EDUCATIONAL DATA SYSTEMS, MORGAN HILL, CA

PUBLISHED BY:
EDS PUBLICATIONS, DECEMBER 2004
WWW.EDDATA.COM

ONE USE OF A NON-UNIDIMENSIONAL SCALING (NOUS) MODEL

TRANSFERRING INFORMATION ACROSS DIMENSIONS AND SUBSCALES

MARK H. MOULTON, PH.D.

ABSTRACT

Test administrators sometimes ask for student performance on test subscales having few items, rendering them unreliable and hard to equate. Worse, subscales sometimes embody orthogonally distinct secondary dimensions as well. Traditional Rasch analysis offers reasonable solutions in some cases, but not all, and is not a general solution. This paper proposes a general solution using a Rasch-derived *non-unidimensional scaling* measurement model, called NOUS, which transfers information across items, subscales, and dimensions. Drawing examples from a recent state exam, it shows that NOUS yields measures for short subscales that are comparable to unidimensional measures computed using long forms of the same subscale. It concludes by discussing applications for multidimensional equating, student-level diagnostics, and measurement of performance on open-ended items.

EDUCATIONAL DATA SYSTEMS
15850 CONCORD CIRCLE, STE A
MORGAN HILL, CA 95037
408-776-7646
MARKM@EDDATA.COM

THE PROBLEM

Test administrators sometimes want to report student performance on test subscales for which there may be only a few items. In such cases, subscale measures become unreliable due to high error, and hard to equate to other forms and administrations due to a lack of common items within the subscale. Worse, they sometimes embody dimensions slightly or substantially at odds with the dominant dimension of the test. Traditional Rasch analysis offers reasonable solutions in some cases, but not all, and cannot be viewed as a general solution for the following reasons:

1. The Rasch Model specifies that items embody a single dimension. To the degree subscales embody dimensions that are poorly correlated to the dominant dimension of the test, the model loses its capacity to predict and measure subscale performance. At the same time, the dominant dimension of the test becomes increasingly tenuous and hard to define, and analysis of fit becomes problematic.

2. When Rasch is focused on a short subscale, the low number of items increases error and decreases reliability, which is a function of the ratio between the standard deviation of the examinees and their root mean square error (RMSE). The RMSE is driven by the number of items in the subscale.

3. Items can be equated across administrations only to the degree they participate in the same dimension. When there are only a few items in a subscale and the subscale is on a separate dimension, it becomes impossible to compare performance on that subscale across administrations.

4. When subscale measures are computed separately, they do not share the same logit metric, leading to problems with comparability.

Multidimensional Rasch models, such as that embodied in the "between-item" multidimensionality component of the ConQuest program (Wu, Adams, Wilson, 1998), were not designed for this type of problem. Their purpose is primarily to disentangle the dimensions in a dataset, not to transfer information across dimensions, although ConQuest EAP measures do exploit distributional information that has the effect of transferring information. There may be other variants of multidimensional IRT Models that explicitly transfer information across dimensions, such as Reckase's Linear Logistic Multidimensional Model (Reckase, 1997), but I have not had the opportunity to study them operationally.

Outside the field of IRT, this type of problem is addressed through multivariate regression models and perhaps most powerfully through neural networks (Bishop, 1995), an advanced form of regression. Such methods have not, so far as I know, been adapted to educational testing, though they might work well enough. As they were not developed according to any explicit specification for "objective measurement," neural networks are strongly dependent on representative samples and have a tendency to produce predictions that describe an initial "training" data set very well, but fail to generalize.

We are thus led to speculate on the possibility of an IRT model that combines the "objectivity" requirements of the unidimensional Rasch model with the multidimensional predictive powers of regression and neural networks. Such an IRT model should possess the following properties:

1. Non-unidimensional. The model should be capable of modeling persons and items in a space of any number of dimensions in such a way that each item is allowed to define its own unique dimension.

2. Information Transfer. In estimating how each person does on a given item or subscale, the model should be able to use all information in the data matrix that has some bearing on that item or subscale and to increase the reliability of the corresponding measures accordingly.

3. Specific Objectivity. The model should define an objective space of n-dimensions in which the relative positions of the persons and items are mutually independent, i.e., separable. Thus, the relative positions of the persons should not be affected by the item sample, and the relative positions of the items should not be affected by the person sample, as a condition of fit between the observed and estimated values.

4. Missing Data. The model should be robust to missing data and possess no theoretical need for complete data sets. This contrasts with conventional statistical methods, including regression and neural networks. The ability to handle missing data is essential to IRT, not just for the practical reason that data are often missing, but because equating designs (the *raison d'etre* of IRT) necessarily involve significant blocks of missing data.

5. Bad Data. The model should clearly identify data values that are not likely to meet the conditions for reproducibility. For instance, if we were deliberately to reverse one of the data values so that we know on *a priori* grounds that it is wrong, the model should not try to adapt itself to the new value. It should yield a clear and significant difference between the reversed value and the model's own estimate for that cell.

6. Random Data. The model should, so far as possible, be robust to the effects of randomness in making its predictions and measurements, yet be able to tell the researcher the degree to which the data truly are random.

Although the model discussed in this paper possesses all six of these properties, I shall only discuss it as it relates to the second property, Information Transfer. In particular, I will apply the model to a test containing a mix of math and language items in which the language items are treated as a subscale. I will assess how well the model can be used to recover "benchmark" language measures when the number of items in the language subscale is very small.

THE METHOD

This paper proposes a general solution in the form of a non-unidimensional scaling measurement model (to coin a phrase) called NOUS, whose distinguishing feature is that it transfers information across items, subscales, and dimensions, even when they are poorly or negatively correlated. It does this by means of an algorithm that computes correlations between item vectors and person-pair vectors computed from the test as a whole. While the algorithm itself is unrecognizable in terms of existing multidimensional algorithms, it reduces to a Rasch-like statement of the performance of each response being a function of a person's ability and an item's difficulty in the dimension implied by that item.

The NOUS model can be summarized as:

$$G_{ni} = B_{ni} - D_{ii} + e_{ni} \qquad \text{Eq. 1}$$

The performance G_{ni} of Person n on Item i is defined in terms of Person n's ability on the iota dimension minus Item i's difficulty on the same dimension plus some normally distributed error, where the iota dimension is defined by the spatial orientation of Item i. For simplicity's sake, I will refer to the iota dimension using the "i" subscript, since the item and the dimension that the item embodies are, for practical purposes, interchangeable. Also, I refer to persons as "rows" and items as "columns" in accord with general practice.

So far, Eq. 1 is consistent with the Rasch model with the exception of the iota subscript and its not being framed as a probability equation. The decision not to frame NOUS as a probability equation is pragmatic; the algorithm is simpler and quicker. Nonetheless, probabilities are easily calculated from NOUS estimates.

NOUS differs mainly in the way it defines and computes the ability and difficulty parameters.

$$B_{ni} = -1 * \sum_{M}^{R} E\left(M - N\right)_i \qquad \text{Eq. 2}$$

$$B_{ni} = -1 * \sum_{M}^{R} \left(MN * \cos q_{MNi}\right) \qquad \text{Eq. 3}$$

The ability B_{ni} (I reserve ? for use with angles) of Person n is defined as the estimated (E) sum of differences within a column across R person-rows between a given person N on item i (representing the iota dimension) and all other persons M on the same iota dimension. Be sure to note that this summation occurs *within* the column of item i, not across columns (the i subscript can be misleading). In order to estimate each of the differences in Eq. 2, we employ the definition of the cosine which states that the difference between two points on a given dimension equals the product of the absolute distance between them (MN) and the cosine of their angle with the given dimension ($\cos?_{MNi}$). The angle in question is not between two items, as one usually sees, but between one item and the line defined by a pair of persons. Almost the entire bulk of the NOUS algorithm is concerned with estimating these distances and cosines.

