner and do a little algebra, we have a statistic that will be small in a statistical way if specific objectivity holds for a group g of examinees:

62. $$G_{gi} = \frac{(O_{g0} - E_{g0})^2}{E_{g0}} + \frac{(O_{g1} - E_{g1})^2}{E_{g1}} = \frac{\left[\sum_{v \in g} x_{vi} - \sum_{v \in g} p_{vi}\right]^2}{\sum_{v \in g} [p_{vi}(1 - p_{vi})]} = \frac{\left[\sum_{v \in g} y\right]^2}{\sum_{v \in g} Var(y)}.$$

If group g had no role in the estimation of the item difficulties and if things work the way we would like, G_{gi} should be a chi-squared statistic with a single degree of freedom. If g is included as part of the larger group used for estimation, the degrees of freedom are something less than

one. G_{gi} will be zero, or close to it[1], if it were computed for the entire group of examinees that was used to estimate the item difficulties. The numerator of the term second from the right, set equal to zero, is the UCON estimation equation for difficulty.

The method for controlling the model is a natural consequence of Specific Objectivity. We have posited a strong model, imagined consequences that follow from the model, and are now exploring our data to see if the consequences do in fact follow. The estimates of item difficulty should apply equally well to any group of examinees we might wish to examine, but we need to make sure.

While we are striving for an estimate of the difficulty that applies to anyone, p_{vi} is computed specifically for individual *v*. We are not relying on an average, one-size-fits-all, population-specific p-value for the item. Every x_{vi} has its own unique control[2] in the form of the p_{vi}. This point will become increasingly significant as we try to compare two or more groups, which may not have the same abilities or distributions of abilities.

Multiple Groups, One Item

The most obvious, and most convenient, criterion for defining groups of examinees is by ability. With two subgroups *g*, e.g., the top half and bottom half of the examinees, the individual chi-squares can be summed to a more general statistic:

$$63. \qquad G_i = \sum_{g=bottom}^{top} G_{gi} \; .$$

If the two subgroups constitute the entire estimation group, we are again in the position of hoping for a chi-squared statistic with a single degree of freedom, i.e., the degrees of freedom are one less than the number of groups. The observed responses, summarized in Table 4-2, are counts of examinees in each cell; the expected responses take the same form as the observed with p_{vi} replacing the x_{vi} and are computed using the estimated difficulties and abilities determined by the entire sample or taken from an existing, calibrated item bank.

Table 4-2: Two-by-Two Table for Two Examinee Groups, Item i

Response	Bottom Group	Top Group
0 = incorrect	$O_{Bi0} = \sum_{v \in B} (1 - x_{vi})$	$O_{Ti0} = \sum_{v \in T} (1 - x_{vi})$
1 = correct	$O_{Bi1} = \sum_{v \in B} x_{vi}$	$O_{Ti1} = \sum_{v \in T} x_{vi}$

If you've been paying attention, you will have noted that we don't really need the separate lines for correct/incorrect. That was just included to make it a 2-by-2 contingency table, which may look familiar to old goodness-of-fit testers. All that is required for the analysis are Σx_{vi}, Σp_{vi}, and $\Sigma p_{vi}(1-p_{vi})$ for each group (see expression 62).

---

[1] There are some technical reasons that I am not ready to discuss why Σx may not be identically equal to Σp for the total estimation group.

[2] Actually the p_{vi} are not all that unique to the person. Every person who took the same set of items and got the same raw score will have the same *p's*. Indexing by score rather than by person could lead to more efficient computer code. The argument for efficiency is becoming less compelling and the argument against fixed forms more compelling.

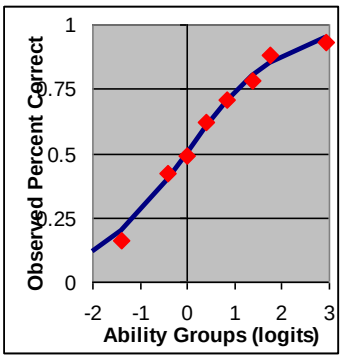
The aggregate G_i is often divided by the degrees of freedom to give something often called a "mean square" for the item with a null expectation of about one[3]. This in turn can be subjected to a cube root transformation (Linacre, 2012) to get a distribution symmetric around zero in the null case.

We are not restricted to two groups; the argument can extend to any number of ability groupings simply by redefining the range of the summation in expression 63. Most computer programs, going back to the 1960s, that included this statistic assume fixed forms and lump together adjacent raw scores to, first, get a reasonable number of examinees in each cluster, and, second, to deal with the physical constraints of computing technology circa 1970. The convenience of clustering will come at the price of a decrease in resolution, i.e., points in a cluster may cancel out if one is above expectation and one below.

