Measuring the Power of Learning.®

Research Report
ETS RR–17-29

Unidimensional Vertical Scaling in Multidimensional Space

James E. Carlson

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Unidimensional Vertical Scaling in Multidimensional Space

James E. Carlson

Educational Testing Service, Princeton, NJ

In this paper, I consider a set of test items that are located in a multidimensional space, S_M, but are located along a curved line in S_M and can be scaled unidimensionally. Furthermore, I am demonstrating a case in which the test items are administered across 6 levels, such as occurs in K–12 assessment across 6 grade levels, and for which a unidimensional vertical scale can be developed. I am limiting my coverage to dichotomously scored items because the models are much simpler than those for polytomously scored items. However, the concepts discussed can be extended to the latter type of item. I also limit my demonstrations to a 2-dimensional space, S_2, so I can geometrically represent my points in a 2-dimensional representation in this article. These concepts can also be extended to a higher-dimensional case.

Keywords Assessment; dimensionality; item response theory; linking; vertical scaling

doi:10.1002/ets2.12157

I start by discussing the concepts and models that are used to represent items that are measuring proficiency in an *M*-dimensional proficiency space (S_M). I then introduce the fact that it is theoretically possible for a set of items to be tightly located along a nonlinear curve in S_M and that such a curve is mathematically unidimensional. Hence, such a set of items is measuring a unidimensional proficiency variable and can be modeled with a unidimensional item response theory (IRT) model. I then show that this is feasible using an artificial dataset generated to represent six levels of test takers using a two-dimensional case. The first demonstration includes calibrating each level independently followed by linking them as usually done in vertical scaling. Later I expand the analysis demonstration by showing results for two different concurrent calibration procedures. The procedures I am demonstrating are easily generalized to a case of higher dimensionality, and I discuss how this is possible. The main motivation for these demonstrations is the assumption, often stated by testing professionals, that a one-dimensional model cannot adequately represent the results of testing in the same educational subject area across a span of multiple grades. Patz and Yao (2007), for example, stated, "When calibrating items from multiple test forms for the purpose of measuring students across a range of grade levels, the IRT assumption of unidimensionality would appear implausible" (p. 260). Briggs and Weeks (2009) stated, "[I]f the dimensional structure changes from test to test over time … the scores along a unidimensional vertical scale will be biased in the sense that they are not measuring the multidimensional construct of interest, but some sort of composite factor" (p. 12). Kolen and Brennan (2014) stated, "One of the most challenging aspects of applying IRT to vertical scaling is the assumption that the same unidimensional ability is assessed across grades" (p. 469) and concluded, "[m]ore research on psychometric structure across grades and on the use of multidimensional IRT in vertical scaling is needed" (p. 469). The demonstration in this paper addresses this last issue somewhat, although not necessarily definitively.

Conceptual Background

Using a Euclidean space, items that are measuring multiple dimensions can, theoretically, be located using location parameters in a multidimensional coordinate space, S_M. Reckase (1989) pointed out the following:

> In principle, test items can be developed to distinguish between persons on each one of the coordinate system dimensions, or combinations of the dimensions. However, a particular test may not contain items that are sensitive to differences on all of the dimensions, or examinees may not differ across all tested dimensions. Therefore, the

Corresponding author: J. Carlson, E-mail: jcarlson@ets.org

> dimensionality of the data generated by the administration of a test to the population of interest may not have the same dimensionality as the full θ-space. (p. 3)

Reckase went on to state that a single item provides measurement in a single dimension (direction) in the space, and the interaction of test takers with the item provides item response data that divide the space into two parts: one with test takers primarily having correct scores and one with those primarily having incorrect scores. That single dimension will typically be some linear combination of the underlying dimensions represented by any set of axes in the coordinate space, and the direction of that dimension in the space represents "the direction of greatest rate of change from incorrect responses to correct responses" (p. 5). This direction is referred to here as the *direction of measurement* (DOM). With respect to a set of items on a test, Reckase (1989) pointed out the following:

