EDUCATIONAL DATA SYSTEMS, MORGAN HILL, CA

PRESENTED AT IOMW XIII, 2006
BERKELEY, CALIFORNIA

MULTIDIMENSIONAL EQUATING

LINKING MULTIDIMENSIONAL TEST FORMS BY CONSTRUCTING AN OBJECTIVE N-SPACE

MARK H. MOULTON
HOWARD A. SILSDORF

ABSTRACT

Form equating methods have proceeded under the assumption that test forms should be unidimensional, both across forms and within each form. This assumption is necessary when the data are fit to a unidimensional model, such as Rasch. When the assumption is violated, variations in the dimensional mix of the items on each test form, as well as in the mix of skills in the student population, can lead to problematic testing anomalies. The assumption ceases to be necessary, however, when data are fit to an appropriate multidimensional model. In such a scenario, it becomes possible to reproduce the same composite dimension rigorously across multiple test forms, even when the relative mix of dimensions embodied in the items on each form varies substantially. This paper applies one such multidimensional model, NOUS, to a simulated multidimensional dataset and shows how it avoids the pitfalls that can arise when fitting the same data to a single dimension. Some implications of equating multidimensional forms are discussed.

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

FOR NOUS DEMO SOFTWARE AND INFORMATION, VISIT
WWW.AOBFOUNDATION.ORG

THE PROBLEM

Test administrators at the state and local levels are under pressure to produce tests that reflect the range of content standards within a curriculum. Psychometricians are under pressure to ensure the reproducibility of test results across test forms and test administrations, which is best achieved when items embody a single, well-defined dimension. The result is a tradeoff between content validity and construct reliability. Either a test adequately embodies the curriculum but cannot be reproduced on other or subsequent test forms, or it is so narrowly focused on a few contents that it does not provide a useful or complete picture of student performance.

A number of methods have evolved to resolve this tension:

1. One test form. Apply the same test form on all occasions to all groups of examinees. This has the great advantage that the test can be of any dimensional complexity. Except for issues of differential item functioning, all students receive exactly the same test and are held to approximately the same standard. No equating methods are necessary. The price of such freedom includes the tendency for the test to lose its freshness and become easier over time, the inability to tailor the test to specific examinee populations (reducing its validity for those examinees), and a general vagueness regarding what, exactly, the test is testing.

2. Benign Neglect. Ignore the problem and assume that all test forms and items for a given content are reasonably unidimensional and can be equated. This method is surprisingly effective due largely to the happy accident that dichotomous educational data tends to be unidimensional anyway, regardless of the wide array of standards to which items are assigned. The method tends to break down when different item formats are used (multiple choice, open-ended, long text passages, etc.). It is also prone to problems with composite dimensionality.

3. Allow Composite Dimensions. This is where a test is known to have more than one dimension but is analyzed with a unidimensional model. The resulting measure is a composite, or average, of examinee performance on each of the dimensions. In itself this is not a problem. What is a problem is that composite dimensions can become quite unstable. For one thing, each test form needs to represent the same amount of each dimension, e.g., have the correct proportion of items assigned to it. This can be difficult to realize in practice, and it is hard to know whether one has succeeded until after the fact. Worse, composite dimensions can reorient themselves according to the aptitudes of the examinee population. A test calibrated on an examinee population with little variation in math ability and lots of variation in language ability will not behave the same way when applied to a population with lots of variation in math ability and little in language ability, even though it has the same items. This can cause longitudinal growth curves to lurch unpleasantly. A unidimensional model requires items in one dimension.

4. Analyze Dimensions Separately. Each content area or set of items found through various diagnostic statistics to lie in a distinct dimension is analyzed separately. This is sound methodologically, but runs up against the practical limitation that there are often too few items per dimension to yield a reliable measure, and time constraints prevent adding more items.

5. Employ Multidimensional Models. There are a number of highly regarded multidimensional models available today, both in IRT and in parallel statistical schools of thought. The model used here, NOUS, is yet another. However, the literature on how to use such models for test equating

remains sparse and they do not seem to be used widely. We propose a fairly intuitive and simple way of using NOUS to equate test forms.

THE NOUS MODEL

In 2004, NOUS (Non-Unidimensional Scaling) was introduced in the context of computing measures for item subscales by trading information across items and dimensions (Moulton, 2004). It was found to be competitive with other methods of refining subscale measures, especially as the number of items in the subscale drops below seven. The algorithm was based on a model which views persons as points floating in n-dimensional space, items as lines or measuring sticks floating in the same space, and data as approximations of where each person point projects perpendicularly onto each item line. Fit to the model, with low error, indicates that the person points and item lines are coherent and reproducible and erect an objective space. Besides subscale measurement, applications to multidimensional equating and open-ended scoring were discussed.

A defining characteristic of the model was that it is geometrical. Fit to the model implies a definite geometrical structure that transcends the observed data. In this sense, it was seen to be philosophically consonant with the Rasch paradigm, and indeed reduced to a simple Rasch-like equation.

$$G_{ni} \equiv B_{n\iota} - D_{n\iota} + \varepsilon_{ni} \qquad \text{Eq. 1}$$

The performance Gni 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.

At the time of that paper, an alternative variant of NOUS (NOUS 2004) was being developed from the same geometric paradigm but realized entirely under the rubric of linear algebra (Silsdorf, 2004). While the older version (NOUS 1996) assumed a geometric space, it did not actually locate items and persons within that space in a coordinate system. It also did not require the user to determine or specify a specific dimensionality for the dataset. The newer version does, providing linearly independent dimensional coordinates for the item and person space for a specified number of dimensions, and these coordinates can be transferred across datasets.[1]

The great advantage of a multidimensional model like NOUS is that it makes it possible to analyze datasets – e.g., to predict their missing values, spot aberrant values, and construct reproducible item and person parameters – even when their columns do not lie in a single dimension. Items can range across any number of dimensions, be expressible in many metrics, be positively or negatively correlated, even have very low correlations, so long as the number of items does not exceed the number of dimensions. (In practice, one would prefer to have at least twice as many items as dimensions.) The ability to predict missing values – ultimately the foundation of all psychometric equating designs – depends on the degree to which each item with a missing value lies in the subspace erected by the remaining items.

[1] "Linearly Independent" means that all common multiples between a set of basis vectors have been removed so that it is impossible to predict from where a given vector projects onto one basis vector where it would project onto another basis vector. This is not the same as being "orthogonal," however, for it is possible for two vectors to be linearly independent without being strictly orthogonal. For purposes of measurement and prediction, nothing is lost by requiring that basis vectors be only linearly independent.

NOUS 2004 models each data point as the Euclidean inner product of its row and column vector coordinates. Let G be the observed data matrix, B the person ability matrix, E the item difficulty matrix and ε is the difference between the observed and estimated values. Then,

$$G = B*E + \varepsilon \qquad \text{Eq. 2a}$$

or spelled out using non-linear algebra notation,[2]

$$G_{ni} = B_{nx}*E_{ix} + B_{ny}*E_{iy} + B_{nz}*E_{iz} + \ldots + B_{nd}*E_{id} + \varepsilon_{ni} \qquad \text{Eq. 2b}$$

where the performance G_{ni} of Person *n* on Item *i* is defined as Person *n*'s ability B_{nx} on the *x* dimension times the "easiness" E_{ix} of Item *i* on the *x* dimension plus Person *n*'s ability on the *y* dimension times Item *i*'s "easiness" on the *y* dimension, and so forth up to *d* dimensions, plus an error term. It will be noted that this, too, reduces to a Rasch-like formulation when only one dimension is considered.

$$G_{ni} = B_{nx}*E_{ix} + \varepsilon_{ni} = B_{nx}/D_{ix} + \varepsilon_{ni} \qquad \text{Eq. 3}$$

The person's performance is seen as the product of a person ability parameter and an item easiness parameter, or alternatively as a person ability parameter divided by an item difficulty parameter which, recast in probabilistic form, is the Rasch model for the *x* dimension. Recall that the original Rasch model was also multiplicative.

One rationale for using the Euclidean inner product to model test scores is that it is an expression of how one vector *projects* onto another vector, and projection is the geometrical correlate of measurement.[3] To measure a person's height, we *project* (drop a perpendicular from) the top of the person's head to a measuring stick. To measure his math ability, we *project* him onto a math item, so to speak. The resulting geometrical picture sees persons and items as vectors emanating from an origin in a Cartesian space of some number of dimensions, and each test score as where a given person vector projects onto a given item vector. Trigonometry tells us that the projection equals the length of the person vector *n* multiplied by the cosine of its angle with the item vector *i*. Let $\|n\|$ and $\|i\|$ represent the lengths of the two vectors from the origin to their arrow-tips. This projection is precisely analogous to Person *n*'s "ability" on the dimension defined by Item *i*.

