EDUCATIONAL DATA SYSTEMS, MORGAN HILL, CA

WEIGHTING AND CALIBRATION

MERGING RASCH READING AND MATH SUBSCALE MEASURES INTO A COMPOSITE MEASURE

MARK H. MOULTON, PH.D.

markm@eddata.com
www.aobfoundation.org

ABSTRACT

While the emergence of Rasch and related IRT methodologies has made it routine to update tests across administrations without altering the original Pass/Fail standard, their insistence on unidimensionality raises a problem when the standard combines performance on multiple dimensions, such as mathematics and language. How combine a student's mathematics and language measures to make a Pass/Fail decision on composite ability when the two scales embody different dimensions and logit units? Using client-determined weights and student expected scores, we review existing methods for combining unrelated subscales, encountered in a recent high-stakes certification exam, to produce composite logit measures without sacrificing the advantages of unidimensional IRT methodologies.

THE PROBLEM

Here is a common psychometric problem, encountered in a recent high-stakes certification exam:

For a number of years, examinees have been administered a 100 item exam with 55 mathematics and 45 reading items mingled together, plus an essay. To pass the exam, students must get at least 67 of the multiple choice items correct, whether on math or reading does not matter (it is a fully compensatory exam), plus a minimum essay score which we shall ignore for the purposes of this paper. The items have not been changed in years. Some are obsolete. Many are too easy and do not reflect current standards. The client wants to switch out some of the old items for new, harder items, but in such a way that the examinees who take the new test are held to the same standard as examinees who took previous versions of the test. How do we update the test without changing the Pass/Fail standard?

(For ease of reference, the old test form is referred to as the Winter 2003 test form while the refreshed test form is referred to as the Spring 2003 test form. An additional test form was administered in Fall 2003.)

This type of problem is routinely dealt with using some form of IRT model such as the Rasch (1-PL) model. Item difficulties are calculated. By linking the Winter and Spring tests with common items, the relative difficulty of the new test can be taken into account when computing person measures. This is common-item equating and allows examinees from both administrations to be held to the same standard.

In this case, since the test remained essentially unchanged for years, there had been no IRT equating, nor any need for it. Examinees passed if they got 67% of the items correct. So long as the test remained unchanged, this was a valid way to hold students from different test administrations to the same standard. However, in the process of updating the test to make the Spring form harder, the 67% rule ceased to reflect the original performance standard. An examinee required to get at least 67% of items correct on an easy test is not being held to the same standard as an examinee required to get 67% correct on a hard test.

When all the test items lie in the same dimension -- measuring the same type of ability -- common-item Rasch equating is straightforward. This is because the Rasch model is unidimensional in its specifications. It requires items to share a single dimension as a condition of fit between the model's predictions and the data, and indeed common-item equating only makes sense in terms of a single dimension. For our exam, this was not the case. Examinees were being held to a *composite* standard of performance on math and reading items -- demonstrably different dimensions, though somewhat correlated. How do we equate a 2-dimensional test across different administrations using a 1-dimensional measurement model? This paper reviews several ways to approach this problem, focuses on one in particular, and presents results from a recent certification exam.

APPROACHES

There are four main approaches to dealing with this problem:

1. **Leave the test as is.** In practice, this is the most common "solution." Regardless of the dimensionality of the test, whether it contains one or two or 10 dimensions, so long as the same test form is administered from year to year the examinees are being held to

the same standard – for the most part. The qualification is necessary because of "item drift," the tendency of items to change their difficulties over time for a variety of causes, usually becoming easier. One cause of drift is that when items are re-exposed from year to year, the examinee population catches on and learns how to prepare, making the items easier. This is particularly worrisome when there are real consequences attached to the test, since the incentive for cheating is much higher.

Because of item drift, there is a tendency for student measures to appear to drift upward through the life of a test form, then to drop precipitously when a new test form is administered, resulting in an artifactual saw tooth pattern in student achievement. This is one reason why it is preferable to refresh tests with new items on a regular basis, but this requires psychometric methods of test equating. Hence, the next approach.

2. **Rasch-analyze the full 2-dimensional test all at once.** This is feasible when the two dimensions are reasonably correlated, as Math and Language are (r = 0.66 in this case). We anchor the test on the Winter item difficulties and set a logit cut-point (0.708 logits) that corresponds to a 67% success rate. Although the Spring students are exposed to a more difficult test form, because they are measured on a scale that has been calibrated relative to difficulty of the Winter test items they are held to the same standard as the Winter students.