> If the direction of maximum rate of change at each point in the space differs across items … the dimensionality of the data generated by the interaction of the population and the test is equal to the dimensionality of the space needed to contain all of the directions specified. (p. 6)

One of the purposes of this paper is to demonstrate a case in which this statement is not true. Reckase (1989) does not consider in his publications (e.g., Reckase 1985, 1989; Reckase & McKinley, 1983, 1991) that a set of test items can, theoretically, be located along a curved line (which is mathematically unidimensional) in an M-dimensional space, and a population of test takers can, also theoretically, be located near that same curved line, and hence those items can be measuring those test takers on a unidimensional proficiency variable. Although Carlson (2001) has demonstrated this fact with several examples in a one-population situation, in this paper more details and the demonstration that such a situation can occur across multiple levels of tests and populations of test takers (e.g., grade levels in school) are provided.

Item Response Theory Models

I first review unidimensional item response theory (UIRT) models and parameterizations of them, then show the generalization to multidimensional (MIRT) models. Although most readers are probably very familiar with UIRT (and probably also MIRT) models, the reason for this review is to relate the unidimensional concepts to the multidimensional, with similar terminology, notation, and parameterizations, to set up the main topic of curved unidimensional proficiency scales in a multidimensional space.

Unidimensional Item Response Theory Models

There are many equivalent ways in which UIRT models have been expressed and parameterized. The commonly used three-parameter logistic (3PL) model may be expressed as

$$P_{ij} = P\left(x_{ij} = 1 | a_i, b_i, c_i, \theta_j\right) = c_i + \left(1 - c_i\right) \frac{e^{f_{ij}}}{1 + e^{f_{ii}}} = c_i + \left(1 - c_i\right) \left(e^{-f_{ij}} + 1\right)^{-1}. \quad (1)$$

$$f_{ij} = a_i \left(\theta_j - b_i\right). \quad (2)$$

In Expressions 1 and 2, P_{ij} represents the probability of a correct response ($x_{ij} = 1$) on item i, at θ_j, a value in the proficiency dimension that varies over the entire space (in the unidimensional case a line theoretically varying from negative infinity to positive infinity); a_i is the item slope (discrimination) parameter; b_i the location (difficulty) parameter; c_i the lower asymptote (pseudo guessing) parameter.

A constant multiplier, $D = 1.7$, with the sole function of making the 3PL as close as possible to the similar normal (Gaussian) model (see, e.g., Birnbaum, 1968, p. 399) is often added to the right side of the exponent. This constant has no effect other than changing the metric of the parameters and their estimates, so to simplify, I am not including it here. The logistic is typically preferred to the normal model because of its mathematical simplicity (Birnbaum, 1968, p. 400). The exponent in Expression 2 is sometimes parameterized in the slope-intercept form as