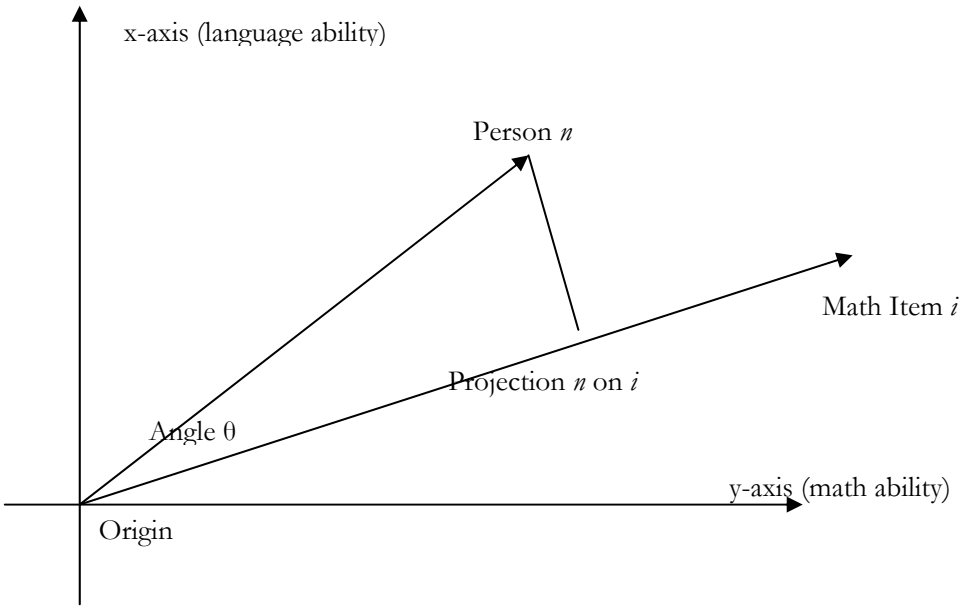
$$\text{Ability of Person } n = \text{Proj}(n \rightarrow i) = \|n\| * \cos\theta_{ni} \qquad \text{Eq. 4}$$

Figure 1 depicts what a projection onto a math item might mean for a person in a 2-dimensional space composed of math and language aptitude.

[2] Although NOUS is based on linear algebra, we use a more explicit notation in order to spell out the operations involved. "I", "E", and other symbols are not to be confused with similar symbols in linear algebra.

[3] It can be proven using theorems of linear algebra that any continuous function can be approximated by a geometric projection to any desired degree of accuracy. In addition, discrete functions can be approximated by projections in the sense that they will appear at intervals on a continuous function.

Figure 1



The notion of a measurement or test score being represented by where a person projects onto an item leaves out an important ingredient, the metric of the item. The magnitude of the test score is a function both of the person and the item. Let the item's contribution to the test score be represented by its vector length, it's "easiness" or tendency to yield a higher test score. Then, we can justify our mathematical intuition regarding the connection between fundamental measurement and the Euclidean inner product of two vectors by invoking the well-known linear algebra formula for a cosine. The cosine of the angle between any two vectors is:

$$\cos\theta_{ni} = (B_{nx}*E_{ix} + B_{ny}*E_{iy} + B_{nz}*E_{iz} + \ldots + B_{nd}*E_{id}) / (\|n\|*\|i\|) \qquad \text{Eq. 5}$$

where $\|n\|$ and $\|i\|$ are the lengths of vectors *n* and *i* from the origin and the numerator is the Euclidean inner product of the two vectors.[4]

The Euclidean inner product is found by multiplying both sides of Equation 5 by $\|n\|*\|i\|$. Suitably rearranging the left side of the equation, we have the NOUS model and a justification for claiming that fundamental measurement can indeed be modeled by a Euclidean inner product.

[4] Notice that this is nearly identical to the formula for the Pearson correlation formula, correlations being the statistical equivalent of cosines. The only difference – and it is often overlooked – is that the Pearson correlation is a true cosine only so long as each term in the numerator corresponds to a value on a linearly independent basis vector, i.e., represents a coordinate value in a Cartesian coordinate system. When correlations are computed statistically across a sample of cases (each case being another term in the numerator) this condition is rarely met, which is why it is possible to get different correlations between the same two variables by using different subsets of the sample. Fortunately, the violation of the linear independence requirement when calculating Pearson correlations can be compensated for by ensuring that the cases are a random representative sample of the target population. Loss of geometrical rigor is compensated by statistical rigor. To remove the need for statistical assumptions on the other hand, we need only ensure that each component in the numerator is on its own linearly independent dimension.

$$(\|n\| * \cos\theta_{ni}) * \|i\| = (B_{nx} * E_{ix} + B_{ny} * E_{iy} + B_{nz} * E_{iz} + \ldots + B_{nd} * E_{id}) \quad \text{Eq. 6}$$

Notice that $(\|n\| * \cos\theta_{ni})$ is equal to the projection of person vector *n* onto item vector *i* (Equation 4), which can be interpreted as the ability of Person *n* as projected onto the dimension defined by Item *I*. Therefore, each data value is modeled as the product of a person ability on a dimension specified by a given item and the item's easiness on the same dimension, given by its vector length. This is, again, a Rasch formulation, but now in terms of the dimension created by the item. Equation 3, on the other hand, derived from the right-hand side of the equation, is a Rasch formulation in terms of one of the component dimensions (*x, y, z,* ... or *d*). Thus, both sides of the equation represent a decomposition of observed scores into person ability and item easiness, a separation of parameters that is an extremely useful and important property of linear models. It may even be surmised, though a proof is not known to me, that only fit to a linear model allows for true parameter separability, i.e., special objectivity.

THE ESTIMATION ALGORITHM

The algorithm is implemented using Gaussian elimination to achieve an iterative Least Squares solution. In the version used for this paper, missing data are assigned dummy values which iterate toward their model value in conjunction with the person and item parameters. Upcoming versions of the algorithm remove the need to assign dummy values to missing cells, instead just ignoring them as is done in Rasch programs.

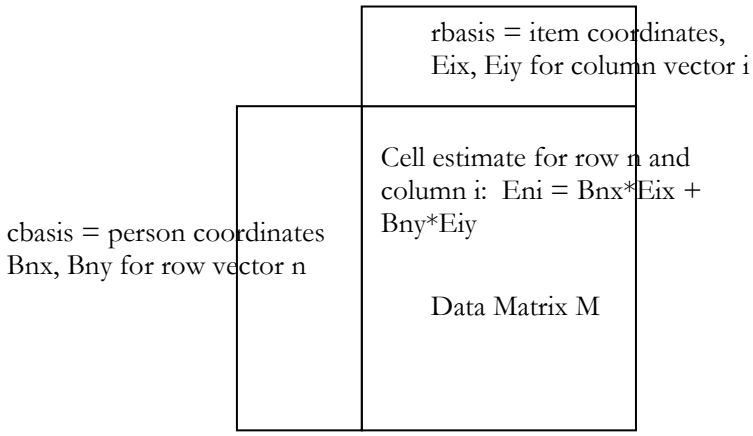
After the user specifies a dimensionality *d*, NOUS works using three matrices: the original data matrix *M*, a matrix of coordinates for each person or row called *cbasis*, (the basis for the column space of the estimates matrix), and a matrix of coordinates for each item or column called *rbasis* (the basis for the row space of the estimates matrix). All the rows in the estimate matrix are formed by linear combinations of the rows of the rbasis. All the columns in the estimate matrix are linear combinations of the columns of the cbasis. It can be proven from linear algebra that cbasis and rbasis will have the same dimension or rank because the row space and the column space are the same. An estimate for each cell is computed by multiplying the cbasis for its row by the rbasis for its column (Equation 2).

NOUS adjusts the two basis matrices to maximize the fit (minimize the squared residuals) between the observed values and the estimated values. The rbasis matrix is calculated by fixing the cbasis at arbitrary starting values and finding that rbasis which, when multiplied by the cbasis, yields estimates that best fit the observed values given the specified dimensionality. This is equivalent to performing a simple linear regression in which we are solving for one set of coefficients given another set of coefficients and the observed data. For a given column, the observations are the data values that go down that column, each of which is associated with a set of row coefficients (cbasis). Note that there is no requirement that the number of observations be the same in each column, which is why NOUS is highly robust to missing data designs. In this way, we compute a preliminary set of column coordinates (rbasis).