The difficulty of Item i can be expressed more conventionally:

$$D_{ii} = -1 * \sum_{M}^{R} X\left(M_i\right) \qquad \text{Eq. 4}$$

The difficulty $D_{i?}$ of Item i along its implied iota dimension is the negative sum of the observed values X(Mi) in the Item i column of M persons, where R is the number of persons with non-empty cells in that column. Eq. 4 is unimpressive from a measurement perspective, but since each item is treated as its own dimension, item difficulties lose much of their meaning relative to each other. Eq. 4 is more than adequate as a way of anchoring the ability distribution for prediction purposes.

The result is an estimate of person performance with standard error for each person/item interaction, corresponding to the expected values matrix in IRT programs. The NOUS estimates are

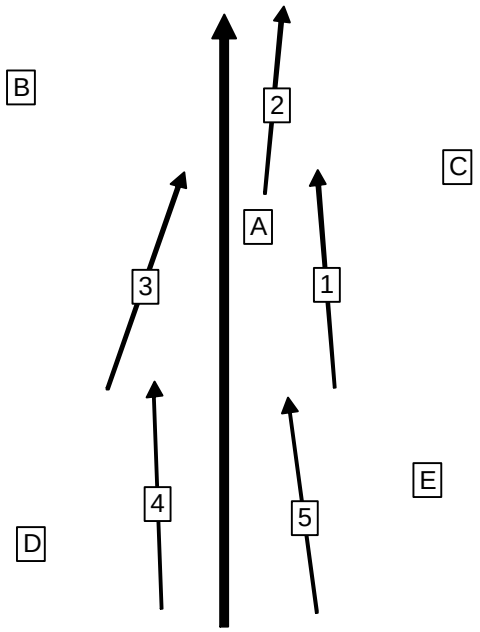
readily converted into linear measures for each item or item subscale. Because they draw information from the full data matrix, they are notably more precise and reliable than could be obtained from a unidimensional analysis of the same subscale. They are also comparable across items and subscales.

The essential innovation is that we have wholly dispensed with the specification that items share a common dimension.

THE GEOMETRY

NOUS constructs a simple geometrical picture. Persons are visualized as points floating around in space. Items are directions visualized as lines pointing in any direction, with demarcation points like a yardstick. Data are where each person projects perpendicularly onto each line. Figure 1 shows the unidimensional dichotomous case.

Figure 1: The Unidimensional Case



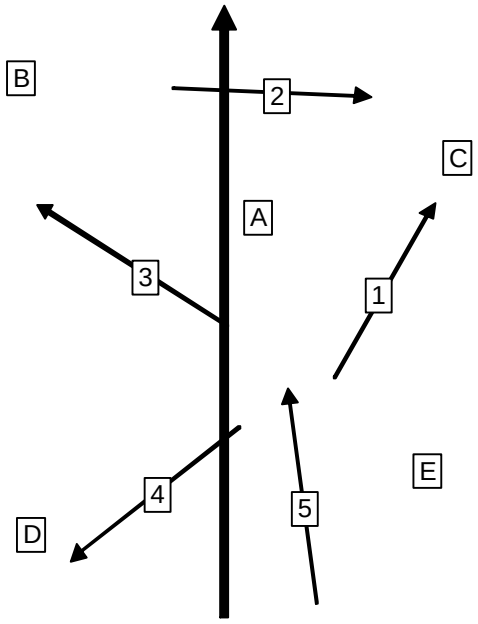
The bold vertical line represents the dominant dimension of the test. Each item is a line (direction, actually, also known as a vector), infinitely extended in both directions and more or less parallel to the dominant dimension, where the location of the label (numbers in the case of items) shows the "difficulty" of that item, i.e. the position below which a person gets a "0" and above which he gets a "1". Each person (indicated by the alpha labels) is a point floating in space of higher dimensionality than the dominant dimension of the test. Each data value is where a person-point projects onto an item-line, a "1" if he falls "above" the item label (on the side with the arrow), a "0" if he falls "below." For example, in the dichotomous case, we would say that Person B gets a "1" on Item 3 because he projects onto the region above the Item 3 label. Person A gets a "0" on Item 2 and a "1" on all the other items. And so on.

Figure 2 gives the 2-dimensional case (since the diagram is restricted to the plane of the paper), in which the unidimensionality specification has been relaxed. We now see that the items are pointing anywhere within the plane. Data are modeled the same way as in the unidimensional case. For

instance, we see that Person B gets a "0" on Item 4, even though it is "higher" up the bold dimension of the test (which is no longer dominant, it should be noted). If any of the items were to point outside of the plane of the sheet of paper, we would have the 3-dimensional case. In the general case we allow the lines to point in any direction in a space of any number of dimensions (though it cannot exceed the number of items), and we imagine the persons as points floating around in that space.

In analyzing data from Figure 2, one can easily see where a unidimensional model will run into trouble. The unidimensional model will pick the "average" orientation of the items as its dominant dimension, and all the items that are not aligned to that dimension (almost all items) will misfit accordingly. If there are a significant number of "reversed" items like Item 4, the person and item measures will tend to collapse to the center of the scale since there *is* no dominant dimension.

Figure 2: The 2-Dimensional Case



NOUS uses a combination of statistical methods and projective geometry to analyze data assumed to originate from a space of some unknown number of dimensions. There is no theoretical or practical limit to the number of dimensions it can analyze at a time (the computational time is the same), although the reliability of estimates decreases the lower the ratio of items to orthogonal dimensions.

These figures are formalized in the following definitions:

1. A "person" is a *point* in n-dimensional space, and can be thought of as a vector. Each point occupies one and only position in that space.

2. An "item" is a *direction* or vector in n-dimensional space, infinite in extent, embodied as any of the set of lines parallel to that direction and displaced orthogonally relative to it. For ease of reference, such directions are often called "lines." Each direction/line possesses boundaries demarcating equally spaced sections on the line, and these boundaries are indicative of "difficulty" along the dimension embodied by that direction/line. Each line

has one and only one orientation in that space and each of its boundary points occurs at one and only one location on that line.

3. A "datum" is the locus of *projection* of a point onto a line, in which the projection line is the perpendicular dropped from the point to the line and the locus of the projection is given by the value assigned to that region of the item-line. Each projection receives one and only one such valuation.

4. "N-dimensional space" is that medium by which points *differ* from each other on items. The number "N" is the minimum number of orthogonal dimensions needed to account for the variance of the projections of all the points on all the item-lines for a given data set. That space is called "objective" whose persons retain constant relative positions across items samples and whose items retain constant relative orientations and difficulties across person samples.

THE ESTIMATION ALGORITHM

Since the specification of the non-unidimensional model in 1996 (Moulton, 1996), at least three markedly different algorithms have been discovered for estimating the item-specific difficulty and ability parameters used to estimate each person/item interaction in the n-dimensional case. They can be grouped under two headings: Numeric Methods and Deterministic Methods.