$$f_{ij} = a_i \theta_j + d_i, \quad (3)$$

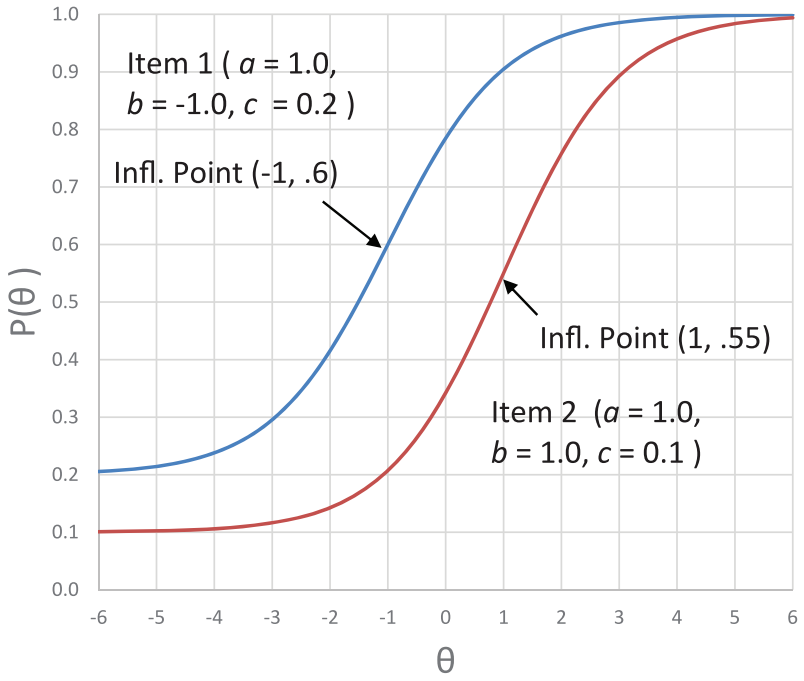

Figure 1 Item characteristic curves for two 3PL items showing inflection points.

where, comparing Expressions 2 and 3, clearly

$$d_i = -a_i b_i. \tag{4}$$

For completeness of coverage, I note that b_i and θ_j are both values of the proficiency metric that is being assessed by the item. Different ways of expressing and parameterizing Expression 1 are used in different publications. Simpler UIRT models include the two-parameter logistic (2PL) model in which all items are assumed to have a lower asymptote of zero, yielding (from Expression 1)

$$P_{ij} = P\left(x_{ij} = 1 | a_i, b_i, \theta_j\right) = \left(e^{-f_{ij}} + 1\right)^{-1}. \tag{5}$$

and one-parameter logistic (1PL, Rasch) models in which all items are assumed to have the same a parameter, which can be set to 1.0 without loss of generality (Birnbaum, 1968, p. 402), so that Expressions 1 and 2 can be written as

$$P_{ij} = P\left(x_{ij} = 1 | b_i, \theta_j\right) = \left(e^{-f_{ij}} + 1\right)^{-1}$$
$$f_{ij} = \theta_j - b_i. \tag{6}$$

Figure 1 displays two 3PL item response curves (IRCs, also called *item characteristic curves*) having different parameters, as shown in the figure. The IRC shows how the probability of a correct response (vertical axis) increases with the value of the proficiency variable, θ (horizontal axis). The IRCs of 3PL items have an inflection point at a θ value equal to the b-parameter value with a probability value halfway between the lower and upper asymptotes,

$$P_{ij} = \left(1 + c_i\right)/2. \tag{7}$$

For the 2PL and 1PL models, the c parameter is set to zero, so the lower asymptote is at zero and the inflection point, still at $\theta = b$, has a probability value of .5; note that this value of b at the inflection point is still equal to $(1 + c)/2$ because this expression is equal to one half with $c = 0$.

Although most writers refer to the j subscript as a test-taker subscript, this is really not necessary. The θ_j simply vary over the entire range of θ values (theoretically from negative infinity to positive infinity) so that Expression 1 is best thought of as the expression for the probability value at any specific value, θ_j, on the horizontal axis, which can be that of many (or no) test takers.

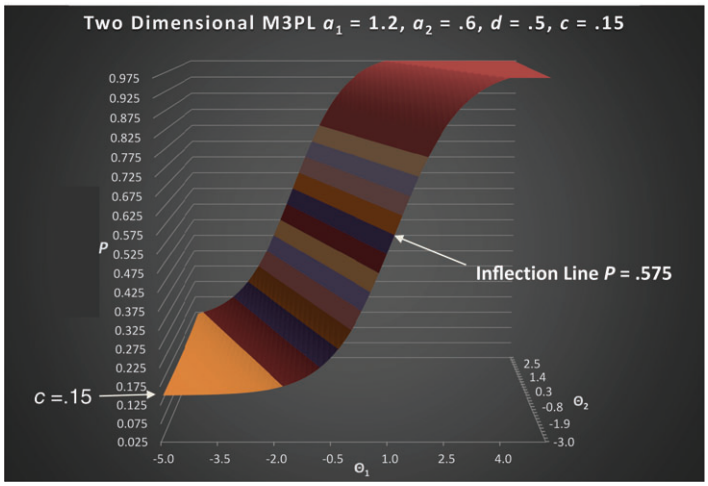

Figure 2 Two-dimensional M3PL IRS with parameters shown and inflection line at $P = .575$.

