VI. Linking and Equating: Getting from A to B

Unleashing the full power of Rasch models means identifying, perhaps conceiving an important aspect, defining a useful construct, and *calibrating* a pool of relevant items that measure it over a meaningful range. So far we have concerned ourselves with processing isolated bunches of items. In this world, linking and equating item sets has been treated as a distinct and unique phase in the process from conception to measurement. With the technology available, this has typically been the most convenient and efficient approach, and may continue to be so.

In the new world, post fixed-form, paper-based instruments, which are more and more passé, building a calibrated pool can be an inherent and natural part of the process and not a separate step. Calibration procedures allow us to combine individual level test data across administrations, perhaps years apart, to check if specific objectivity holds across time or distance. This is just another *between-groups* comparison, which gives more opportunities for control and investigation of the process.

For the moment, we will continue to talk linking and equating the old fashioned way. The terms *link* and *equate* are often used interchangeably, sometimes not. In other texts and contexts, the distinction seems to be that *equated* implies that the tests measure the same construct while *linked* implies the tests are correlated but may not be synonymous. In this sense, measures from *equated* tests are interchangeable; measures from *linked* tests may be useful as validations, as predictors, or as mileposts but they are not interchangeable.

In my world of Rasch measurement, we consider only tests that measure the same construct. The distinction I will make between *link* and *equate*, it is that two forms are:

- *Linked* if they are literally connected through some common, overlapping element, either overlapping items or overlapping examinees or both.
- *Equated* if the (logit) scores are on the same scale so that scores from one form are equivalent to and can be compared legitimately to (logit) scores from the other form.

Linked implies a direct physical connection between the forms; then and only then, they are *equate-able*. *Equated* implies the forms were *linked* and all the necessary arithmetic has been done.

Logit abilities can be estimated from any selection of calibrated items, ideally a unique, carefully tailored set specific for each examinee. An ability estimate from any subset of the calibrated pool can be compared directly to estimates based on any other subsets taken from the same pool. In that sense, all possible forms that have been or ever could be built from the pool are *equated*. All this assumes that the pool conforms to Rasch's requirements; that is, composed of *equally valid and reliable items*.

Because when we started this by estimating the item difficulties, we had one more unknown than we had equations, we needed to impose a convention. The convention we choose, from among many possibilities, was $\Sigma d_i = 0$. Any collection of items calibrated as a group will be centered at zero: a form for pre-school reading readiness will be centered at zero; a form for post-graduate study of Middle English literature will be centered at zero. The *equating* problem is to determine how far it is from the "Recognizing Letters" zero to the "Analyzing Chaucer" zero, which we can only solve if the *linking* has been taken care of.

The Conventional Equating App

Linking and Equating are converting temperatures on the Kelvin scale to temperatures on the Celsius scale; they use the same units but have different zeros[1]. In terms of what points on the scale mean, it doesn't matter which numeric labels we attach; the effect of liquid nitrogen on your skin will be the same whether it's labeled 77ºK or -196ºC. We just need to agree on the scale to communicate with each other. Neither the nitrogen nor your skin much care about the labels.

The steps for equating tests or thermometers are:

1. Pick some anchor points.
2. Observe measures for those points on both the new and old scales.
3. Average each set of measures.
4. Subtract the new average from every point on the new scale.
5. Add the old average to every point on the new scale.

This effectively subtracts out the new and adds in the old.

Anchor Points	*Kelvin*	*Celsius*	*Difference*
boiling point of nitrogen	77	-196	273
freezing point of salt water	255	-18	273
freezing point of pure water	273	0~	273
triple point of water	273	0	273
normal human body temperature	310	37	273
boiling point of water at sea level	373	100	273
Average	**260.17**	**-12.83**	**273**

If we subtract -12.83 and add 260.17, we have placed the Celsius values on the Kelvin scale and done nothing to the Kelvin values. We could have subtracted the difference from the Kelvin values and ended up on the Celsius scale just as well. If we had chosen different anchor points, the means would be different but not the difference. Anything we do to the scale values of the anchor points we must do to every point on the scale. The hard part is remembering when to add and when to subtract. The process is just as simple though perhaps not as tidy in our world of mental measurement. Replace the Kelvin values with the logit difficulties from the previous calibration and the Celsius values with the logit difficulties for the same items from the current calibration[2]. Then the idea is the same and seems too simple to bother talking about.