The Numeric Method, in the tradition of the Newton-Raphson method used in many maximum likelihood routines, establishes a coordinate system of some specified number D dimensions to describe the person/item space. Within that space, it assigns starting positions to each person-point, and starting orientations (suitably normalized around the origin) to each item-line, and computes projection estimates for each point onto each line. These estimates are compared to the actual data values and fit statistics are computed. Each person-point is then moved through the D-dimensional space until a position is found that maximizes the fit of the projections to the data. Similarly, each item-line is moved (in an angular sense) through space until a best-fit orientation is found. The process is repeated until a global best-fit solution is found (Silsdorf, 2001). All Numeric methods are iterative trial-and-error routines that maximize fit to a model, though they are capable of great refinement using tools from linear algebra.

A Deterministic Method attempts to solve the parameter estimation problem mathematically and does not include an iterative routine to maximize fit. It asks: What is the mathematical function that relates the value in one cell to the values in the remaining cells of the matrix? This function is used to estimate each cell directly. There is no trial-and-error adjustment of parameters to maximize fit.

The algorithm used in this paper is a variant of the Deterministic algorithm worked out in 1996. This variant, which refines the initial distance estimates, turns out to describe significantly more variance than the 1996 version when it comes to analyzing dichotomous test data (ordinarily the two versions are not so different). The refined version has not yet been posted on the web, but it can be requested of the author. Both deterministic versions are astonishingly complex, but they have proven themselves to be effective and robust in analyzing a wide variety of data-types, ranging from psychological profiles to commodity prices. Currently, all NOUS software has been written in this Deterministic vein. The Numeric approach has progressed only as far as prototypes.

The Numeric and Deterministic Methods have their advantages and disadvantages. Despite its complexity, the Deterministic approach has a computational advantage because: a) it does not require successive iterations; and b) computational time is unaffected by the number of orthogonal

dimensions. It also does not run the risk of degenerate solutions or multiple local solutions. The Numeric approach has the advantage of assigning each person and line unambiguous locations in a Cartesian coordinate space, and thus clearly specifying a point-line model, unlike the Deterministic approach which assumes such a space but is unable to label or identify positions within it. Much work is being done to amalgamate the two approaches.

While it is not possible to present the full Deterministic algorithm in this paper, a brief synopsis can be attempted:

1. Type-A Differences. The initial raw data matrix, called the Base array, is used to compute differences between every pair of persons for each item.

2. Type-A Correlations. Correlations are computed between item pairs in the Base array.

3. Type-A Distances (2D). The Type-A Differences and Type-A Correlations are combined to produce multiple estimates of the distance between each pair of persons, a separate estimate for each pair of items. The distance is called "2D" because the geometrical formula behind it assumes that the person-pair and item-pair used to compute each estimate lie in the same plane. To the degree they do not, the distances are underestimated. In order to correct this underestimation, the distances are refined using the higher dimensional formula given in the next step.

4. Type-A Distances (3D). A formula very similar to the 2D formula is employed, but instead of estimating person-pair distances using differences on lines, we use distances on planes (which are given by the 2D distances) and correlations between planes (which are computed by correlating the 2D distance estimates for each item pair). The 3D formula makes its own assumptions: a) the two persons lie in the 3-dimensional space created by the two planes; b) the two persons lie in a plane perpendicular to the intersection of the two planes. Violations of the first assumption cause the distance estimate to be underestimated; violations of the second causes it to be overestimated. To remove the effects of such distortions, each distance is estimated multiple times with different combinations of planes. Because it is reasonable to assume that the distortions will not be systematic across combinations of planes, we take a weighted average of these estimates to approximate the "true" distance between each pair of persons, or at least a distance estimate that suffers the same amount of shrinkage or expansion as all the other distances, which is sufficient for our purposes.

5. Projections. Using the Type-A Distances, we calculate where each person-point projects onto each line defined by a pair of person-points. The projection formula, derived from classical geometry, makes no assumptions regarding the dimensionality of the space.

6. Type-B Correlations. The Base array already gives us raw data approximations of where each person-point projects onto each item-line. Step 5 gives us estimates of where each person-point projects onto each line defined by a pair of person-points. Using these two pieces of information, we compute a correlation between each item line and each person-pair line. This correlation is an estimate (subject to distributional assumptions) of the cosine between an item-line and a pair of persons.

7. Type-B Distances. The distance between each pair of persons is calculated anew using a formula that makes no dimensional assumptions. Each person-pair distance is computed as the sum of Type-A Differences for that pair divided by the sum of Type-B Correlations for that pair (positive and negative differences are handled separately).

8. Type-B Differences. Each Type-B Correlation is multiplied by the Type-B Distance corresponding to that person-pair to compute an estimate of the difference between those two persons *that is specific to each item*. To a given person N, there is an array of differences between that person and all the other persons M in the sample. (The negative sum of these NM difference estimates for a given item corresponds to the ability of person N.)

9. Unsummed Estimates. Each Type-B Difference between target person N and reference person M is subtracted from the raw data value of each person M to create an array of independent estimates of how person N will perform on a given item. (The negative sum of raw data values across reference persons M equals the "difficulty" of the item, though this difficulty is not reported and is not of much use except as a way to anchor the person distribution.)

10. Final Estimates. Step 9 gives us a set of independent estimates for how Person N is likely to perform on a given item I. From these, a set of output statistics is computed:

    a. The *Mean* of these estimates gives the final estimate of how Person N will perform on item I.

    b. Its *Standard Error* is given by the standard deviation of these estimates divided by the square root of the number of estimates for that cell. Note that it is *not* directly driven by the number of items but rather by the number of persons.

    c. The P*robability* of success is calculated most directly as the percentage of estimates that fall above a specified threshold, such as 0.5 in the case of dichotomous data.

    d. The *Cell Residual* is the difference between each estimate and the raw data value for that cell, if it exists.

    e. The *Fit* for each cell is given by the cell residual divided by the standard deviation of the estimate for that cell. Misfit values can be aggregated across rows and columns to give person and item fit statistics.

    f. The *Person Separation* for each item is given by the standard deviation of the final estimates for each item column (adjusted for error) divided by the Root Mean Square Error (RMSE) of the estimates in that column. Item Separation does not have much meaning in the NOUS universe.

    g. The *Reliability* of an item is given by its Person Separation squared, divided by (1 + the Person Separation squared).

Due to the way the Type-B Differences are computed, the sum of expected values will automatically equal the sum of observed values for each row (or equivalently, the sum of residuals will equal zero), constituting an acceptable and practical best-fit solution. With older versions of NOUS, including the version used in this paper, the sum of residuals will not quite equal zero when the data are incomplete. Upcoming versions correct this defect at the person level.

The Deterministic algorithm naturally has theoretical limitations that are too involved to discuss here. The most important limitation is that it is forced to make statistical assumptions at crucial junctures. An example is the use of correlations. Correlations, used as a way to estimate the cosine of an "objectively" existing angle, can severely over-estimate a small cosine if the person-points are not

symmetrically distributed in space. Therefore, the Deterministic algorithm works best if a symmetrical (i.e. multivariate normal) distribution of points in space can be assumed. Where such a distribution does not exist, the algorithm will generally correct distortions through averaging procedures, but not always. Distorted correlations that are not corrected manifest either as cell misfit or high error or low variance explained, indicating that the objectivity requirements of the model have been violated.