Multidimensional Item Response Theory Models

As in the case of the UIRT models, MIRT models are expressed by different authors in a number of different ways. One common way is the generalization of Expression 1 to the multidimensional 3PL model (M3PL) in a space, S_M, with M dimensions. The model is

$$P_{ij} = P\left(x_{ij} = 1 | \mathbf{a}_i, d_i, c_i, \boldsymbol{\theta}_j\right) = c_i + \left(1 - c_i\right) \frac{e^{f_{ij}}}{1 + e^{f_{ij}}} = c_i + \left(1 - c_i\right) \left(e^{-f_{ij}} + 1\right)^{-1}. \tag{8}$$

$$f_{ij} = \mathbf{a}_i' \boldsymbol{\theta}_j + d_i = \sum_{m=1}^{M} a_{im} \theta_{jm} + d_i, \tag{9}$$

where $\mathbf{a}_i$ is a vector of M slope (discrimination) parameters, one related to each dimension; d_i is a location parameter; c_i is the lower asymptote of the item response surface (IRS), similar to that of UIRT models; and $\boldsymbol{\theta}_j$ is a vector of M proficiency values, with elements, θ_{jm}, each a value on one of the M proficiency variables. Geometrically $\boldsymbol{\theta}_j$ contains the coordinates of a point in S_M.

Figure 1 displays the IRS for a two-dimensional M3PL item with parameters shown. Also shown is the line of inflection across the surface at $P = .575$ (the line between the orange and blue shaded areas). This line shows the set of points having probability of correct response exactly halfway between the lower and upper asymptotes of $c = .15$ and 1.0, respectively. Therefore, the probability values along the line are

$$P_{ij} = \left(1 + c_i\right)/2 = .575, \tag{10}$$

a generalization of Expression 7. The subscript j in Expression 10 can be thought of as identifying the set of all possible points along a line representing an orthogonal (perpendicular) projection from the inflection line onto S_M. Because the figure displays a two-dimensional case, of course $M = 2$.

As mentioned previously for the 3PL, the j subscript in Expressions 8 and 9 does not necessarily have to indicate a specific test taker. Given the M-dimensional θ space (S_M), we can think of $\boldsymbol{\theta}_j$ as the jth of the infinite number of unique vectors that can occur in S_M. In the two-dimensional case illustrated in Figure 2, this is S_2, the "floor" of the figure with axes θ_1 and θ_2, both theoretically ranging from minus to plus infinity. Different test takers may have the same values on the M θ variables and hence the same vector. Different vectors may have different values of the probability of a correct response, the perpendicular distance between S_M and the IRS (above the "floor" in S_2), but, as can be seen by examination of the IRS in Figure 2, there are lines of constant probability so any vector representing a point with an orthogonal projection from such a line has the same probability value.

And each $\boldsymbol{\theta}_j$ vector does not necessarily have any test takers with its specific elements. Hence the IRS is perhaps best thought of as the locus of probabilities, as computed from the model parameters, of correct responses at the locations

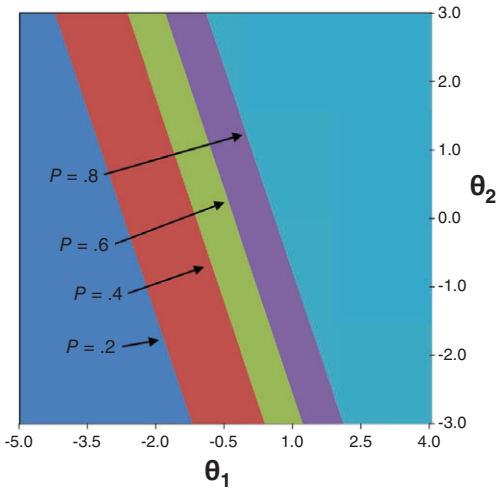

Figure 3 Contour plot of the item response surface in Figure 3.

of all possible unique $\boldsymbol{\theta}_j$ vectors. In the two-dimensional case illustrated, each such vector has two elements that are the coordinates of a point in S_2 which, as mentioned previously, is the floor in Figure 2. For individual test takers, of course, given the values in their $\boldsymbol{\theta}$ vectors, we could compute their probabilities of achieving a correct response on the item. And given sample data from the interaction of N test takers with each item on a test, we can estimate the model parameters and use them to estimate the test takers' proficiencies.