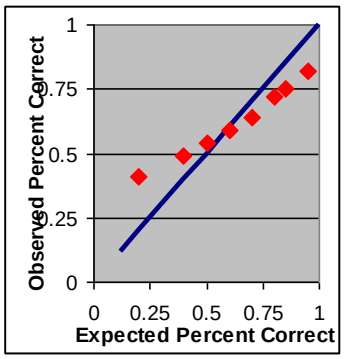
The ability group analysis directly confronts the central tenet of Specific Objectivity: *the estimates of item difficulty must be independent of the abilities and ability distribution of the examinees*. If G_i is large for an item, then there is a relationship between ability and difficulty that we didn't want. This is how we identify items that aren't uniformly valid and reliable. More visually, it means the item characteristic curve (ICC) is not the right shape, perhaps too flat, perhaps too steep, perhaps the wrong asymptotes, perhaps not smooth enough.

The shape of the ICC is an easy thing to examine; not so easy to diagnosis. It is generally confounded with any number of other factors that may be more explanatory. Figures 1a-f present some sample ICCs and speculations that might be the beginnings of further investigations. The solid line represents the theoretical expectation Σp_v. The diamond plotting symbols are observed values Σx_v for examinees in ability groupings. Figure 1a uses logit ability for the horizontal axis; the others use the expected percent correct for the group. It makes remarkably little difference in the utility of the picture, but working with expected percent correct here makes the computing slightly easier and some audiences slightly less hostile.

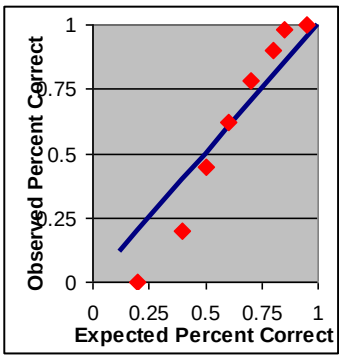
Figure 1a: Near Perfect



1b: Too Flat

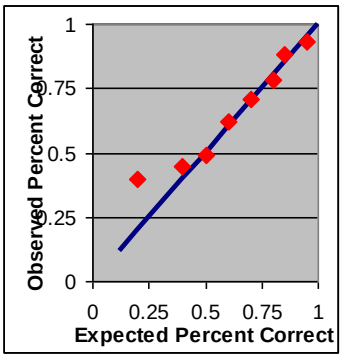


1c: Too Steep

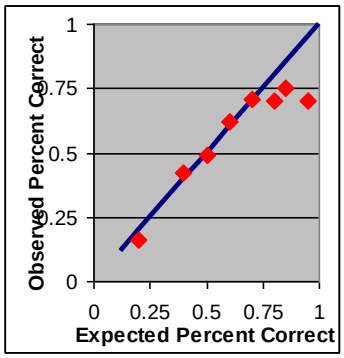


[3] Given the nature of the data that we typically see, I don't lose too much sleep worrying about the exact distribution for the null case.

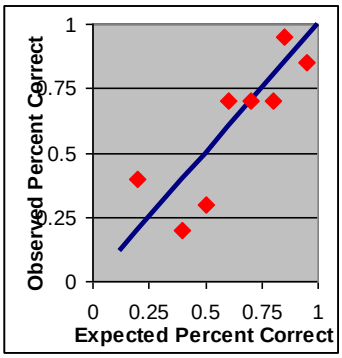
1d: Lower Asymptote



1e: Upper Asymptote



1f: GOK



The first, Figure *1a*, is what we want all our ICCs to look like. The ICC in Figure *1b* is too flat, which may be an item that is influenced by other attributes of the students or things from outside of school unrelated to ability. In *1c*, it is too steep, which may be an item that is strongly affected by instruction that only the better students have received or have absorbed. The item in *1d* may have a low ability way to get it right, or it may be a topic best known or only taught to a low scoring group; *1e* may have a high ability way to get it wrong, such as overthinking a simple question or applying the wrong principle. I can't explain figure *1f*, which can only be described as *GOK* ("God only knows") or your guess is as good as mine. These are only speculations; they are certainly not the only possibilities and may have nothing to do with what actually happened.

Some Rasch skeptics may suggest here that we should have included more parameters in our model to account for some or all of these disturbances. One Rasch response to additional parameters is that validity would be compromised for a possible gain in reliability and comes at the expense of the sufficient statistics, would introduce estimation complications, doesn't avoid the confounding, and may mask the real problem. Another response (Andrich, 2014) is that we began with a model that permits measurement; a conflict between the model and the data indicates a problem with the data, which must be investigated. Abandoning the model would mean giving up on the possibility of achieving measurement and the possibility of a better understanding of the world. Simply fitting the data is not progress.

In measurement, *validity* trumps *reliability*.

Including discrimination in your measurement model is like putting the bathwater to bed with the baby. It has its place but that isn't it. These are interactions, in the analysis of variance sense, which make it problematic to attempt interpreting the main effects (i.e., person ability and item difficulty.) The interactions do not deserve parameters and probably represent, and possibly mask, violations of unidimensionality. Leaving them out of the model does not mean we blithely deny their existence. It does mean we absolutely must consider them in the subsequent and obligatory analyses for model control. It's when they are put in the model that they tend to escape further scrutiny.