In principle, two forms, like two thermometers, can be linked with one anchor point (one item or one person). Because it is the same item, any difference in the item's logit difficulty estimate, beyond random error, is due to a difference in the arbitrary origins of the calibrations. With an easy form, centered on zero, the link item could have a positive logit, implying it is harder than the form average. With a difficult form, also centered on zero, the same link item may have a negative logit, implying it is easier than the average. It doesn't affect how hard the item is any

[1] Converting Celsius to Fahrenheit, with different units, is a slightly different discussion.

[2] The terms *current* and *previous* suggest we are equating this year to last year, which is common. We might also equate the fifth grade scale to the fourth grade scale or this year's test to an existing bank. You might also want to change the row and column labels to something more appropriate.

more than using Kelvin or Celsius affects how cold nitrogen is. The goal of the equating exercise is to eliminate the difference due to the convenient but arbitrary convention of centering on zero by adjusting the logit estimates on one form so that the link item has the same numeric value on both.

Assuming our link item has a difficulty estimate of d_A when calibrated with form A and an estimate of d_B when calibrated with form B, then the form B estimate can be made to equal the form A estimate by subtracting d_B and adding d_A.

The adjusted difficulty for the linking item is:

26. $d_B^* = (\mathrm{d_B} + \mathrm{t}) = \mathrm{d_B} + (\mathrm{d_A} - \mathrm{d_B}) = \mathrm{d_A}$.

So far, this is just as trivial as it seems. The two forms are equated by adding $t = d_A - d_B$ to every logit on form B. If we do it to one item on form B, we need to do it to every item on form B to maintain their relative positions. And we can adjust the form B logit abilities in exactly the same way by adding t to each of them without going through the bother of re-computing them from the adjusted difficulties.

In practice, it may not be a good idea to equate through a single item. With several link items, the equating constant t is simply the difference between the means of the link items from the two calibrations.

27. $$t = \bar{d}_{ALink} - \bar{d}_{BLink} = \frac{\sum\limits_{i \in Link} d_{iA}}{n_{Link}} - \frac{\sum\limits_{i \in Link} d_{iB}}{n_{Link}} = \frac{\sum\limits_{i \in Link} \left(d_{iA} - d_{iB}\right)}{n_{Link}},$$

where n_{Link} is the number of link items.

The equating constant t is added to every item and ability on Form B as before. After adjustment, the mean, not each individual item, of the link set will be identical in both contexts.

Link Length

The number of items that should be in the link, like any sample size, depends on how precisely you need to know the answer. Using the Wright-Douglas approximation, the typical standard error for t is:

28. $$se_t = \sqrt{\frac{1}{n_{Link}} \left\{ \frac{6}{N_A} + \frac{6}{N_B} \right\}},$$

where N_A is the number of students used in the form A calibration and N_B is the number used in the form B calibration. Assuming a modest 500 students per form and 15 link items leads to a standard error for the equating constant of a rather loose 0.04 logits. Doubling the sample size to 1000 students, reduces the standard error by 25% to 0.03.

Turning the relationship around provides the link length needed for a given standard error:

29. $$n_{Link} = \frac{1}{se_t^2} \left\{ \frac{6}{N_A} + \frac{6}{N_B} \right\}.$$

If you want to know the equating constant to plus or minus 0.01 logits and have 5,000 students per form, a link length of $12 / (5{,}000 \times 0.01^2) = 24$ items is required, which is often pushing the limits of the form builders.

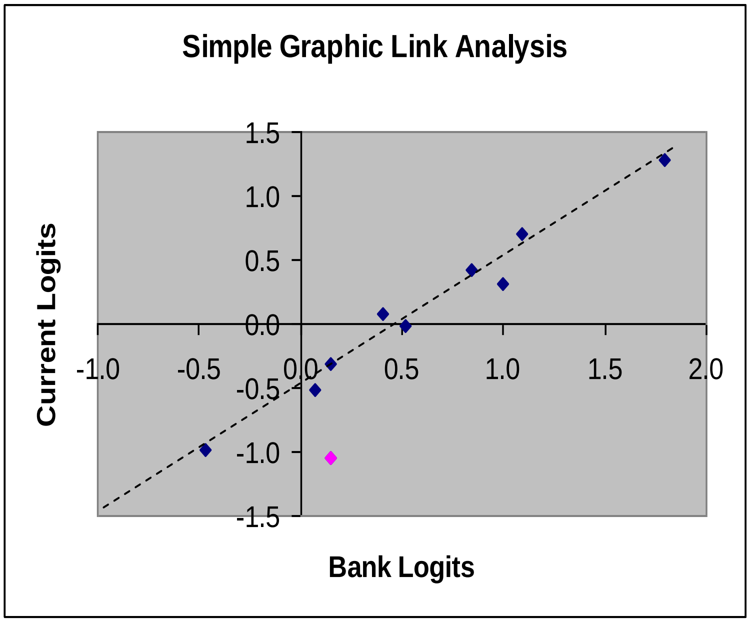
Inconveniently, limitations on the test length, content balance, and item exposure often have more to do with the size of the link then with the magnitude of a standard error needed to make the psychometrician comfortable. And most get very uncomfortable when n_{Link} is as low as 10 items, given the unimpeachable logic of equation 29.