Multidimensional Item Difficulty

Reckase (1985) defined MID in terms of the direction in the θ space of the steepest slope of the item's IRS, referred to here as the DOM of the item. The DOM is defined as a direction of nondecreasing probabilities because we assume that as θ_1 and/or θ_2 increase, the probability of a correct response to the item is nondecreasing. Referring to Figure 2, it should be clear that as we consider moving along different straight lines in S_2 (the floor of the figure), the orthogonal distances of the IRS from S_2 (which are the probabilities) may be increasing, remaining steady, or decreasing. The DOM must be in a direction in which the probability is nondecreasing and the slope of the IRS is a maximum compared to other directions as we move along the line in S_2 defining that direction. Looking at the figure, it should be clear that there are many such lines parallel to each other. In this paper I denote such a line that passes through the origin (coordinates of $\theta_1 = 0, \theta_2 = 0$) as the DOM, θ_i^*. It represents a proficiency variable that is a linear combination of θ_1 and θ_2, defining the proficiency measured by item i and its direction. The DOM is perpendicular to the orthogonal projection of the inflection line onto S_2.

The inflection line is so referred to because, if you consider all points along that line, as we move in a positive direction along the DOM, the slope is increasing before that line is reached and decreasing after it is reached; in Figure 2 that line has $P = .575$, as indicated. Orthogonally projecting the lines of constant probability of the IRS in Figure 2 onto the two-dimensional θ space (S_2, the floor in Figure 2) yields the contour plot displayed in Figure 3. In that figure, the orthogonally projected lines for probabilities of .2, .4, .6, and .8 are shown. Hence the projection of the inflection line is slightly toward the .4 line from the .6 line.

For the two-dimensional case, the MID is the signed distance between the orthogonal projection of the inflection line onto S_2 and the origin ($\theta_1 = 0, \theta_2 = 0$). In the general case in S_M, it is given by

$$D_i = \frac{-d_i}{\sqrt{\sum_{m=1}^{M} a_{im}^2}} = b_i^*, \tag{11}$$

which I denote as b_i^* because it can be considered a location parameter of item i along the DOM, θ^*. It is the location parameter analogous to the b parameter in the one-dimensional model, and for the item being used to illustrate, it is

computed from Expression 11 to have a value of −.373. It is the distance from the origin to the inflection line, positive if in the first quadrant and negative if in the third. An important aspect of the MIRT models, as mentioned previously, is that the IRS must be such that the probabilities represented in the IRS are nondecreasing in the positive directions of all M proficiency dimensions and also in the dimension defined by θ_i^*, the DOM. In the two-dimensional case in the illustrations herein, that means that the location defined by the DOM and the D_i or b_i^* parameter must always be in the first or third quadrant of S_2.

Reckase (1985) defined the MID direction in terms of the angles, ω_{im}, between the DOM of item i and each of the M θ axes. These angles are the elements of the vector $\boldsymbol{\omega}_i$, in which ω_{im}, can be determined from

$$\cos\omega_{im} = \frac{a_{im}}{\sqrt{\sum_{m=1}^{M} a_{im}^2}}, \quad m = 1, \ldots, M. \tag{12}$$

Reckase (1985) also provided an expression for the exponent in Expression 9 in terms of polar coordinates in S_M as

$$f_{ij} = \sum_{m=1}^{M} a_{im}\theta_{jm}\cos\left(\omega_{jm}\right) + d_i. \tag{13}$$

I introduce this parameterization here because I use it later in my demonstration of curvilinear unidimensional proficiency. I should point out that Reckase typically discussed the M2PL model (c parameter set to 0.0) rather than the M3PL, although he did mention that model briefly (Reckase, 1989).