THE RESULTS

NOUS was applied to dichotomous data from a recent high school mathematics and language exam. There was also an essay, ignored for the purposes of this paper. Out of 55 math items and 45 language items, a new hypothetical dataset was constructed consisting of 50 randomly selected math items and 15 language items, one of which was artificially constructed to be maximally aligned with the language dimension, having a point-biserial correlation of 0.65. The remaining 14 items were chosen on the basis of their point-biserials (ranging from 0.36 to 0.65) to represent "language." The 15 language items together represented a hypothetical language "subscale" within a larger 65 item test dominated by the mathematics dimension. Thus, the new "test" consisted of 50 math items plus the 15 language items.

Mathematics and Language measures were computed from the original test (all items) using the complete sample of students, approximately 10,000. Of those, 100 students were chosen at random. The measures of these 100 students were set aside to represent Math and Language "benchmarks" against which to measure the success of NOUS and other methodologies in predicting overall Language performance on the basis of the 15 items in the Language subscale. The hypothetical dataset therefore consisted of 100 persons and 65 items (50 math + 15 language).

The small student sample allows a comparison of methodologies for small n, which makes it possible to expose weaknesses in methodologies that are highly sample dependent.

To assess NOUS's ability to recover the benchmark language measures with subscales of varying widths, a series of 15 runs was conducted, the first with 15 language items, the second with 14, and so on, until there was only one item in the language subscale. The order of language item deletions was governed by their point-biserial correlations. The items with the lowest point-biserials were deleted first; the item with the highest point-biserial was left for last.

Because NOUS only computes estimates at the level of the individual item, there are several methods for extracting a NOUS measure from multiple language items:

Method 1. Compute a language score *before* running NOUS by averaging the language successes and failures for each student first. This is treated as a composite language "item." Run the matrix through NOUS, including the composite item. The resulting NOUS estimates for that composite item are the NOUS language subscale measures. Note that I say "estimates" rather than "expected scores." This is because the expected scores (probabilities in the case of dichotomous data) are bounded by 0 and 1, and are therefore nonlinear. The NOUS "estimates" on the other hand, though they *look* like probabilities, are not bounded by 1 and 0 and occasionally spill outside. In fact, when graphed against logit measures they appear to be linear with the exception of a few students at the extremes. Nonetheless, to facilitate comparison with WinSteps, the NOUS estimates are treated like probabilities and converted into logits through simple adjustments of the upper and lower bounds.

It is preferable to do this with actual probabilities, but the version of NOUS used in this paper does not compute them. The formula used to do the logit conversion was:

$$B_{ni} = \ln\left(\frac{p_{ni}}{1-p_{ni}}\right) \qquad \text{Eq. 5}$$

where B_{ni} is the student's ability and p_{ni} is computed as

$$p_{ni} = \frac{E_{ni} - Min_{Eni}}{Max_{Eni} - Min_{Eni}} \qquad \text{Eq. 6}$$

E_{ni} being the NOUS estimate, the Max and Min terms being the maximum and minimum NOUS estimates, respectively, in the Item i column.

Method 2. Choose the language item that on *a priori* grounds best embodies the language dimension. The NOUS estimate for this item, put through Equations 4 and 5, becomes the language measure. This method works surprisingly well and is the easiest. But it does not work quite as well as Method 3 in recovering the benchmark language measures.

Method 3. Compute the language score *after* running NOUS by summing the estimates across the subscale for a given person, dividing by the number of subscale items, and converting into logits using Equations 5 and 6.

Method 3 was the one chosen for this paper because of its parity with WinSteps, the selected Rasch program. The source of this parity is that the sum of NOUS estimates is constrained to equal the sum of corresponding observed values. Therefore, by converting the sum of NOUS estimates into logits, we are exactly duplicating the WinSteps person ability computed in the first step of its estimation routine. The result is that when NOUS and WinSteps are used to compute person estimates from the same complete dataset, they are very close. This mitigates one artifactual source of difference between NOUS and WinSteps when computing subscales.

On a reporting level, Method 3 has an intuitive simplicity that may cause it to become preferred. A person's predicted raw score, on the whole test or a subscale, becomes a simple average of estimated values (preferably converted into real probabilities first) which can be converted into logits for measurement purposes or retained as a subscale prediction fully comparable to a raw score.

RECOVERING THE LANGUAGE BENCHMARK

Our objective is to recover, using a language subscale consisting of 1 to 15 items, the language "benchmark" student measures that were computed with WinSteps from a 45 item language test. Figure 3 shows the correlations between the language subscale student measures computed using NOUS and the language "benchmark" measures computed using the longer test. For purposes of comparison, language measures have also been computed using several other methods:

1. WinSteps, using only the language subscale items (no math items).

2. WinSteps, combining both the language subscale items and the math items in a composite dimension.

3. ConQuest MLE. These are the student Maximum Likelihood Estimates (MLE) for the language dimension computed using the ConQuest 2-dimensional between-item model. The estimation algorithm is similar to that used by WinSteps, except that person measures are allowed to be different for the math and language dimensions.

4. ConQuest EAP. These are the student expected *a posteriori* (EAP) estimates for the language dimension computed using the ConQuest 2-dimensional between-item model. These estimates draw information from repeated samples of the student distribution.

5. Neural Network. Whereas the other methods are not specifically tailored to transfer information across dimensions, neural networks are. They work like complicated regression models in which each item becomes a predictor variable for a specified dependent variable via one or more layers of intermediate variables. In this case, the dependent variable is performance on a single language item as predicted by 14 math items (the demo version I used, Alyuda Forecaster which can be downloaded from the web, would only accommodate 15 columns at a time). I ran a parallel NOUS run for comparison purposes. It remains to do a proper comparison study with neural networks, but such a study is hindered by the constraint that neural nets will only predict one dependent variable at a time. Also, they require much larger datasets than was chosen for this small-sample study.

It should be noted that the purpose of including additional methodologies in this test is not to conduct a comparison study. The intention is to provide context and to show that these models are not identical and do not have the same objectives. For instance, although the ConQuest generalization of Rasch models includes ways to model between-item and within-item multidimensionality, this is not quite the same as information transfer across correlated dimensions (though the EAP measures do allow transfer to a significant extent). The model whose objectives most closely match those of NOUS is the neural network, but it is not yet possible to compare the two properly.

CORRELATIONS, ERRORS, AND OTHER COMPARISONS

The first set of comparisons, given in Figure 3 and Table 1 (see Appendix of Tables), looks at the correlation (not disattenuated for error) between the measures produced using each method and the language benchmark. These are "unweighted" in the sense that they have not been adjusted to account for the reliability of the person measures.