To compute an improved set of row coordinates (cbasis), the procedure is transposed. Now the observed values go across each row. We solve for the row coefficients taking the column coefficients (rbasis) that were calculated in the previous step as fixed. In this way, we perform an alternating least squares regression back and forth between the columns and rows until the basis matrices and estimates converge. Convergence is guaranteed regardless of the starting values of the matrices and the dummy starting values used for missing cells, which are recalculated and improved with each iteration. (Again, remember that the use of starting values for missing cells is a temporary expedient dispensed with in upcoming versions of the algorithm.)

This alternating regression is similar to the process by which row and column parameters are calculated in Rasch algorithms, especially in the handling of missing cells (once NOUS upgrades from the starting estimate method), and it is quite similar to a procedure that has been around since the late 1970's called Alternating Least Squares (ALS).[5] The main difference between the NOUS 2004 algorithm and ALS is that NOUS handles all the dimensions in a single equation for each iteration. ALS steps up through the dimensions in a sequence of 1-dimensional runs applied to the residuals matrices from the previous dimensional runs, which is similar to how WinSteps calculates principal components. There are also similarities to singular value decomposition (SVD), another method for performing row/column decomposition. One advantage of NOUS is that while it does not approach basis orthogonality as SVD does, it produces good linearly independent bases vectors with low correlations that yield estimates identical to standard ALS and SVD algorithms, but in a more efficient manner. In particular, it handles the missing data problem quite effectively. The relative merits and deficiencies of the various approaches have yet to be studied, but the empirical results are similar.

Figure 2: Layout of a 2-Dimensional Data Matrix and Associated Bases



LIMITATIONS

The NOUS 2004 software is young and suffers limitations which are worth bearing in mind:

[5] It can be shown from the theorems of Linear Algebra that ALS is guaranteed to converge to estimates that represent the smallest Euclidean distance from the observed values. NOUS exploits the same property.

1. NOUS assumes an interval (non-ordinal) metric. Ordinal data can often be analyzed *as is*, but the resulting NOUS estimates will spill outside the floor and ceiling limits. Conversion of ordinal data to logits prior to analysis is recommended.[6]

2. We have not yet programmed the calculation of standard errors for each cell estimate and person/item parameter. One useful work-around is to NOUS-analyze the residuals matrix (absolute(Observed – Estimate)) at one dimension to compute an *expected* absolute residual for each cell, which is equivalent to the standard deviation of the cell estimate assuming that the data value is not missing.

3. Missing values are assigned dummy starting values which are refined as the program iterates. This is not a problem with orderly data, but there exist types of data for which no valid geometrical solution exists for predicting a missing cell, and in these situations the choice of starting value may affect the final estimate for that cell. Upcoming versions remove this issue.

4. NOUS makes it possible to anchor a given run to a specified set of row or column coordinates. For instance, one can calibrate the items in a specified dimensional space using one data set and apply the resulting rbasis file to another data set to force the new person coordinates into the same coordinate system as those from the first data set. However, one can anchor only to an entire rbasis file as a whole, not to selected items within it, so it is not yet as flexible as item-anchoring in the WinSteps sense, though simple work-arounds exist.

5. There are circumstances when the geometrical specifications of the model are sufficiently violated that NOUS returns unlikely answers. Methods for conditioning such estimates with statistical assumptions to yield reasonable answers are being developed.

Notwithstanding these limitations, the current version works sufficiently well to yield results that are at least as useful as those produced by existing psychometric software packages.

---

[6] To convert multidimensional ordinal data sets to logits it is necessary to "logitize" the data for each column separately in a way analogous to standardization or conversion into z-scores (appropriate for interval data). The raw logit value corresponding to a given value is given by the log of the number of cases in the column expected to score below that value (for that column) divided by the number expected to score above, minus the log difficulty of the item which is the log of the number of cases at or above a specified value divided by the number below that value. When a given value is compared to other matching values, it is assumed to score better than half of them. The resulting formula for dichotomous data for column i is:

$$\text{Logit}(1)i = \ln\left((\text{count}(0)i + \text{count}(1)i/2)/(\text{count}(1)i/2) \right) - \ln\left(\text{count}(0)i/\text{count}(1)i \right), \text{ while} \qquad \text{Eq. 7}$$

$$\text{Logit}(0)i = \ln\left((\text{count}(0)i/2) \,/\, (\text{count}(1)i + \text{count}(0)i/2) \right) - \ln\left(\text{count}(0)i/\text{count}(1)i \right) \qquad \text{Eq. 8}$$

After running the raw logitized values through NOUS, convert the resulting logit estimates into expected values using the logit-to-probability formula:

$$Pni = \exp(\text{logit NOUS Estimate}) \,/\, (\, 1 + \exp(\text{logit NOUS Estimate})) \qquad \text{Eq. 9}$$

MULTIDIMENSIONAL EQUATING

All equating designs ultimately ask the question: "Given the observed performance of a student on one test form, how *would* the student have performed on a hypothetical test form administered to all students?"

It is sometimes assumed that equating is only meaningful across unidimensional tests that share a common dimension. But note that there is nothing in the definition above that specifies or assumes unidimensionality. It is merely necessary that performance on one test form be sufficient to predict performance on a hypothetical test form administered, in theory if not in practice, to all students. Let us call the test form that a student actually takes the Observed Form. Call the hypothetical test form in terms of which all students are to be compared the Reference Form. The definition above leaves open the possibility that so long as we can accurately predict how all students in a population would perform on the Reference Form based on their performance on an Observed Form, no matter how multidimensional these forms might be, the test forms are equated and we can compare students on a common metric. This allows the psychometric burden to shift from measurement to interpretability.

NOUS is designed to predict performance on a Reference Form based on performance on Observed Forms that may happen to be multidimensional. It does this by being able to predict missing values, whether they be randomly missing (more likely on adaptive test designs) or non-randomly missing (true of almost all test equating designs and especially challenging in vertical equating designs). The missing values may include the items on the Reference Form that a student does not take, or they may be the composite score for the entire Reference Form or a subset of the Reference Form. Thus, the effectiveness of any equating model depends on its ability to predict randomly and non-randomly missing data. In fact, all equating designs are missing data prediction problems at some level.

We hasten to add that there are numerous multidimensional methodologies besides NOUS such as CONQUEST, NOHARM, Singular Value Decomposition, Alternating Least Squares, Neural Networks, various Bayesian methods, and Multivariate Regression, not to mention a host of missing cell imputation techniques, which can in principle be used to predict missing cells from multidimensional data sets, or at least to compute parameters to predict missing cells. It is not the purpose of this paper to attempt comparisons at this time, but rather to show how any multidimensional model – and NOUS in particular – can be applied to this particular problem.

As stated earlier, NOUS as a *geometrical* model can predict any missing cell to the degree the item vector containing the missing cell is described by the subspace erected by the remaining items. Thus, if we are trying to predict a composite test score from a Reference Form, the Observed Form needs to contain items that capture (or exceed) the same dimensions as those that go into the composite score on the Reference Form. To the degree the Reference Form contains dimensions that are not in the Observed Form the predictions will be incorrect, for there is no geometrical solution to the problem. However, the judicious application of *statistical* assumptions may assist NOUS to yield at least a reasonable solution in such cases.

Mere ability to predict missing values is not, however, the defining or most important feature of NOUS, or the Rasch model for that matter. There are many cases where the observed values might simply be wrong. This is another way of saying that they reflect the intrusion of dimensions (such as a scoring error) that are irrelevant to the multidimensional dimensional construct(s) of the test and

should be disregarded anyway for purposes of measurement. Thus, the purpose of NOUS is less to predict actual missing cells than to erect a multidimensional measurement structure that can be applied across multiple test forms and situations and measure those aspects of a person in which we are most interested, as well as do a "reasonable" job of explaining his or her performance on our test instruments.

STEPS FOR APPLYING NOUS TO AN EQUATING PROBLEM

Before proceeding with actual examples, here is a thumbnail sketch of a possible equating procedure.

1. Define a Composite Dimension. We begin with a complete or incomplete dataset of person scores on a set of items constituting an item bank. We select a subset of items to define the dimension or composite dimension we are interested in, weighted accordingly. It could even be the entire item bank. Call this the Reference Form. The goal of the equating, then, is to predict how every student who ever takes any subsection of the test (the Observed Form) would perform on the Reference Form. The logit score of that person on the Reference Form becomes that person's measure. To the degree we can accurately predict examinee performance on the Reference Form, we can compare all examinees on the same metric – even if they have taken different forms, even if the forms are multidimensional.