Multidimensional Item Discrimination

Reckase and McKinley (1983, 1991) discussed discrimination in M2PL models and the slope of the IRS in the measurement direction at the inflection line; the slope was derived by Reckase (1985) as

$$Slope = \frac{1}{4}\sum_{m=1}^{M} a_{im}\cos\omega_{im} = \frac{1}{4}\sum_{m=1}^{M}\left(a_{im}\frac{a_{im}}{\sqrt{\sum_{m=1}^{M} a_{im}^2}}\right) = \frac{1}{4}\sqrt{\sum_{m=1}^{M} a_{im}^2}. \tag{14}$$

For the M3PL, the slope formula has a multiplicative factor of $(1-c_i)$. For the unidimensional 2PL model, Expression 14 reduces to the slope at the point of inflection being equal to $(1/4)a_i$, as would be expected. For the 3PL, it is $(1/4)a_i(1-c_i)$. The multidimensional item discrimination (MDISC) was defined (Reckase, 1986, 1989) as

$$\text{MDISC}_i = \sqrt{\sum_{m=1}^{M} a_{im}^2} = a_i^*, \tag{15}$$

which I denote as a_i^* because it represents a slope (discrimination) parameter for the proficiency variable, θ_i^* measured by the item and is analogous to the a parameter in the unidimensional model. For the illustrative item, it is computed from Expression 15 to have a value of 1.342. Note also, from Expressions 11 and 15,

$$b_i^* = \frac{-d_i}{a_i^*}, \tag{16}$$

which I use later in developing the concept of the curvilinear unidimensional proficiency in a multidimensional space.

Relation of Parameters to the Proficiency Space

I will use an alternative definition of location parameters in place of the Reckase D (my b^*), which is a point in S_M defined by distance in the measurement direction (defined by $\boldsymbol{\omega}_i$). I define location as the rectangular coordinates of that point in

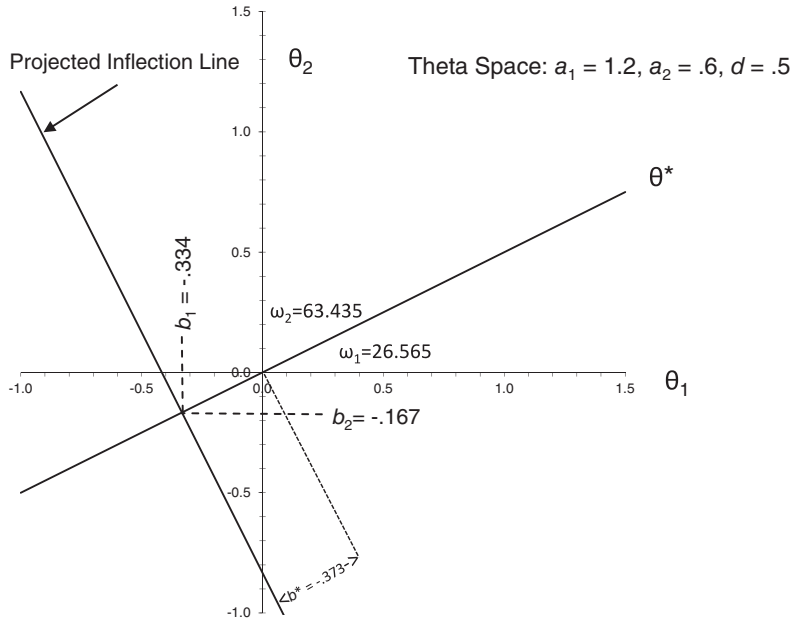

Figure 4 Two-dimensional proficiency space showing geometric interpretation of parameters for the item response surface in Figure 3.