Link Control

Multiple link items, in addition to increasing the precision of the estimate of the equating constant, are necessary to control the process. Each item pair provides an estimate of the equating constant. Like statistics everywhere, the estimates will not be identical; we expect some variation and would be concerned if we didn't have it. The problem is to recognize and eliminate outliers from the calculations of the means[3]. There are a number of more or less heuristic techniques for dealing with the uncertainty.

Figure VI.4 shows the simplest possible link analysis. The points plotted are the logits obtained from a calibration of the current administration against the logit difficulties from the bank. The data should follow a slope one line with the intercepts representing the required equating constant. This example is a very clean link, with one outlier, and an x-intercept of 0.5 (or a little less), ignoring the outlier. Adding 0.5 to every current logit will shift the plotted points vertically so that the slope one line passes (almost) through the origin and the current administration has been equated to the bank[4]. Spotting the outlier in this case is trivial when done visually.

Figure VI.4: Sample Link Analysis



[3] It should also be noted that n_{Link} is the number of items we want in the link after the outliers are dropped.

[4] This process is symmetrical. The y-intercept, -0.5, could be added to the bank logits to shift the plot horizontally and place them on the current scale if that made sense to anyone.

Worst First

The same analysis and slightly more precision can be achieved with the appearance of much more rigor if we use tables and numbers. The simplest non-graphical method is drop the item with the biggest discrepancy between the adjusted current logit and bank logit, $t_i = d_{iA} - d^*_{iB}$, and to continue doing dropping items until it doesn't have a noticeable effect on the result. Or it may just vacillate up and down depending on whether the last item dropped had a positive or negative discrepancy. At that point you are just analyzing noise. One of these situations typically happens with discrepancies around 0.25 logits. The procedure is widely criticized because it doesn't specify what *noticeable effect* means or which side to pick when the result is vacillating. On the other hand, some policy makers like this wiggle room.

Constant Criterion

An approach that at least appears more objective is to choose a criterion logit value and reject any item from the link if its estimate $t_i = d_{iA} - d^*_{iB}$ is larger than the criterion in absolute value. The most commonly mentioned criterion is 0.3 logits. While drawing this line in the logit sand may be more objective, it is just as arbitrary but based on a lot of experience and not much different in practice. This is easy to apply but widely criticized because it ignores the standard errors of estimation and applies the same criterion regardless of how well we think we know the item's difficulty.

Student's t

A little more statistically pure strateg performs a *Student's t-test* on each item using the standard errors of calibration. Items are rejected if the *t-statistic* exceeds the psychometrician's tolerance level. This is widely criticized because it considers the standard errors and is more tolerant of discrepancies when they come from poorly estimated items although all have same impact on the equating constant.

Robust Z

Our final strategy, which seems to incorporate the weaknesses of all of the above, uses a simple robust estimate of the standard error of the differences and applies it uniformly across the items. A *robust-z* is computed for each link item as:

30. $$z_i = \frac{t_i - Q_2}{0.74(Q_3 - Q_1)},$$

where Q_1, Q_2, and Q_3 are the first, second, and third quartiles of the distribution of the t_i. An item is eligible for rejection when z_i is greater than *1.645* in absolute value. It's a good thing when the items are very consistent, but in that situation $Q_3 - Q_1$ is very small so small discrepancies can make for a large *z*. And, if we truncate the distribution enough, the quartiles will eventually be very close together. This approach is criticized because it will always find a biggest discrepancy based on the empirical distribution and not on any theoretical or philosophical considerations.

To provide a rational stopping rule, no (more) items are dropped when:

- ratio of standard deviations of the two sets of difficulties is between 0.9 and 1.1,
- correlation between the two sets of difficulties is at least 0.95, and
- you are in danger of running out of link items.

A large *z* makes an item eligible for dropping but does not insist that it must go. Some in the business also set a minimum number of link items that they won't go below but sometimes you just have to admit you have a problem.

The Arithmetic

Table 9 contains a summary of these analyses for the same data used in Figure 4. The *Pool* and *Current* logits are given. The *Difference* is the *Pool – Current;* the *Adjusted* is the *Current + average Difference;* and the *Discrepancy* is the difference between the *Pool* and *Adjusted* logits. The *Student's t-statistic* is *Discrepancy* divided by the standard error. The *Robust Z* is defined by equation (30).