2. Calibrate Item Coordinates. We run the dataset through NOUS and find the dimensionality D that best optimizes the prediction of missing cells in the Reference Form. Analysis of fit is used to clean the data relative to dimensionality D. This dimensionality is recorded for future reference. All items, plus any composite scales or subscales of interest to the user, are analyzed together to compute item, scale, and subscale coordinates, saved in the rbasis file. These are recorded along with the dimensionality and any additional item parameters that might be needed.

3. Construct New Test Forms. We construct test forms to meet the following test specification: Each must consist of items which individually or collectively represent every dimension D that is represented in the Reference Form and its associated scale and subscales. That means the items in each form should, individually or collectively, have non-zero values for each of the dimensions specified by the Reference Form. In addition, it is necessary to have a content expert review the items to make sure that all dimensions are adequately represented. Statistical criteria alone do not suffice to establish adequate dimensional coverage, but an obvious method for checking is to see how well the items on a proposed new form predict the Reference Form data when run through NOUS. Reliability is maximized by increasing the number of items on the form and targeting their difficulty on the examinee population appropriately. While the overall dimensional coverage needs to match, the proportion of items assigned to each dimension does not need to match that of the Reference Form. In addition, none of the items on the new form need to belong to the Reference Form, so long as they all are calibrated together.

4. Administer Test Forms. Each test form is administered and scored for a sample of students to be measured on a common scale or subscale.

5. Compute Person Coordinates. We run the new data through NOUS anchoring the items at their pre-calibrated coordinate values. NOUS calculates person coordinates as well as estimates for how each person would have performed on the Reference Form. Analysis of fit is performed and the person coordinates and Reference Form estimates are adjusted accordingly. Again, note that our prediction of performance on the Reference Form does not rely on the form's having Reference Form items.

6. Compute Person Measures. Sum the person estimates for the Reference Form, convert into a percentage, and compute a logit measure for how that person would have performed on the Reference Form. Alternatively, simply predict the composite score on the Reference Form directly, as well as any subscale scores. These are the equated measures. Note that we are not reporting the person's coordinate measures, though such is theoretically possible. The reason is that such coordinates are abstract and hard to interpret, whereas the Reference Form is tangible and easy to explain and represents what the test administrator means by the test, regardless of its dimensional richness. Person measures can always be related back to expected performance on the Reference Form. As long as the Reference Form is held constant, the person measures are comparable. An additional reporting metric involves computing person measures using Equation 6, in which the cell estimates are divided by the length of the Item vector (its easiness). This effectively removes item easiness from the reported results.

The use of NOUS to equate test forms is still in its infancy and optimal procedure have yet to be determined.

4-DIMENSIONAL SIMULATED DATA

To illustrate how an equating procedure might work, we step through a simulated 4-dimensional equating problem. See Tables 1-??. The data set was created by populating a 4-dimensional rbasis file and cbasis file with random numbers and computing their Euclidean inner product. The resulting numbers were rounded to add a little statistical noise. Data were deleted to mimic a situation in which 30 students are administered Form A and Form B. Although this is simulated data, similar procedures have been applied to real data with similar results.

STEP 1: ASSEMBLE DATA

Table 1 provides the data matrix. There are 30 "students" down the rows, each of whom took either Form A or Form B. There are 30 "items" across the top, separated into subscales A, B, and C, representing different types of contents. We see that each student took 20 out of the 30 items, but a different 20 depending on which form they were given. Off to the right is space for an as-yet non-existent Reference Form which will consist solely of the average scores for subscales A, B, and C, plus an average score across all 30 items. They are called composite scores. Because each row contains missing data, we are unable to calculate the composite scores in a way that holds each student to the same standard.

STEP 2: RUN INPUT DATA THROUGH NOUS TO ASSESS DIMENSIONALITY

Table 2 (actually a graph) depicts how well the program predicts the values of known cells made missing for each of 10 dimensions. Two statistics are graphed: a) the mean absolute residual between the true value and the predicted value; b) the variance of the true values that is explained by the predicted values. Both statistics tell the same story. This dataset is most accurately modeled with four dimensions, which was expected given that we generated the data with a 4-dimensional cbasis and rbasis.

Another important finding is that NOUS at four dimensions is doing a good job of predicting missing cells, even with large blocks of the dataset missing. The mean residual is 0.12 out of a scale that ranges across 10 points or so. The variance explained is 0.95. Therefore, we conclude that it is reasonably safe to accept NOUS's predictions for the truly missing cells as reasonably close to the true, or most likely, values. This is important because it is the foundation of our claim that we can compare students who take different test forms. We are saying, in effect, that we can predict how they will do on items they have not in fact taken.

STEP 3: RERUN NOUS AT DIMENSION 4 TO OBTAIN CELL ESTIMATES

Having found that four dimensions are optimal to model this data, we run NOUS at that dimensionality to compute values for missing cells and replace the observed data values with modeled values. This yields a fully populated matrix. Since all students now have "data" on each item, we can calculate subscale averages and overall averages to create a "Reference Form" based on which all students will henceforth be compared.

STEP 4: RERUN NOUS AT DIMENSION 4 ON CELL ESTIMATES PLUS THE REFERENCE FORM

By rerunning NOUS on the estimates matrix plus the newly populated Reference Form, we can compute coordinate values for each item, including those for each of the four Reference Form variables. The resulting matrix is called rbasis. This is exactly analogous to an "item anchor" file from WinSteps, except that it is in four dimensions rather than one. It is this file that makes it possible to equate the two forms and compare students without having to rerun NOUS on the whole data set every time a new student is added.

STEP 5: RERUN NOUS ON THE ORIGINAL DATA, PLUS THE BLANK REFERENCE SET, USING THE ANCHORED RBASIS

Now all the items, including the composite variables on the Reference Form, have fixed coordinate values that can be applied to the original data matrix. They can not only be used to fill in the missing cells of the items the students did not take, they can even fill in the missing cells of the Reference Form – even though the columns are completely devoid of data! NOUS performs this remarkable calculation simply by multiplying the stored rbasis values for the composite variables by the cbasis values computed for each person based on his observed data. It is these predicted values on the Reference Form for each composite variable that becomes the student's measure and diagnostic profile, overall and for the three special contents labeled A, B, and C. This makes it possible to compute a series of comparable person measures based on his or her scores on any subset of items in the 30-item bank – so long as these items adequately embody the four dimensions of this particular achievement space.

But does it work? Can NOUS really calculate reasonable measures for the Reference Form based only on some student data and the rbasis. To answer this question, we compare the new Reference Form measures computed from the original data and rbasis, much of the data missing, with the original Reference Form numbers we computed in Step 3 based on a full matrix. If the two are in substantial agreement, we can conclude that the rbasis values are sufficient to equate students who take different forms and can be used henceforth for new students.

STEP 6: COMPARE REFERENCE FORM MEASURES USING THE TWO METHODS

We find that the Reference Form measures are in substantial agreement, $r = 0.98$. The measures computed using an anchored rbasis are sufficient to equate students who take different forms. While the new measures do not have quite the same standard deviation as those based on the fully populated matrix (whose predictions we trust based on the findings in Step 2), this does not matter since all students receive their measures using the anchored rbasis method. It is enough that they have a strongly linear relationship.