   Unfortunately, this approach suffers several theoretical shortcomings that arise from trying to analyze 2-dimensional data with a 1-dimensional IRT model.

   a. Analysis of fit becomes difficult. This is because item misfit is now driven by the disturbing presence of the secondary dimension. In our case math is the dominant dimension and language is the secondary dimension. In attempting analysis of fit we observe that a number of language items are misfitting, so we suspend them from the analysis and recalibrate. But this has causes the composite dimension to shift closer to the pure math dimension, making a new crop of language items misfit. We suspend these, too. The composite dimension shifts even further toward math, causing yet more language items to misfit. Eventually, we have no language items left and are left with a unidimensional math test.

      Many analysts prefer to skip the analysis of fit stage altogether. Then, of course, they forego the benefits of such analysis – to identify bad items and make the test such that it yields reproducible results. A better alternative is to perform analysis of fit on the math and language dimensions separately. When the dimensions have been separately "cleaned," the math and language items are combined in a composite run, at which point the fit statistics are ignored as an artifact of multidimensionality.

   b. When equating, it is hard to control the composite dimension. Equating a 2-dimensional test is theoretically straightforward if we can assume that all math items equally represent the math dimension and all language items equally represent the language dimension. When this is the case, the orientation of the composite dimension is controlled by adjusting the relative weights of the two dimensions. Unfortunately, items only approximate the dimension at which they are aimed, and often the correlations are surprisingly low (on our test, the point-biserials range from 0.05 – 0.44 for language, from 0.09 – 0.48 for math), lower than the correlation between the two content dimensions (r = 0.66). Therefore, when it comes time to update the test with new items, it is hard to be certain that the new composite dimension exactly corresponds to the old composite dimension. If the composite dimensions are even a little different, the relative pass rates on the two tests may shift significantly in a way that is purely artifactual.

It helps to anchor the scale on the difficulties of the original items, but this does not guarantee that the new composite dimension is the same as the old one. Nor can we apply analysis of fit to force the dimensions to be the same, since the fit statistics are now confounded by the effects of multidimensionality. In the end, we are forced to go back and analyze the math and language dimensions separately in order to take advantage of the analysis of fit procedure which makes it possible to bring the old math dimension into line with the new math dimension, and the old language dimension into line with the new language dimension.

This brings us to the next approach.

3. **Analyze the math and language subscales separately, then combine them somehow.** Treating the subscales separately is appealing from a theoretical perspective because it fits better with the specifications of the Rasch model and holds out the promise of true unidimensional equating. In our case, this means equating the Spring form with the Winter form in terms of common items for math and language separately. Fit statistics and point biserials are used to identify those items which best embody their respective dimensions and which best fit the dimension of the Winter subscale. Items with poor fit or low point biserials are suspended from analysis. A similar analysis may be performed at the examinee level in order to clarify the latent dimension.

   The problem is to figure out how to combine the math and language measures to determine the Pass/Fail status of an examinee in a way that is comparable with the original standard.

   One approach that does *not* work is to combine the logits from the math and science subscales. This is because the two logit subscale metrics are not directly comparable. Although logits retain a constant *probabilistic* meaning across tests, their scaling can change due to the fact that probabilities are a function both of the latent variable and of administration-specific variance or "noise." Thus, the same test can be administered to the same students and yield different logit metrics on different occasions, even though the students maintain the same relative positions on the scale. This means that logit measures from different subscales cannot simply be combined or averaged in some way if the subscales lie on different dimensions. (If they lie on the same dimension, a simple scaling factor equates them.)

   A simple solution becomes apparent when we reconsider what is meant by test equating. Consider the following definition:

   Two tests forms (from different administrations, in this case) are considered equated if, from an examinee's response vector on the second test, it is possible to predict what the same examinee's response vector *would have been* on the first test.

The definition seems sensible enough. It states that we can compare examinees from two test administrations if we can somehow predict how each group of examinees would have performed on the other test. This allows the two groups to be compared in terms of the same test, which is exactly what conventional Rasch equating does by means of common test items. Note, however, that the above definition makes no mention of unidimensionality. In other words, unidimensionality is *not* a necessary precondition of equating, however useful it may be as a convenient assumption for facilitating prediction across forms.

Therefore, in order to equate the Winter and Spring 2-dimensional forms, it should be enough to predict how many items an examinee *would have gotten correct* on the Winter form based on his performance on the Spring form. This prediction can be obtained by converting the examinee's logit measures on the math and language subscales into expected scores for each subscale, and it is these *expected scores* (which are simply predictions of raw scores) that can be added and weighted to yield a hypothetical composite test score for the Spring examinee on the Winter test. There are two methods for calculating expected scores, described in detail in the next section. The Pass/Fail decision is made by determining whether the composite *expected* test score as a percentage of the whole test exceeds the 67% threshold.

4. **Employ a non-unidimensional IRT model to compute the necessary expected values.** Because the Rasch model specifies unidimensionality as a condition of fit, it is ideally used on only one dimension at a time in order to retain its most important properties. A non-unidimensional scaling model (NOUS), on the other hand, relaxes the unidimensionality requirement for a given test, allowing the computation of expected scores for each cell in the matrix even when the items participate in any number of different dimensions. It is distinguished from existing multidimensional Rasch models by its method of sharing information across item subscales.