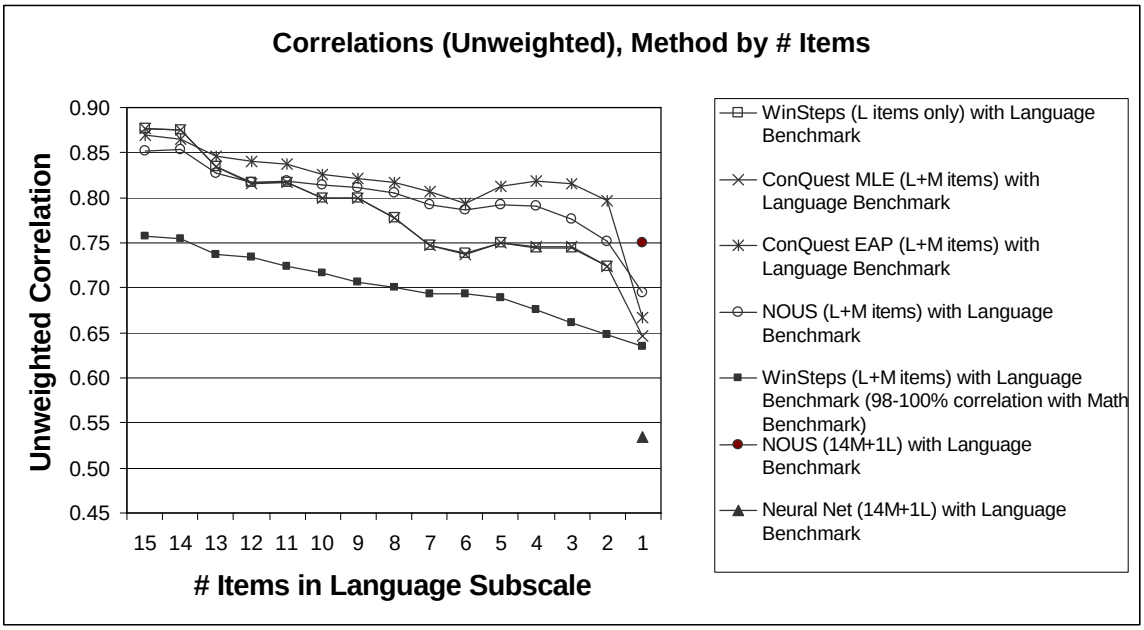
We find that all methods yield similar correlations, with the exception of WinSteps (combined language and math) and the neural network. When WinSteps combines language and math items, it constructs a composite dimension that is weighted according the number of items in each dimension. Since there are 55 math items and 15 or less language items, the combined WinSteps measures are hopelessly dominated by the math dimension ($r = 0.98 – 1.0$). The fact that its correlations are in the 0.65 – 0.75 range arises from a pre-existing 0.63 correlation between the language and math benchmark dimensions.

The neural network achieves only a 0.53 correlation with the benchmark language dimension, but it was exposed to a worst-case scenario in which there were only 14 math items as predictors and 1 language item as the dependent variable. Faced with the same scenario, NOUS manages a 0.75 correlation.

Otherwise, the methods yield similar correlations, the ConQuest EAP measures being the highest, followed by the NOUS measures. The WinSteps (language items only) and ConQuest MLE correlations are virtually identical, highlighting the fact that ConQuest in effect performs a separate WinSteps-style run for each dimension specified by the user.

The downward trend reflects the deterioration in language measures corresponding to a shrinking language subscale.

Figure 3: Unweighted Correlations with Language Benchmark (refer to Table 1)



It is soon apparent that Figure 3 is uninformative. For instance, it reports a 0.72 correlation for WinSteps (language items only) when there are only two items in the subscale. Graphing the WinSteps 2-item measures against the benchmark, we see that the relatively high correlation masks the unreliability and lack of precision of the 2-item measures (Figure 4). In other words, WinSteps is essentially restating the original dichotomous data back at us. Change a student's raw score on the two items from a 1 to a 0 and his "measure" will change accordingly. There is nothing to contradict it.

Therefore, each method needs to be assessed in light of its reliability, which is a function of the Root Mean Square Error (RMSE) of the students. Figure 5 shows the RMSE for each method. Figure 6 shows the Person Separation for each method (the ratio of the adjusted standard deviation of the measures to their RMSE). These are used to compute a Reliability coefficient (R = P.Sep.^2/(1+P.Sep^2)) which, ranging from 0 to 1, is multiplied by the unweighted correlations of Figure 3 to give the Reliability-weighted correlations shown in Figure 7.

Figure 4: Benchmark Measures vs. 2-Item Subscale Measures

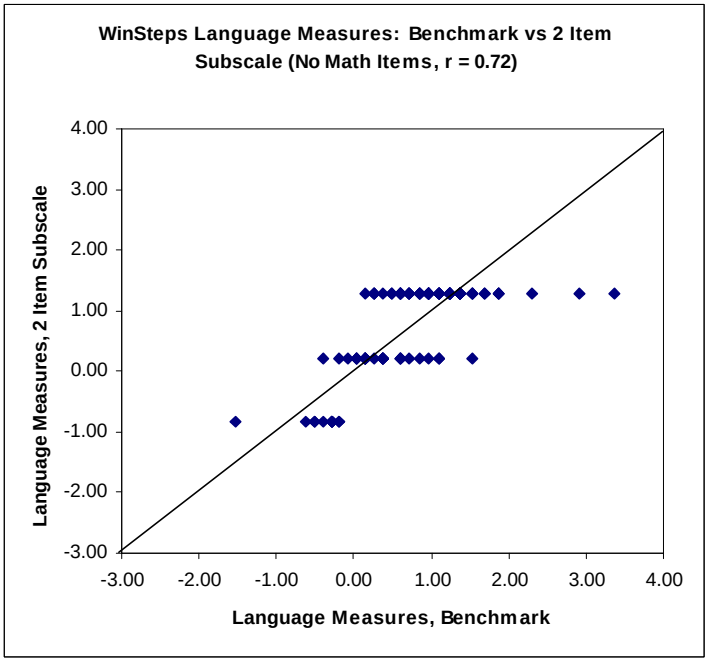
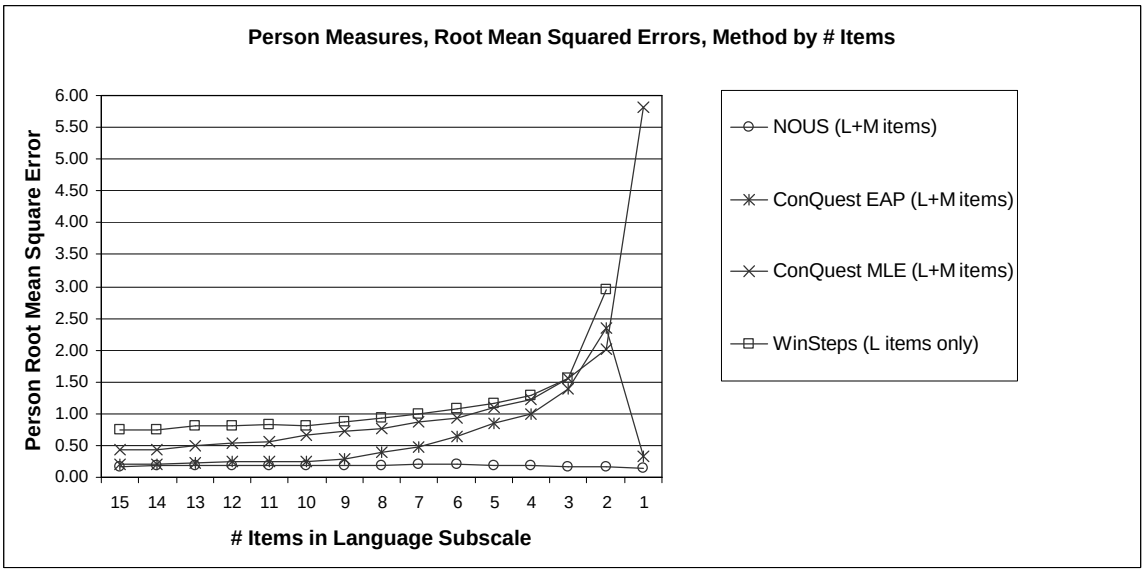


Figure 5: Root Mean Square Error across the Persons



Note: The drop in the ConQuest EAP RMSE at Items = 1 is an artifact of a shift in favor of the math dimension.