Table 1: Input Data, 30 students by 30 items, 2 forms, no reference form

		Subscales										Item Bank																				Reference Form			
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	Composite Scores			
Form		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	A	B	C	All
A	1	6	3	6	6	7	5	3	2	5	7	2	5	3	4	5	4	6	6	3	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	2	2	2	4	4	5	4	2	2	3	4	2	3	3	2	4	4	3	5	3	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	3	3	3	2	3	3	3	2	3	3	4	2	5	3	2	5	2	4	4	2	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	4	5	3	4	4	5	3	2	2	3	5	3	5	3	3	5	3	5	5	3	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	5	4	2	3	3	3	2	2	1	3	4	2	4	2	3	4	2	4	3	2	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	6	3	4	3	4	4	2	2	2	1	4	4	4	3	1	4	4	4	4	4	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	7	4	3	3	4	5	4	2	3	4	5	2	5	3	4	6	2	5	6	2	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	8	5	2	6	6	7	6	3	2	5	7	2	5	4	4	5	4	6	6	3	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	9	5	4	5	6	6	4	3	2	3	6	4	5	4	3	5	5	6	5	4	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	10	5	3	4	5	6	5	3	4	6	7	2	6	4	5	8	2	6	7	3	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	11	3	3	2	3	4	3	2	2	2	4	3	5	3	2	5	3	4	4	3	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	12	4	3	4	5	5	4	2	2	2	5	3	4	3	2	5	5	5	5	4	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	13	5	2	6	6	7	5	3	2	4	6	2	4	3	4	5	4	5	6	3	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	14	4	2	4	5	6	5	2	3	5	6	2	5	3	4	6	3	5	6	3	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	15	5	4	5	6	6	4	3	2	3	6	4	5	4	3	6	5	6	6	4	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.
B	16	.	.	.	.	.	.	.	.	.	.	2	4	4	4	6	4	5	7	3	6	8	3	6	6	6	5	6	4	8	3	.	.	.	.
B	17	.	.	.	.	.	.	.	.	.	.	3	4	4	3	5	4	5	6	3	5	7	2	6	5	6	5	5	4	6	3	.	.	.	.
B	18	.	.	.	.	.	.	.	.	.	.	3	7	5	5	8	5	7	9	4	6	10	4	8	7	7	6	8	6	8	5	.	.	.	.
B	19	.	.	.	.	.	.	.	.	.	.	3	3	3	2	4	3	4	4	3	4	5	1	5	3	5	4	4	3	4	3	.	.	.	.
B	20	.	.	.	.	.	.	.	.	.	.	2	4	2	3	4	1	4	3	2	2	4	3	4	4	3	1	4	4	3	3	.	.	.	.
B	21	.	.	.	.	.	.	.	.	.	.	3	4	3	2	4	5	5	5	4	4	6	2	6	4	5	4	4	3	5	3	.	.	.	.
B	22	.	.	.	.	.	.	.	.	.	.	3	5	3	3	5	3	5	5	3	3	6	3	6	5	5	3	5	5	4	4	.	.	.	.
B	23	.	.	.	.	.	.	.	.	.	.	3	6	3	3	6	3	6	5	3	3	7	4	6	5	5	3	6	5	4	5	.	.	.	.
B	24	.	.	.	.	.	.	.	.	.	.	1	3	2	2	3	2	4	4	2	3	5	2	4	3	3	3	4	3	4	2	.	.	.	.
B	25	.	.	.	.	.	.	.	.	.	.	2	4	3	3	5	2	4	5	2	3	6	2	4	4	5	4	5	4	5	2	.	.	.	.
B	26	.	.	.	.	.	.	.	.	.	.	2	4	3	3	4	4	4	5	3	4	6	2	5	4	4	4	4	3	5	3	.	.	.	.
B	27	.	.	.	.	.	.	.	.	.	.	3	5	3	3	5	4	5	5	3	4	7	3	6	4	5	4	5	4	5	4	.	.	.	.
B	28	.	.	.	.	.	.	.	.	.	.	2	4	4	3	5	4	4	6	3	5	7	2	5	5	5	5	5	3	7	2	.	.	.	.
B	29	.	.	.	.	.	.	.	.	.	.	4	6	4	3	7	5	6	6	4	5	8	3	8	5	7	5	7	6	6	5	.	.	.	.
B	30	.	.	.	.	.	.	.	.	.	.	2	4	2	2	5	2	4	4	2	2	5	3	4	4	4	2	4	4	3	3	.	.	.	.

Table 2: Graph showing which level of dimensionality does the best job of predicting the true values of cells made missing

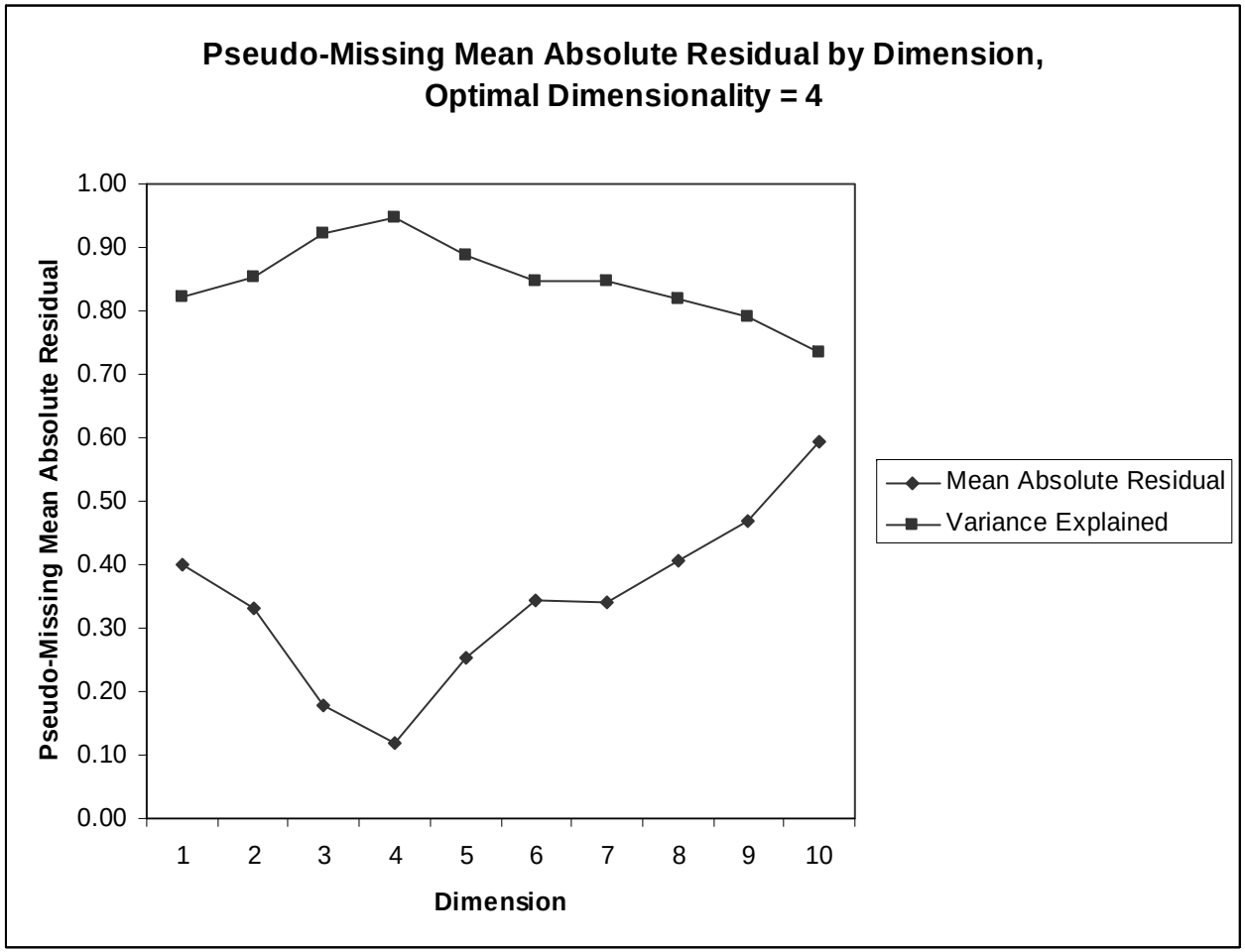


Table 3: NOUS estimates for each cell, missing and non-missing. The Reference Form numbers are averaged from the estimates as a whole and for each subscale