   Unfortunately, such models are new in the field of Item Response Theory and have yet to be studied properly. (Moulton, 1996, 2001, www.aobfoundation.org.) Nonetheless, the application is fairly straightforward. A data matrix is constructed consisting of Winter and Spring items, both math and language, each row corresponding to an examinee from one of the two administrations. Blank cells exist for items belonging to an administration the examinee did not attend. The NOUS model analyzes all the data together, trading information across subscales, and computes expected values for each cell of the matrix including the missing cells. The sum of expected values across the Winter items becomes the score for each examinee, regardless of test administration. Thus, NOUS yields a very direct answer to the question, "How would a Spring examinee have performed on each item given in the Winter administration?" In this way, the Spring examinees are made comparable to the Winter examinees, whose scores are also sums of expected values on the Winter items.

   It seems likely that problems of this type, which involve multidimensional datasets, will eventually be handled by some form of non-unidimensional IRT model. Until such models are fully understood, however, the unidimensional models are the tool of choice.

COMBINING SUBSCALES USING EXPECTED SCORES: TWO METHODS

As mentioned above under the third approach, there are two methods for converting logit subscale measures for math and language into expected scores that answer the question, "How many items on the Winter test form would this Spring examinee have gotten correct?"

Method I

1. **Equate the two administrations for each subscale.** Place the mathematics items for the two administrations on a common scale by anchoring the Spring items to the Winter math items through common items, where the zero point of the logit scale is set by convention at the mean of the Winter items. Perform analysis of fit to ensure that the math items have stable and generalizable difficulties and embody a single dimension. Do the same with the language items.

2. **Compute expected proportion correct for each subscale.** Each examinee has a logit ability measure for each subscale. Calculate the difference ($\theta_{nM} - d_{0M}$) between examinee *n*'s logit measure and the mean logit difficulty d_{0M} of the Winter test for one of the subscales, say, Math (the M subscript). Convert the difference into an expected proportion correct of items on the math subscale using the logistic probability formula:

   Expected Proportion Correct$_{\text{Math}}$ = $\exp(\theta_{nM} - d_{0M})/(1 + (\theta_{nM} - d_{0M}))$ Eq. 1

   This is the Rasch formula for dichotomous data used to calculate the probability of an examinee with a certain ability succeeding on an item of a certain difficulty. In this case, the examinee's ability is given by his logit ability measure. The item difficulty is the mean difficulty of the items in the Winter administration for that subscale, set at zero by convention. This difficulty may be interpreted simply as the "difficulty of the test," or when subtracted from the examinee's ability measure, as the expected proportion of items that an examinee at that ability level may be expected to get correct on that test.

   Compute the Expected Proportion Correct (EPC) for each examinee for both the math and language subscales.

3. **Compute examinee Expected Proportion Correct for the whole test.** This requires having the client provide weights for each subscale according to its perceived importance in meeting the standard. Such weights are a matter of definition and cannot be empirically determined. Are math and language equally important? Assign a weight of .5 to each. Is language more important? Give it a weight of .6 and math .4. In this case, the client assigned math a weight of .55 and language a weight of .45 by simply designing the test to have 55 math items and 45 language items. The weights and the examinee's expected proportion correct are put into the following formula to get the examinee's expected proportion correct for the whole test:

   EPC[Whole Test] = (EPC[Math])(W[Math]) + (EPC[Lang])(W[Lang]) Eq. 2

   where EPC refers to the Expected Proportion Correct and W refers to the Weight assigned to that subscale. The weights are defined to add to 1.

4. **Compare with the Standard.** It is this number, the expected proportion correct, that is to be compared with the original standard to determine whether the examinee passed or failed

the test. In this case, that standard was 67%. Therefore, if an examinee's expected proportion correct on the test met or exceeded 67%, the examinee passed the multiple choice section (we are ignoring the effect of the essay score).

Method II

Method II is like Method I in all respects except in how it computes the expected proportion correct. In Method II, the expected proportion correct is simply the average of the expected scores of the items from the Winter administration, for that subscale. These expected scores are computed automatically by the Rasch software (WinSteps, in this case) as the predictions for each cell which are compared to the observed values. They can be accessed through the IPMatrix the XFile output files. However, to get the expected scores for missing cells it is necessary to "fool" WinSteps into believing that the matrix is complete. To do this, anchor all the items and persons at their measured logit values and fill the data matrix with nonsense data that is nonetheless analyzable. WinSteps will then report the correct expected values for each cell. Expected scores can also be computed manually using Equation 1 applied to each item individually.