Figure 6: Person Separation, Used to Compute Reliability

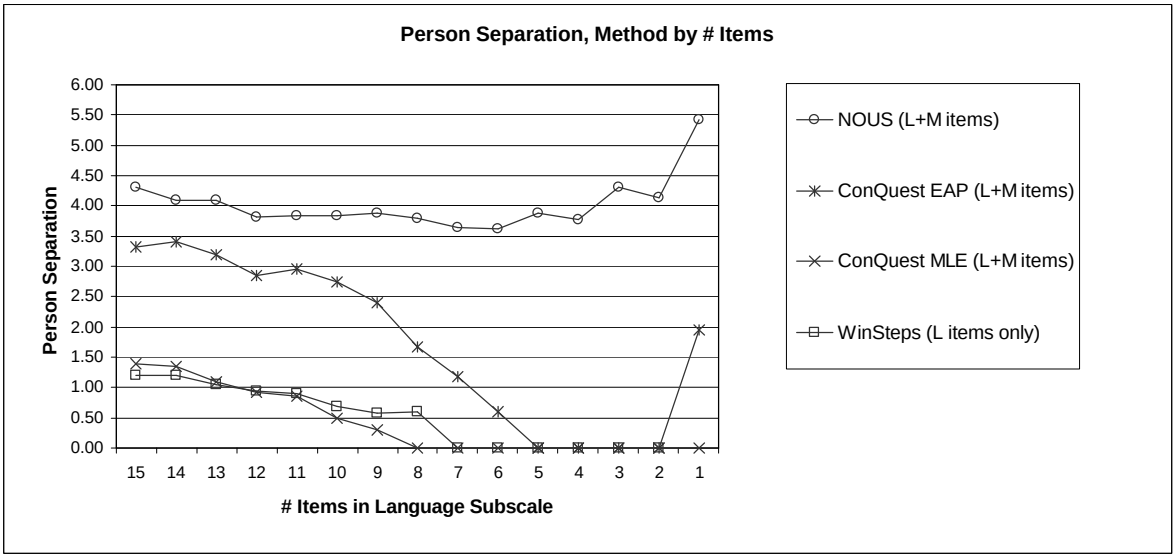
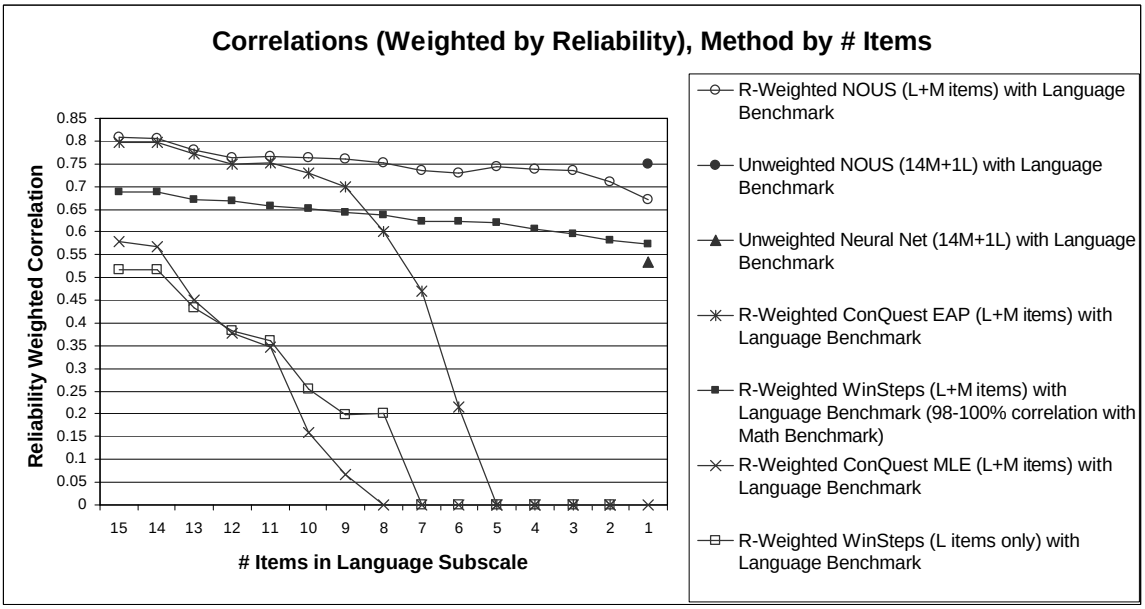


Figure 7: Correlations Weighted by Reliability



Figures 5-7 are quickly summarized. NOUS maintains a low RMSE and a high Person Separation regardless of the size of the language subscale. As a result, its measures retain a strong Reliability-weighted correlation with the benchmark language dimension as the subscale shrinks. The WinSteps combined language and math measures look respectable only because they are dominated by the math dimension, which is well-represented and reliable. The neural network appears to be the next best performer, but it is impossible to tell because it has not been weighted by reliability. It cannot offer standard error estimates at the individual person level -- only residuals. The other IRT methods suffer high error and low reliability as the subscale shrinks, the EAP measures holding out the longest because of their ability to capture some information from the math items.

What are we to make of these findings? On the surface, NOUS appears to be the most reliable and consistent of the methods considered here for subscale estimation, but this finding rests heavily on the RMSE and Separation statistics. Are the NOUS RMSE statistics comparable to the ConQuest and WinSteps statistics? Not entirely, for the NOUS RMSE is not computed as a direct function of the number of items but rather of persons. Also, it is best interpreted as a way to assess the internal consistency of the NOUS measurement structure, i.e., the degree to which different subsets of the dataset agree in their predictions for each cell, which is not necessarily the same as "reliability" in the ConQuest and WinSteps sense.

Until the errors have been made strictly comparable, we are forced to retreat to what error and reliability mean on an intuitive level. Reliability is signified by being able to delete a data value and predict it successfully from the rest of the data matrix, and being able to do this repeatedly with many data values. It is also signified by being able to enter deliberately incorrect data values and have the model stubbornly insist on the correct value. This is the type of experiment that is performed regularly with NOUS datasets, and in general a low error and a high separation corresponds to the ability to predict missing data values and correct wrong ones. However, the experiment has yet to be performed in conjunction with other methods, and it involves some labor.

CONTAMINATION FROM THE MATH DIMENSION

Before closing, there is another important comparison to perform. To what degree do NOUS estimates suffer contamination from the math dimension when constructing the language subscale? This can be assessed by comparing the correlations between the NOUS estimates and the math benchmark measures with the "true" correlation between the language benchmarks and the math benchmarks, which is 0.63. If the NOUS correlations are higher than 0.63, it signifies contamination from the math dimension. As before, the same comparisons are performed with the other methods.

Figure 8 shows that the NOUS estimates do indeed suffer a certain amount of contamination from the math dimension. Relative to the "true" correlation between language and math of 0.63, the NOUS estimates range from a maximum of 0.78 to a minimum of 0.67. What is interesting is that the pattern is not linearly related to the size of the subscale. When the subscale consists of 4 items, the subscale estimates almost succeed in shaking themselves loose from the math dimension. One could theorize that this is because the first items to be deleted are those that have a more tenuous relationship to the latent language dimension, leaving the subscale more vulnerable. The contamination decreases as we approach the hard core of the language subscale, but it increases again as this core is broken apart, leaving a single language item to define itself in the shadow of 50 math items.