		Subscales										Item Bank										Predicted values										Reference Form			
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C			Composite Scores	
Form		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	A	B	C	All
A	1	5.9	2.5	6.1	6.1	6.9	5.0	3.0	1.8	4.8	6.9	2.2	5.0	3.3	4.3	5.2	4.0	6.0	6.0	3.2	4.7	7.5	4.2	6.2	6.0	4.0	3.3	5.3	3.9	6.2	4.1	4.4	4.4	5.5	4.8
A	2	2.1	1.9	3.7	4.3	5.0	4.2	1.8	2.1	2.7	4.2	2.1	2.9	3.1	2.0	3.9	3.7	3.4	4.8	2.8	4.3	5.5	0.9	4.6	3.3	4.8	4.6	3.9	2.6	5.3	1.9	3.0	3.5	3.7	3.4
A	3	2.9	2.8	1.9	2.9	3.4	2.8	1.7	2.7	2.8	4.0	2.4	4.5	2.8	2.4	5.3	1.9	4.0	4.4	2.3	2.6	5.5	2.3	4.6	3.8	5.1	3.2	5.1	4.6	3.8	3.1	3.1	3.2	3.9	3.4
A	4	4.8	3.2	3.7	4.3	4.6	3.1	2.4	1.9	3.2	5.2	2.8	4.9	2.8	3.0	5.1	3.1	5.2	4.7	2.9	3.3	6.2	3.3	5.7	4.6	4.5	2.9	5.0	4.4	4.2	4.1	3.8	3.7	4.4	4.0
A	5	4.2	2.3	2.9	3.1	3.3	2.1	1.8	1.2	2.7	4.1	1.8	3.9	1.8	2.7	3.8	1.8	4.1	3.3	1.9	2.0	4.5	3.2	4.0	3.8	2.6	1.3	3.7	3.5	2.7	3.3	2.8	2.5	3.4	2.9
A	6	3.1	3.8	2.9	3.9	3.9	2.3	2.0	1.8	0.9	3.8	3.9	4.0	3.1	1.1	4.3	4.4	4.2	3.9	3.8	3.9	5.4	0.8	6.2	2.6	6.1	4.6	4.3	3.7	3.6	3.6	3.5	3.7	3.4	3.5
A	7	4.0	2.6	3.1	3.9	4.8	4.0	2.2	3.0	4.4	5.3	2.0	5.2	3.1	3.7	6.1	2.0	4.9	5.5	2.3	3.2	6.7	3.5	5.0	5.3	4.9	3.2	5.9	5.1	5.0	3.5	3.6	3.7	5.0	4.1
A	8	5.2	2.2	6.1	6.1	7.2	5.6	2.9	2.2	5.0	6.8	2.0	4.7	3.5	4.2	5.3	4.1	5.6	6.4	3.1	5.0	7.7	3.8	5.9	5.9	4.3	3.9	5.3	3.7	6.7	3.5	4.3	4.5	5.5	4.8
A	9	5.2	3.8	5.1	5.8	6.0	3.9	2.9	1.9	2.8	6.0	3.8	5.1	3.6	2.8	5.2	5.1	5.8	5.5	4.2	4.9	7.2	2.6	7.3	4.6	5.9	4.6	5.3	4.3	5.4	4.5	4.5	4.6	4.9	4.7
A	10	5.0	3.0	3.9	4.9	6.1	5.3	2.8	3.9	6.0	6.8	2.1	6.5	3.9	4.9	7.7	2.2	6.1	7.1	2.6	3.9	8.5	4.7	6.0	6.9	5.9	3.8	7.5	6.4	6.5	4.2	4.5	4.6	6.4	5.2
A	11	3.1	3.3	2.3	3.3	3.6	2.6	1.9	2.4	2.2	4.0	3.0	4.4	2.9	2.0	5.0	2.9	4.1	4.3	2.9	3.1	5.5	1.8	5.3	3.4	5.5	3.8	4.9	4.4	3.7	3.4	3.3	3.4	3.8	3.5
A	12	3.7	3.3	4.2	5.0	5.3	3.7	2.4	2.0	2.3	5.0	3.4	4.2	3.4	2.1	4.6	4.7	4.7	5.0	3.8	4.7	6.4	1.5	6.4	3.7	5.8	4.9	4.7	3.6	5.1	3.5	3.9	4.2	4.2	4.1
A	13	4.8	2.1	5.9	5.9	6.8	5.2	2.8	1.8	4.4	6.3	2.0	4.2	3.3	3.8	4.6	4.2	5.2	5.9	3.0	4.9	7.1	3.3	5.7	5.3	4.0	3.8	4.8	3.2	6.3	3.3	4.1	4.2	5.1	4.5
A	14	4.1	2.3	4.2	4.8	5.9	5.0	2.5	3.0	4.8	5.9	1.9	4.9	3.5	3.9	5.9	2.8	5.0	6.2	2.6	4.1	7.3	3.4	5.3	5.6	4.9	3.9	5.8	4.5	6.1	3.2	3.9	4.1	5.3	4.4
A	15	5.1	4.0	5.0	5.8	6.1	4.1	2.9	2.3	3.1	6.1	3.9	5.4	3.9	2.9	5.7	5.1	5.9	5.8	4.2	5.0	7.6	2.6	7.4	4.8	6.4	4.9	5.7	4.7	5.7	4.5	4.7	4.8	5.2	4.9
B	16	3.8	2.0	5.5	6.1	7.5	6.5	2.8	3.2	5.3	6.6	1.9	4.5	4.1	4.1	5.9	4.1	5.1	7.2	3.1	5.6	8.2	2.8	5.9	5.9	5.6	5.3	6.0	4.0	7.8	2.8	4.3	4.8	5.8	5.0
B	17	3.5	2.7	4.5	5.2	6.0	4.9	2.4	2.6	3.6	5.5	2.7	4.3	3.7	2.9	5.3	4.2	4.8	5.9	3.4	4.9	7.1	2.0	6.0	4.6	5.8	5.0	5.3	3.9	6.2	3.1	4.0	4.4	4.8	4.4
B	18	5.8	3.7	6.1	7.0	8.3	6.7	3.5	3.8	5.9	8.1	3.3	6.7	4.9	4.9	7.9	4.8	7.1	8.4	4.1	6.1	10.1	4.2	8.0	7.3	7.3	5.8	7.9	6.1	8.3	4.7	5.6	5.9	7.2	6.2
B	19	2.7	2.7	3.0	3.8	4.1	3.0	1.8	1.8	1.9	3.9	2.7	3.5	2.8	1.7	4.0	3.6	3.7	4.1	2.9	3.6	5.2	1.2	5.0	3.0	4.9	4.0	4.0	3.2	4.1	2.7	3.1	3.3	3.4	3.3
B	20	3.9	2.3	2.3	2.7	3.0	2.0	1.7	1.6	2.8	3.9	1.7	4.1	1.9	2.7	4.1	1.4	4.0	3.3	1.7	1.7	4.5	3.2	3.8	3.8	2.9	1.3	4.0	3.8	2.7	3.2	2.7	2.5	3.4	2.9
B	21	4.2	2.9	4.8	5.2	5.5	3.7	2.5	1.4	2.4	5.1	3.1	3.9	3.1	2.3	4.0	4.7	4.8	4.7	3.6	4.6	6.1	1.9	6.1	3.8	4.8	4.2	4.1	3.1	4.9	3.5	3.8	4.0	4.1	4.0
B	22	4.6	3.4	3.4	4.1	4.4	3.0	2.3	2.1	3.1	5.1	2.9	5.1	2.9	2.9	5.3	3.0	5.2	4.7	2.9	3.2	6.2	3.1	5.8	4.5	4.9	3.1	5.2	4.7	4.1	4.1	3.8	3.7	4.5	4.0
B	23	5.4	3.8	3.6	4.3	4.6	2.9	2.6	2.2	3.4	5.6	3.1	5.9	3.0	3.4	5.9	2.9	5.8	5.0	3.1	3.1	6.7	3.9	6.3	5.1	5.0	2.8	5.8	5.4	4.1	4.8	4.1	3.9	4.9	4.3
B	24	2.8	1.6	3.1	3.4	4.0	3.2	1.7	1.7	2.8	3.9	1.5	3.0	2.3	2.4	3.6	2.4	3.4	3.9	2.0	3.0	4.7	2.0	3.7	3.5	3.3	2.7	3.6	2.7	4.0	2.2	2.6	2.8	3.4	2.9
B	25	2.3	2.0	2.4	3.4	4.3	4.0	1.7	3.0	3.6	4.2	1.7	3.9	3.0	2.8	5.2	2.1	3.6	5.1	2.1	3.3	5.8	2.0	4.2	4.1	5.0	3.8	5.1	4.0	4.9	2.3	3.0	3.4	4.2	3.5
B	26	3.7	2.2	4.5	4.8	5.4	4.1	2.2	1.7	3.1	5.0	2.2	3.6	2.9	2.7	4.1	3.8	4.3	4.9	2.9	4.2	5.9	2.2	5.1	4.1	4.2	3.8	4.1	2.9	5.1	2.8	3.5	3.7	4.1	3.8
B	27	4.6	3.2	4.3	4.9	5.3	3.7	2.5	2.0	3.1	5.4	3.0	4.7	3.2	2.9	5.0	3.9	5.2	5.1	3.3	4.1	6.6	2.8	6.1	4.5	5.0	3.8	5.0	4.1	4.9	3.9	4.0	4.0	4.6	4.2
B	28	2.9	1.7	4.8	5.3	6.5	5.7	2.3	2.7	4.3	5.5	1.8	3.6	3.7	3.2	4.9	3.9	4.2	6.2	2.9	5.1	7.0	1.9	5.1	4.8	5.2	5.1	5.0	3.2	6.9	2.2	3.7	4.2	4.8	4.3
B	29	5.1	4.4	4.4	5.5	5.9	4.1	2.9	2.9	3.4	6.3	4.2	6.2	4.2	3.1	6.7	4.7	6.3	6.2	4.3	4.9	8.1	2.9	7.8	5.2	7.2	5.2	6.7	5.8	5.7	4.9	4.9	5.0	5.6	5.2
B	30	3.5	2.5	2.2	2.8	3.2	2.3	1.7	2.0	2.7	3.9	2.0	4.2	2.2	2.5	4.5	1.7	4.0	3.7	2.0	2.1	4.9	2.7	4.2	3.7	3.8	2.1	4.4	4.1	3.1	3.2	2.8	2.8	3.6	3.1