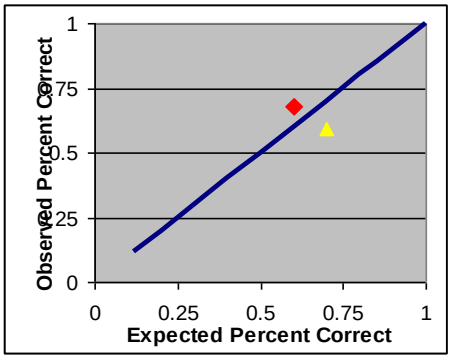
Specific Objectivity doesn't stop with asserting independence from the distribution of ability; it also asserts independence from any other attribute of the examinees. We need to define groups using any and all factors we could imagine making trouble down the line. Gender, ethnicity, region, age, grade, language proficiency, economic status of the examinees are common and obvious choices but hardly exhaust the possibilities for threatening our measurements. The groups will be defined differently but the arithmetic will be identical to that we have just described.

A gender analysis, with two groups, is a restatement of expression 59:

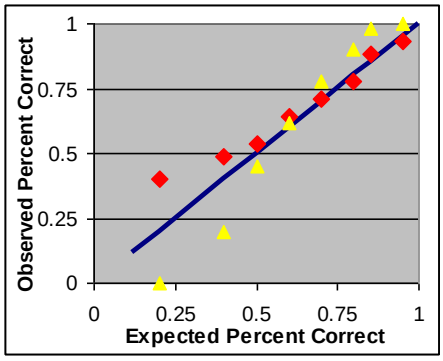
64. $$G_i = \sum_{g=Males}^{Females} G_{gi} \ .$$

This is a check for a main effect of gender on the *difficulty* estimate[4]. It is sensitive to a simple shift in difficulty of the item between the two groups, which is a violation of Specific Objectivity. After controlling for group ability, the item was harder for one group (below the diagonal) than the other (above the diagonal.)

Figure 2a: Gender Main Effect



2b: Gender x Ability Interaction



Returning to an earlier point, because each observation x_{vi} has its own control, its own expected value p_{vi}, it does not matter for these purposes if there is a gender difference in overall ability. We are not comparing one group to the other group but each is compared to the diagonal line.[5] All we need to know is that one point in Figure *2a* is (significantly) above the line and one is below.

To translate this into other people's language, in *Differential Item Functioning* (DIF) words, our main effect for Gender is their *Uniform DIF* and our Gender-by-Ability interaction is their *Non-uniform DIF*.

Multiple Groups, Multiple Factors, One Item

To begin to know what's going on, we need to check for interactions involving gender, beyond the simple main effect. The arithmetic is no different than we have been doing; we just need to define groups appropriately. Instead of just male or female, we have males and females, each divided into ability groups. A sample analysis is illustrated in Figure 2b. The aggregate statistic (see expression 60) involves a double summation:

65. $$G_i = \sum_{r=Low}^{High} \sum_{g_r=Males}^{Females} G_{g_r i}$$

[4] This is not a check for a main effect of gender on the ability estimates, which may be important and interesting for later study once we have valid measures, but isn't a violation of our principles.

[5] The gender main effect for ability is reflected in the offset between groups on the horizontal axis, if you are interested in that sort of thing. The thing that is bothering us is that the higher (based on the total test) performing group (yellow, if you are watching this in color) found the specific item more difficult that did the lower performing group.

When validating an item, we would like to make more general statements than describing the idiosyncratic behavior of individual examinees, although teachers might find that information very helpful. When reporting student scores, our focus may be quite different than it is while validating items. For now, we want to know if the item can be allowed into our bank, if it needs modification, or if there are classes of examinees for which it is not appropriate. In theory, you might have different logit difficulty estimates for use with different groups of examinees but you would need to be very certain the aspect being measured actually manifests itself differently for those groups so that we are really talking about the same thing.

The two items described in the box at right, taken from different projects, both failed one of the checks for Specific Objectivity. For the first, it was the *across years* of administration check. The administrators were confident they understood the issue and chose to continue using the item with difficulty estimates determined by whether or not the ad in question was running; in effect, treating it as two different items depending on what was on television at the time.

For the second, the problem was found in the *among gender-ethic group* check. No one had a good explanation why just this and only this subgroup appeared affected. After much debate and many suggested edits, this item was dropped from consideration for operational use.

Middle School Vocabulary Items

*1. What does the word "***maize***" mean in this passage?*

a) Color of plant seeds
b) European explorer
c) Native American
d) Type of corn

The item functioned well until an ad for a non-dairy spread featured a young woman dressed as a Native American saying, "You call it corn; I call it maize." The item was significantly easier that year, and the following year, across all ability levels, than ever before.

2. What does the word "***village***" mean in this passage?

a) City
b) Park
c) School
d) Town

This item was relatively difficult for the low-scoring portion of one gender-ethic group, who picked *a* over *d*; associating the word with the Village People and Greenwich Village in New York City.