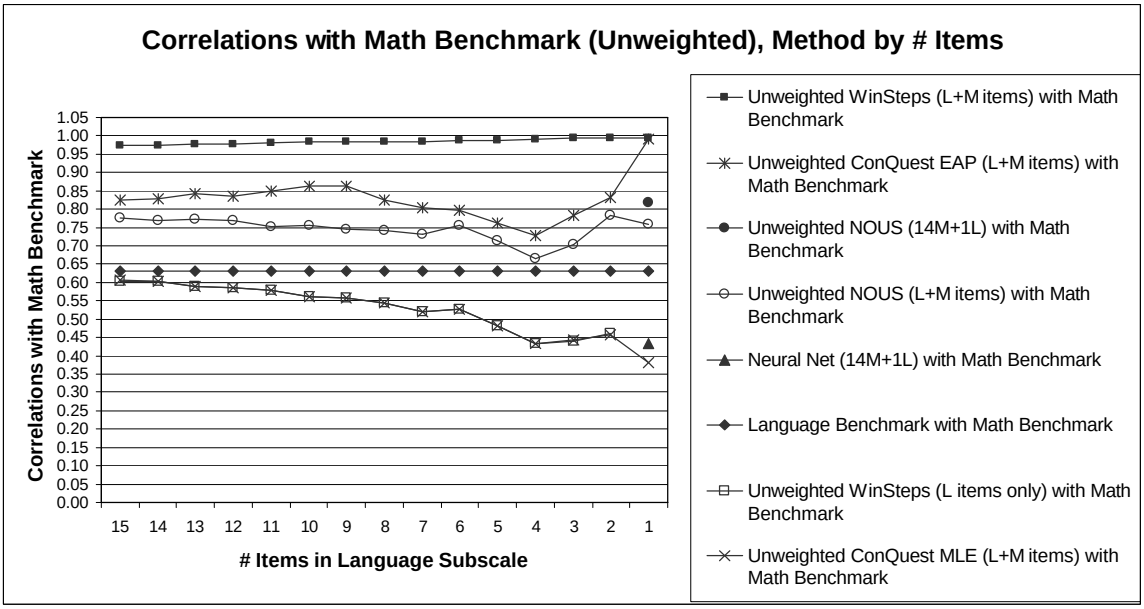
The pattern is repeated with the ConQuest EAP estimates which uniformly suffer a higher degree of correlation, finally approaching r = 0.99 with the last item, signifying a complete abdication to the math dimension. The fact that the EAP curve exceeds the 0.63 threshold proves that the EAP

language measures do in fact draw information from the mathematics items. The WinSteps (combined math and language) correlations are predictably very high, ranging from 0.98 to 1.0 due to the increasing dominance of the math dimension in the composite measure. The MLE and WinSteps (language only) estimates are fully free of the math dimension because they ignore it entirely. The neural net value also seems to be free of it, though it is hard to be sure.

Is there a theoretical reason for the NOUS estimates to be contaminated? Yes. The model specifies the computation of objective distances between person-points, and that specification has not been fully realized with this data, and is rarely perfectly realized. It appears that the prevalence of math items relative to language items has caused person-points lying in the math dimension to be somewhat overestimated relative to those lying in the language dimension. This has the effect of subtly but systematically distorting the Type-B correlations (relating item lines to person-pairs) to favor the math dimension. Based on this diagnosis, there are theoretical remedies for closing the gap, but they will require changes and additions to the algorithm and improvement is likely to be asymptotic.

Is contamination by the dominant dimension fatal in the computation of language measures? Probably not, but it depends on the extent of the contamination. On a purely pragmatic level, these language subscale measures seem to be sufficiently representative of language to serve most purposes of subscale reporting. They are certainly superior to the *ad hoc* methods psychometricians resort to at present – reporting percent items correct or assigning 1 to 3 stars for subscale performance. Nonetheless, inter-dimensional contamination is a serious issue that has yet to be sufficiently explored.

Figure 8: Correlations with the Math Dimension (Table 5)



PSYCHOMETRIC APPLICATIONS

This paper has offered some simple illustrations suggesting that an algorithm embodying a Rasch-like non-unidimensional model can recover latent dimensions using subscales that may be as narrow as a single dichotomous item. So far as I know, this cannot be done by other methods. The problem of estimating subscales, while not at the top of the IRT research agenda, is certainly a common and thorny problem in daily practice. To that extent, NOUS has shown itself worthy of further research, both in terms of understanding the geometrical paradigm of points and lines floating in an objective space and in terms of the algorithms needed to realize this paradigm.

But a brief consideration will reveal that the problem in question is in fact broader than estimating subscales. NOUS, despite its limitations, shows promise of being the IRT equivalent of multivariate regression. Any problem that can be phrased as a regression problem can also be phrased as a NOUS problem. But NOUS has this advantage which it shares with all the Rasch family of IRT models: *it rests on geometry, not statistics*. Just as the Rasch Model requires that persons occupy one and only one position in the narrow space of one dimension as a condition of fit, regardless of how the other persons and items are distributed in that space, so NOUS requires *as a condition of fit* that persons occupy one and only one position in a space of n-dimensions, regardless of how the other persons and items may be distributed in that space. Distributions are the work of statistics; positions are the work of geometry. While both Rasch and NOUS use statistics as a means to an end, perfect fit is only achieved when the geometrical ideal is met.

The geometrical paradigm of NOUS is what makes missing data designs and multiple subscale designs possible. Thus, non-unidimensional Rasch models are relevant to the following problems:

1. Multidimensional Equating. Inasmuch as test equating is defined simply as the ability to predict performance on one test from performance on another, there is no reason why test equating should be constrained to unidimensional test designs. Let Test A consist of items erecting a D-dimensional space. Then, so long as any other Test N has items in common with Test A that in combination erect the same D-dimensional space, NOUS can predict how a student who took Test N *would have performed* on Test A.

   This addresses a perennial difficulty in psychometrics – how vertically to equate grades (or instructional units within a year) that have qualitatively different contents. We construct a hypothetical meta-test containing all possible contents across all grades. Then, based on student performance on grade-specific multidimensional tests (where the latent orthogonal dimensions erect the same space as the meta-test and link to it through common items), we use NOUS to predict how the student *would* perform on the meta-test. Since all students across grades are now modeled in terms of the same meta-test, they can now be compared with each other in terms of the hypothetical sum of expected scores they would get across all its contents.

2. Educational Diagnostics. Testing is intended to serve at least two educational needs: a) measure achievement for accountability purposes; b) diagnose specific strengths and weakness of individual students. Large-scale testing serves the first need fairly well, the second need quite poorly. The non-unidimensional model makes it feasible to design tests with as many subscales as there are items and to compute individual student measures for each subscale comparable in reliability to the measures computed from large-scale assessments on a few broad dimensions. Success depends on the degree to which: a) the

dimensionality of each item is fully enfolded in the D-dimensional space erected by the remaining items on the test (which can be determined statistically); and b) the item can be related clearly to its specified latent diagnostic attribute (the problem of construct validity). Such a test, administered on a large scale or as part of an equated network of local tests, can provide diagnostic information at a level of refinement that teachers rarely see. It also invites the creation of new diagnostic variables to tease out the full multidimensional complexity of children, a complexity currently obscured by the dominance of a few composite, politically-driven dimensions.