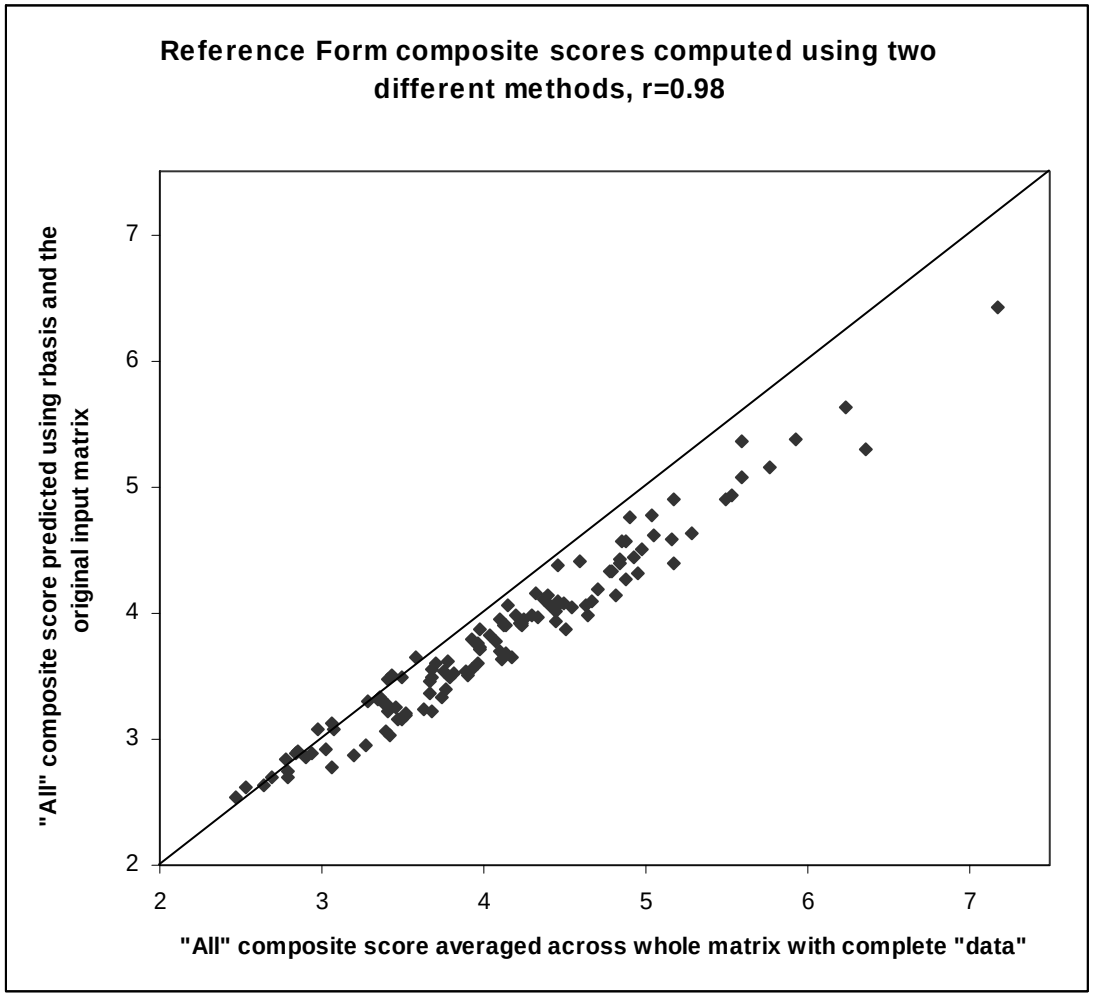
rbasis = coordinates for each item vector

	Subscales												Item Bank																		Reference Form			
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	Composite Scores			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	A	B	C	All
Dim 1	0.6	0.4	0.6	0.7	0.8	0.7	0.3	0.4	0.6	0.8	0.3	0.7	0.5	0.5	0.8	0.5	0.7	0.8	0.4	0.6	1.0	0.4	0.8	0.7	0.7	0.6	0.8	0.6	0.8	0.5	0.6	0.6	0.7	0.6
Dim 2	0.0	-0.8	0.2	-0.1	0.2	0.5	0.0	0.1	1.0	0.2	-1.0	-0.2	-0.2	0.7	0.0	-0.8	-0.2	0.2	-0.7	-0.3	0.0	0.7	-0.8	0.6	-1.0	-0.7	0.0	-0.2	0.4	-0.5	-0.2	-0.2	0.2	-0.1
Dim 3	1.0	0.4	0.1	0.0	-0.2	-0.6	0.2	-0.3	0.0	0.3	0.2	0.6	-0.2	0.3	0.2	-0.2	0.6	-0.3	0.0	-0.5	0.0	0.8	0.3	0.3	-0.4	-0.8	0.1	0.5	-0.6	0.7	0.2	-0.1	0.1	0.1
Dim 4	0.0	-0.4	1.0	0.7	0.8	0.6	0.1	-0.4	0.0	0.2	-0.2	-0.6	0.0	-0.1	-0.7	0.8	-0.1	0.1	0.2	0.7	0.0	-0.3	0.0	-0.1	-0.4	0.3	-0.6	-1.0	0.6	-0.4	0.0	0.1	-0.1	0.0

Table 4: Original Data to be re-analyzed using the anchored rbasis values given in the first four rows