Method II, while somewhat laborious to implement, has an important advantage over the other methods: It leads to composite scores that *approximately* match the original raw score pass rate (the match is increasingly exact as the number of iterations increases). Both Method I and Approach 2 in the previous section (Rasch-analyze both dimensions simultaneously) lead to expected proportions correct that have a slight positive bias, increasing the pass rate by as much as 78 examinees (out of 1616). The reason for Method II's ability to match the original raw score pass rates lies in the WinSteps Maximum Likelihood Estimation algorithm, which adjusts the marginal parameters (person abilities, item difficulties) until the sum of residuals for each row and column approximates zero. This is equivalent to having the sum of expected values equal the sum of observed values, which causes the match with the original pass rate.

While Method II is preferable, Method I results are shown below as it is easier to use.

Converting Back to Logits

Expected scores are useful when the passing standards are set in a raw score metric. However, like raw scores, they are non-linear and not useful for actual measurement. It is for that reason that pass/fail standards are generally not set in a raw score metric but in a logit or scale score metric. Can our method be generalized to logit standards? Yes. Just convert the examinee's expected proportion correct on the whole test back into a logit metric:

Person Measure[Whole Test] - 0 = logn(EPC[Whole Test] / (1 – EPC[Whole Test]))

where EPC refers to the examinee's expected proportion correct. The zero term is the mean difficulty of the Winter test. If the person measure meets or exceeds the standard set in that logit metric, the student passes. In this case, the 67% standard corresponds to a cut-point of 0.708 logits.

Note that with both methods we have inadvertently equated the logits from two separate subscales. This was only possible because we tacitly assumed the expected proportions correct for the two subscales to be comparable. We assumed in effect that the difference between getting 70% and 50% of math items correct has the same meaning as the difference between getting 70% and 50% of language items correct. In subscales that are similarly "noisy," as one would expect of subscales drawn from the same test administration, this is a reasonable assumption.

RESULTS WITH THE WINTER/SPRING/FALL ASSESSMENT DATA

The raw score pass rates for the Winter assessment are shown here, plus the pass rates for three of the theoretical methods described above. These numbers show that all three theoretical methods approximate the true rate quite closely, with a maximum of 78 out of 1616 false positives, i.e., examinees who pass based on their theoretical scores but who fail based on their observed raw scores. Method II yields exactly the same pass rate.

Table 1: Pass Rates, Observed and Theoretical, for the Winter Administration

Approach	*Pass Rate (% of 1616)*	*False Positives*	*False Negatives*
Raw Score	0.638	N/A	N/A
Both Dimensions Rasch-Analyzed at Once, 0.708 cut-point	0.686	78	0
Method I	0.683	73	0
Method II	0.638	0	0

Table 2 shows the raw scores and expected scores (Method I) for three administrations (Winter 2003, Spring 2003, and Fall 2003) for Math and Language. We see from the raw scores that the Spring and Fall tests were much more difficult than the Winter 2003 test due to the introduction of harder items. Without some form of equating, such differences in raw scores would lead to the mistaken conclusion that the examinee aptitudes in Math and Language dropped significantly. The Average Expected Scores tell a truer story. For all three administrations, they were essentially the same as one would expect for a time-frame less than a year.

Note that the Winter 2003 Average Raw and Expected scores are similar but not identical. This is a statistical artifact of Method I. Using Method II, the two become identical.

Table 2: Comparison of Raw and Expected Scores, 3 Administrations

	Mathematics		Language	
	Average Raw Score	Average Expected Score	Average Raw Score	Average Expected Score
W03 Administration	36.1	37.6	34.1	36.3
S03 Administration	32.4	36.4	31.0	36.7
F03 Administration	32.4	35.9	28.9	36.6

Conclusion

The Rasch logit metric necessarily embodies a single dimension, valid only for a specific type of item. Raw scores or percentages can apply to any number of item types. By exploiting the ability to go back and forth between the logit metric and the score metric by way of theoretical expected scores, and by incorporating client-defined weights, it is possible to resolve performance on unrelated subscales into a composite expected score or logit measure which can be compared to a standard. This makes it possible to update tests and change their overall difficulty without losing the ability to compare students on the original composite standard.

References

Moulton, M. H,. *N-Dimensional Replacement: Implications of a Rasch Geometry*, Dissertation: The University of Chicago, 1996

Moulton, M. H., *Transcendence: Four Thought Experiments with a Non-Unidimensional Quasi-Rasch Model*, 2001, www.aobfoundation.org.

Rasch Measurement Transactions: www.rasch.org/rmt/

Rasch, G., *Probabilistic Models for Some Intelligence and Attainment Tests,* Chicago: The University of Chicago Press, 1980.

Wright B.D. & Stone M.H., *Best Test Design*, Chicago: MESA Press, 1979.