Table 9: Sample Link Calculations

First Round: All Items	Pool d_{iA}	Current d_{iB}	Difference d_{iA}-d_{iB}	Adjusted d_{iB}+t	Discrepancy d_{iA}-(d_{iB}+t)	Student's t-statistic	Robust Z
1	1.089	0.705	0.383	1.266	-0.177	-0.76	-1.23
2	0.149	-1.043	1.193	-0.483	**0.632**	**2.35**	**6.57**
3	0.148	-0.311	0.459	0.250	-0.102	-0.43	-0.50
4	0.844	0.415	0.429	0.976	-0.132	-0.57	-0.79
5	0.074	-0.519	0.592	0.042	0.032	0.13	0.78
6	0.402	0.081	0.320	0.642	-0.240	-1.04	-1.84
7	-0.472	-0.979	0.508	-0.419	-0.053	-0.20	-0.03
8	0.515	-0.015	0.529	0.546	-0.031	-0.13	0.18
9	0.998	0.320	0.678	0.881	0.118	0.51	1.61
10	1.792	1.278	0.514	1.838	-0.046	-0.19	0.03
Mean	0.554	-0.007	t = 0.561	0.554	Q2 = -0.05	-0.03	
Std Dev	0.644	0.732	0.244		Q3 = 0.02	0.90	
Ratio SDs		0.88			Q1 = -0.12		
Correlation		0.94					

Second Round: Drop Item 2							
1	1.089	0.705	0.383	1.196	-0.107	-0.46	-1.67
2	0.149	-1.043		-0.553			
3	0.148	-0.311	0.459	0.179	-0.031	-0.13	-0.65
4	0.844	0.415	0.429	0.906	-0.062	-0.27	-1.06
5	0.074	-0.519	0.592	-0.028	0.102	0.42	1.14
6	0.402	0.081	0.320	0.572	-0.170	-0.74	-2.52
7	-0.472	-0.979	0.508	-0.489	0.017	0.07	0.00
8	0.515	-0.015	0.529	0.476	0.039	0.17	0.29
9	0.998	0.320	0.678	0.810	0.188	0.82	2.29
10	1.792	1.278	0.514	1.768	0.024	0.10	0.71
Mean	0.599	0.108	**t = 0.490**		Q2 = 0.02	0.00	
Std Dev	0.667	0.674	0.108		Q3 = 0.04	0.44	
Ratio SDs	0.99				Q1 = -0.06		
Correlation	0.99						

Using all items, the ratio of standard deviations is 0.88 and the correlation is 0.94. Under Huynh's ground rules, we are forced to do something. Only item two is eligible for elimination with a *robust z* larger than 1.645 (and a discrepancy larger than 0.3 logits and a *Student's t* twice as big as anything else.) Dropping this item changes both the SD ratio and the correlation, coincidently, to 0.99, which implies we are finished. The end result, no matter how we did it, was we dropped one item and added 0.49 (almost the 0.5 from the plot) to the current logits to place them on the pool logit scale. (The *Student's t* was the weakest test in this situation but the calibration was based on only 100 examinees.)

Because all the criteria for a satisfactory link (no discrepancy larger than 0.3, correlation greater than 0.95, ratio of standard deviations between 0.9 and 1.1) are met, there is no reason to have computed the *robust z* statistics for the second round. However, having done it, there are two that are larger than 1.645. This is the nature of this calculation with very consistent items; there will always be a most extreme value. If all the other criteria are met, we end the process, report the scores, and go to dinner.

Dropped but not Forgotten

But before we go, there is an annoying question about what to do with item two. We have excluded it from the link, but do we keep it in the Bank, on the form, and in the examinee scores? If we want to keep it on this form or in the Bank, what logit difficulty should it have? One school of thought is that we should continue to use the Bank value because it typically is based on a larger, more defensible sample and analysis. The alternative view is that the value from the current administration is a better reflection of the current context and if you aren't willing to use that, the item isn't appropriate for this situation.

Because the logit difficulty for that item is not consistent across forms (or administrations); the item still may have functioned acceptably in both situations, albeit differently. The first thing to check is that the difficulties have been matched correctly. If so, it may be the item was unusually amenable to instruction or it may mean it interacted with popular culture (e.g., movies, commercials, music lyrics, current events) in some way others did not. Or it may mean a security breach or other malfeasance.

It is generally good for the psychometrician's peace of mind and often informative to identify the source of any disturbance.