3. Open-ended Items. Open-ended (OE) items such as essay prompts pose serious difficulties for measurement and equating. This is because it is too time-consuming and expensive to have more than a few OE items on a test and the OE dimension is often orthogonally distinct from the multiple-choice dimension. This causes a dilemma. Either the OE item is analyzed as if it were simply another MC item, in which case we lose its unique diagnostic properties, or it is analyzed separately as its own dimension, in which case we are reduced to assigning a single raw score that is both unreliable and hard to equate with other test administrations. By transferring information from the MC dimension, a non-unidimensional scaling model makes it possible to report scale score measures for the OE items that embody the OE dimension while at the same time being equated to OE items from other test administrations through common MC linking items. The equating is valid to the degree both test forms erect the same MC/OE dimensional space, even if operationalized by different items.

So essential are these applications to the continued progress of psychometrics, that NOUS-flavored Rasch models are likely to become a standard addition to the psychometrician's toolbox over the next two decades. To that end, they warrant substantial and dedicated research beyond what has been possible to date.

BIBLIOGRAPHY

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Oxford: University Press.

Moulton, M. H., (1996). *n-Dimensional Replacement: Implications of a Rasch Geometry*. Doctoral Dissertation. Chicago: University of Chicago. (Available upon request through The Andrea O'Brennan Foundation, www.aobfoundation.org)

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M.D. (1997). A Linear Logistic Multidimensional Model for Dichotomous Item Response Data. From *Handbook of Modern Item Response Theory*, W.J. van der Linden, R.K. Hambleton, Eds. New York: Springer-Verlag.

Silsdorf, H. (2001). Program Prototype for Computing NOUS Estimates Using Numeric Least-Squares Methods. San Jose: The Andrea O'Brennan Foundation.

Wright, B.D. and Masters, G.N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Wu, M.L., Adams, R.J., Wilson, M.R. (1998). *ACER ConQuest: Generalised Item Response Modeling Software*. Software Manual. Australia: Australian Council for Educational Research.

APPENDIX OF TABLES

Table 1: Unweighted Correlations with the Language Benchmark Measures by Method and # Items in Language Subscale

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
WinSteps (L items only) with Language Benchmark	0.88	0.88	0.83	0.82	0.82	0.80	0.80	0.78	0.75	0.74	0.75	0.74	0.74	0.72	.
ConQuest MLE (L+M items) with Language Benchmark	0.88	0.88	0.83	0.82	0.82	0.80	0.80	0.78	0.75	0.74	0.75	0.75	0.75	0.72	0.65
ConQuest EAP (L+M items) with Language Benchmark	0.87	0.87	0.85	0.84	0.84	0.83	0.82	0.82	0.81	0.79	0.81	0.82	0.82	0.80	0.67
NOUS (L+M items) with Language Benchmark	0.85	0.85	0.83	0.82	0.82	0.81	0.81	0.81	0.79	0.79	0.79	0.79	0.78	0.75	0.69
WinSteps (L+M items) with Language Benchmark (98-100% correlation with Math Benchmark)	0.76	0.75	0.74	0.73	0.72	0.72	0.71	0.70	0.69	0.69	0.69	0.68	0.68	0.66	0.65
NOUS (14M+1L) with Language Benchmark															0.75
Neural Net (14M+1L) with Language Benchmark															0.53

Table 2: Root Mean Square Error (RMSE) across Persons by Method and # Items in Language Subscale

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
NOUS (L+M items)	0.17	0.18	0.18	0.19	0.19	0.19	0.19	0.19	0.20	0.20	0.19	0.19	0.17	0.18	0.14
ConQuest EAP (L+M items)	0.22	0.21	0.22	0.25	0.24	0.26	0.29	0.39	0.48	0.64	0.85	1.01	1.40	2.35	0.34*
ConQuest MLE (L+M items)	0.44	0.44	0.51	0.55	0.57	0.67	0.72	0.78	0.87	0.93	1.10	1.22	1.55	2.01	5.81
WinSteps (L items only)	0.74	0.74	0.82	0.82	0.84	0.82	0.87	0.94	1.00	1.07	1.16	1.28	1.55	2.94	NA

* The ConQuest EAP RMSE is very low for Item 1. This is because it has collapsed on the language dimension and shifted to the more reliable math dimension.

Table 3: Person Separation by Method and # Items in Language Subscale

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
NOUS (L+M items)	4.30	4.09	4.08	3.82	3.83	3.84	3.88	3.78	3.64	3.62	3.87	3.78	4.31	4.13	5.43
ConQuest EAP (L+M items)	3.33	3.41	3.20	2.86	2.95	2.75	2.40	1.67	1.18	0.61	0	0	0	0	1.95*
ConQuest MLE (L+M items)	1.39	1.36	1.08	0.93	0.86	0.50	0.30	0	0	0	0	0	0	0	0
WinSteps (L items only)	1.21	1.20	1.04	0.94	0.89	0.68	0.57	0.59	0	0	0	0	0	0	NA

* The relatively high ConQuest EAP Separation is driven by the low RMSE in the previous table.

Table 4: Correlations with Language Benchmark, Weighted by Reliability, Methods by # Items in Language Subscale

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
R-Weighted NOUS (L+M items) with Language Benchmark	0.81	0.8	0.78	0.76	0.77	0.76	0.76	0.75	0.74	0.73	0.74	0.74	0.74	0.71	0.67
Unweighted NOUS (14M+1L) with Language Benchmark	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.75
Unweighted Neural Net (14M+1L) with Language Benchmark	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.53
R-Weighted ConQuest EAP (L+M items) with Language Benchmark	0.80	0.80	0.77	0.75	0.75	0.73	0.70	0.60	0.47	0.21	0	0	0	0	0.53*
R-Weighted WinSteps (L+M items) with Language Benchmark (98-100% correlation with Math Benchmark)	0.69	0.69	0.67	0.67	0.66	0.65	0.64	0.64	0.62	0.62	0.62	0.61	0.61	0.60	0.58
R-Weighted ConQuest MLE (L+M items) with Language Benchmark	0.58	0.57	0.45	0.38	0.35	0.16	0.07	0	0	0	0	0	0	0	0
R-Weighted WinSteps (L items only) with Language Benchmark	0.52	0.52	0.43	0.38	0.36	0.26	0.20	0.21	0	0	0	0	0	0	.

*The high ConQuest EAP correlation in this cell is driven by the abnormally low RMSE, an artifact of an implicit shift in favor of the mathematics dimension.

Table 5: Unweighted Correlations with Math Benchmark, Methods by # Items in Language Subscale

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Unweighted WinSteps (L+M items) with Math Benchmark	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00
Unweighted ConQuest EAP (L+M items) with Math Benchmark	0.82	0.83	0.84	0.84	0.85	0.86	0.86	0.83	0.80	0.80	0.76	0.73	0.78	0.83	0.99
Unweighted NOUS (14M+1L) with Math Benchmark	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.82
Unweighted NOUS (L+M items) with Math Benchmark	0.77	0.77	0.77	0.77	0.75	0.76	0.75	0.74	0.73	0.75	0.71	0.67	0.70	0.78	0.76
Neural Net (14M+1L) with Math Benchmark	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.43
Language Benchmark with Math Benchmark	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**
Unweighted WinSteps (L items only) with Math Benchmark	0.60	0.60	0.59	0.59	0.58	0.56	0.56	0.56	0.54	0.52	0.53	0.48	0.43	0.44	0.46
Unweighted ConQuest MLE (L+M items) with Math Benchmark	0.60	0.60	0.59	0.59	0.58	0.56	0.56	0.54	0.52	0.53	0.48	0.43	0.44	0.46	0.38