		Subscales										Item Bank																				Reference Form			
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	Composite Scores			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	A	B	C	All
		0.6	0.4	0.6	0.7	0.8	0.7	0.3	0.4	0.6	0.8	0.3	0.7	0.5	0.5	0.8	0.5	0.7	0.8	0.4	0.6	1.0	0.4	0.8	0.7	0.7	0.6	0.8	0.6	0.8	0.5	0.6	0.6	0.7	0.6
		0.0	-0.8	0.2	-0.1	0.2	0.5	0.0	0.1	1.0	0.2	-1.0	-0.2	-0.2	0.7	0.0	-0.8	-0.2	0.2	-0.7	-0.3	0.0	0.7	-0.8	0.6	-1.0	-0.7	0.0	-0.2	0.4	-0.5	-0.2	-0.2	0.2	-0.1
		1.0	0.4	0.1	0.0	-0.2	-0.6	0.2	-0.3	0.0	0.3	0.2	0.6	-0.2	0.3	0.2	-0.2	0.6	-0.3	0.0	-0.5	0.0	0.8	0.3	0.3	-0.4	-0.8	0.1	0.5	-0.6	0.7	0.2	-0.1	0.1	0.1
Form		0.0	-0.4	1.0	0.7	0.8	0.6	0.1	-0.4	0.0	0.2	-0.2	-0.6	0.0	-0.1	-0.7	0.8	-0.1	0.1	0.2	0.7	0.0	-0.3	0.0	-0.1	-0.4	0.3	-0.6	-1.0	0.6	-0.4	0.0	0.1	-0.1	0.0
A	1	6.0	3.0	6.0	6.0	7.0	5.0	3.0	2.0	5.0	7.0	2.0	5.0	3.0	4.0	5.0	4.0	6.0	6.0	3.0	5.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	2	2.0	2.0	4.0	4.0	5.0	4.0	2.0	2.0	3.0	4.0	2.0	3.0	3.0	2.0	4.0	4.0	3.0	5.0	3.0	4.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	3	3.0	3.0	2.0	3.0	3.0	3.0	2.0	3.0	3.0	4.0	2.0	5.0	3.0	2.0	5.0	2.0	4.0	4.0	2.0	3.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	4	5.0	3.0	4.0	4.0	5.0	3.0	2.0	2.0	3.0	5.0	3.0	5.0	3.0	3.0	5.0	3.0	5.0	5.0	3.0	3.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	5	4.0	2.0	3.0	3.0	3.0	2.0	2.0	1.0	3.0	4.0	2.0	4.0	2.0	3.0	4.0	2.0	4.0	3.0	2.0	2.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	6	3.0	4.0	3.0	4.0	4.0	2.0	2.0	2.0	1.0	4.0	4.0	4.0	3.0	1.0	4.0	4.0	4.0	4.0	4.0	4.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	7	4.0	3.0	3.0	4.0	5.0	4.0	2.0	3.0	4.0	5.0	2.0	5.0	3.0	4.0	6.0	2.0	5.0	6.0	2.0	3.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	8	5.0	2.0	6.0	6.0	7.0	6.0	3.0	2.0	5.0	7.0	2.0	5.0	4.0	4.0	5.0	4.0	6.0	6.0	3.0	5.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	9	5.0	4.0	5.0	6.0	6.0	4.0	3.0	2.0	3.0	6.0	4.0	5.0	4.0	3.0	5.0	5.0	6.0	5.0	4.0	5.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	10	5.0	3.0	4.0	5.0	6.0	5.0	3.0	4.0	6.0	7.0	2.0	6.0	4.0	5.0	8.0	2.0	6.0	7.0	3.0	4.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	11	3.0	3.0	2.0	3.0	4.0	3.0	2.0	2.0	2.0	4.0	3.0	5.0	3.0	2.0	5.0	3.0	4.0	4.0	3.0	3.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	12	4.0	3.0	4.0	5.0	5.0	4.0	2.0	2.0	2.0	5.0	3.0	4.0	3.0	2.0	5.0	5.0	5.0	5.0	4.0	5.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	13	5.0	2.0	6.0	6.0	7.0	5.0	3.0	2.0	4.0	6.0	2.0	4.0	3.0	4.0	5.0	4.0	5.0	6.0	3.0	5.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	14	4.0	2.0	4.0	5.0	6.0	5.0	2.0	3.0	5.0	6.0	2.0	5.0	3.0	4.0	6.0	3.0	5.0	6.0	3.0	4.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A	15	5.0	4.0	5.0	6.0	6.0	4.0	3.0	2.0	3.0	6.0	4.0	5.0	4.0	3.0	6.0	5.0	6.0	6.0	4.0	5.0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
B	16	.	.	.	.	.	.	.	.	.	.	2.0	4.0	4.0	4.0	6.0	4.0	5.0	7.0	3.0	6.0	8.0	3.0	6.0	6.0	6.0	5.0	6.0	4.0	8.0	3.0	.	.	.	.
B	17	.	.	.	.	.	.	.	.	.	.	3.0	4.0	4.0	3.0	5.0	4.0	5.0	6.0	3.0	5.0	7.0	2.0	6.0	5.0	6.0	5.0	5.0	4.0	6.0	3.0	.	.	.	.
B	18	.	.	.	.	.	.	.	.	.	.	3.0	7.0	5.0	5.0	8.0	5.0	7.0	9.0	4.0	6.0	10.0	4.0	8.0	7.0	7.0	6.0	8.0	6.0	8.0	5.0	.	.	.	.
B	19	.	.	.	.	.	.	.	.	.	.	3.0	3.0	3.0	2.0	4.0	3.0	4.0	4.0	3.0	4.0	5.0	1.0	5.0	3.0	5.0	4.0	4.0	3.0	4.0	3.0	.	.	.	.
B	20	.	.	.	.	.	.	.	.	.	.	2.0	4.0	2.0	3.0	4.0	1.0	4.0	3.0	2.0	2.0	4.0	3.0	4.0	4.0	3.0	1.0	4.0	4.0	3.0	3.0	.	.	.	.
B	21	.	.	.	.	.	.	.	.	.	.	3.0	4.0	3.0	2.0	4.0	5.0	5.0	5.0	4.0	4.0	6.0	2.0	6.0	4.0	5.0	4.0	4.0	3.0	5.0	3.0	.	.	.	.
B	22	.	.	.	.	.	.	.	.	.	.	3.0	5.0	3.0	3.0	5.0	3.0	5.0	5.0	3.0	3.0	6.0	3.0	6.0	5.0	5.0	3.0	5.0	5.0	4.0	4.0	.	.	.	.
B	23	.	.	.	.	.	.	.	.	.	.	3.0	6.0	3.0	3.0	6.0	3.0	6.0	5.0	3.0	3.0	7.0	4.0	6.0	5.0	5.0	3.0	6.0	5.0	4.0	5.0	.	.	.	.
B	24	.	.	.	.	.	.	.	.	.	.	1.0	3.0	2.0	2.0	3.0	2.0	4.0	4.0	2.0	3.0	5.0	2.0	4.0	3.0	3.0	3.0	4.0	3.0	4.0	2.0	.	.	.	.
B	25	.	.	.	.	.	.	.	.	.	.	2.0	4.0	3.0	3.0	5.0	2.0	4.0	5.0	2.0	3.0	6.0	2.0	4.0	4.0	5.0	4.0	5.0	4.0	5.0	2.0	.	.	.	.
B	26	.	.	.	.	.	.	.	.	.	.	2.0	4.0	3.0	3.0	4.0	4.0	4.0	5.0	3.0	4.0	6.0	2.0	5.0	4.0	4.0	4.0	4.0	3.0	5.0	3.0	.	.	.	.
B	27	.	.	.	.	.	.	.	.	.	.	3.0	5.0	3.0	3.0	5.0	4.0	5.0	5.0	3.0	4.0	7.0	3.0	6.0	4.0	5.0	4.0	5.0	4.0	5.0	4.0	.	.	.	.
B	28	.	.	.	.	.	.	.	.	.	.	2.0	4.0	4.0	3.0	5.0	4.0	4.0	6.0	3.0	5.0	7.0	2.0	5.0	5.0	5.0	5.0	5.0	3.0	7.0	2.0	.	.	.	.
B	29	.	.	.	.	.	.	.	.	.	.	4.0	6.0	4.0	3.0	7.0	5.0	6.0	6.0	4.0	5.0	8.0	3.0	8.0	5.0	7.0	5.0	7.0	6.0	6.0	5.0	.	.	.	.
B	30	.	.	.	.	.	.	.	.	.	.	2.0	4.0	2.0	2.0	5.0	2.0	4.0	4.0	2.0	2.0	5.0	3.0	4.0	4.0	4.0	2.0	4.0	4.0	3.0	3.0	.	.	.	.

Table 5: Does the rbasis method of computing Reference Form measures work? Yes.



IMPLICATIONS

By being able to predict missing cells by locating persons and items in an objective multidimensional space, NOUS enables a wide variety of new equating designs. One such design was presented in this paper. Other equating designs include:

- Multidimensional Computer Adaptive Testing. Instead of storing 1-dimensional item parameters, the CAT algorithm would store the full D-dimensional profile of each item to students. This profile would make it possible to compute the student's probability of success on each item as well as locate that student most efficiently in that space.

- Benchmark Equating. Currently, little is being done to equate the benchmark exams that are at the center of formative assessments administered at the district level, largely due to their multidimensional complexity over time. NOUS offers several equating designs that would make it possible to measure student growth within and across grades on any number of diagnostic variables simultaneously. It even makes it possible to equate exams over time without common items – a complex topic for another paper.

In addition, by allowing rigorous control of multidimensional data sets, NOUS invites test writers to design richer tests to capture a broader array of educational standards without sacrificing comparability across test forms and administrations. It raises afresh the possibility of true vertical equating. Vertical equating has traditionally been hampered by the fact that many educational content areas (history and science, for example) do not lie within a well-defined unidimensional construct but lurch from dimension to dimension across the grades. By erecting a multidimensional space and a Reference Form that includes all these dimensions, either aggregated or broken out by individual contents, it is theoretically possible to track movement through that space solely in terms of predicted performance on the Reference Form, regardless of changes in content across grades. The practical challenge is making sure that every test form contains the same dimensionality as the Reference Form.

In light of these findings, there seems to be some justification for pursuing the implications of NOUS-based equating models.

REFERENCES

References to linear algebra are based on well-established theorems that can be found in most linear algebra textbooks and on the internet. For further information regarding Rasch Models, Alternating Least Squares, Singular Value Decomposition, and their relation to NOUS, please contact the authors. Information regarding NOUS and its applications, as well as software and source code, can be obtained from the authors or found on www.eddata.com and www.aobfoundation.org.

Mark Moulton can be contacted using the information on the cover page.