S_M. The coordinate along dimension m is given by Reckase (1985) as

$$\theta_{im} = D_i \cos\omega_{im} = b_i{}^* \cos\omega_{im} = \frac{-d_i a_{im}}{\sum_{m=1}^{M} a_{im}^2} = b_{im}, \tag{17}$$

where, again, I denote these coordinates as b_{im} because they represent location parameters with respect to the M dimensions of the proficiency space. It is important to note that for any item characterized by the MIRT model, these b parameters must all have the same sign. The reason is that, as discussed above, the IRS must be nondecreasing in the proficiency variables, and for this to be true the bs must be in the first or third quadrants, so must have the same sign. This restriction is discussed further below in conjunction with discussion of characteristics of the proficiency space. Note also, from Expressions 15, 16, and 17,

$$b_{im} = \frac{-d_i a_{im}}{\left(a_i{}^*\right)^2} = \frac{b_i{}^* a_{im}}{a_i{}^*}, \tag{18}$$

which I use later in my development of the curvilinear unidimensional proficiency concept. As stated previously, the location of an item can be described in terms of the relationship of the parameters of an item to lines and points in the proficiency space, S_M. I limit this discussion to the two-dimensional space because in S_2 the proficiency space is a plane that is easily represented in a figure. Figure 4 is used to illustrate. In what follows, I omit the item subscript, i, because I am only dealing with one item, as shown in the figures.

As shown, θ_1 is the horizontal axis and θ_2 the vertical axis. The figure illustrates the following:

1. The line θ^*, which is in the proficiency measured by the item and is in the DOM in the proficiency space.
2. The orthogonal projection of the inflection line onto the space.
3. The location parameters b_1 and b_2 as distances along θ_1 and θ_2, respectively (as computed from Expression 18).
4. The b^* as the signed distance from the origin to the projected inflection line along θ^* (computed from Expression 16).

Note that for this item, b^* is negative because it is in the negative direction (along the DOM) from the origin of S_2. Using trigonometry or Pythagoras's theorem, it is easily shown that

$$b^* = \text{sign}\left(b_1\right)\sqrt{b_1^2 + b_2^2} = \text{sign}\left(b_2\right)\sqrt{b_1^2 + b_2^2}. \tag{19}$$

A plane orthogonal to S_2 located along the θ_1 axis cuts the IRS in Figure 3 in a curve that is a 3PL response curve with parameters a_1, b_1, and c; similarly, a plane orthogonal to S_2 along the θ_2 axis cuts the IRS in a 3PL response curve with parameters a_2, b_2, and c. The DOM is given by the angles defined in Expression 12. For the two-dimensional item in the figures, the measures of the angles between the DOM and the two axes are

$$\omega_1 = \arccos\left(a_1/a^*\right) = \arccos\left(1.2/1.342\right) = 26.565^\circ, \tag{20}$$

and

$$\omega_2 = \arccos\left(a_2/a^*\right) = \arccos\left(.6/1.342\right) = 63.435^\circ. \tag{21}$$

Clearly in the two-dimensional case, because the axes are orthogonal, the measures of these two angles must sum to 90°, and they do. Using trigonometry, one can see from the figure that

$$\begin{aligned} b_1 &= b^* \cos\omega_1, \\ b_2 &= b^* \sin\omega_1. \end{aligned} \tag{22}$$

It is also true that the plane orthogonal to S_2 along the θ^* DOM line cuts the IRS in a 3PL IRC. The parameters of this curve are a^*, b^*, and c. In other words, the DOM of a single item characterized by an M3PL model always defines a unidimensional 3PL response curve when projected orthogonally onto the multidimensional IRS.

Returning to the fact that the b parameters of any item characterized by the MIRT model must have the same sign, consider Figures 2 and 4. Remembering that Figure 4 is the floor of Figure 2, in the two-dimensional case it can be seen, for example, that if b_1 is positive and b_2 is negative, the probability (height of the IRS) will decrease as θ_2 increases, a violation of the nondecreasing assumption. The same will be true if the signs of these two parameters are reversed. Although it cannot be displayed in a three-dimensional figure such as Figure 2, this same restriction applies to the M-dimensional case. From Expression 18, because the a parameters are always positive, it can be seen that the b parameters will always have the same sign, that of b^*.

Curvilinear Unidimensional Proficiency

As the beginning of my illustration of curvilinear unidimensionality, consider a curved line in two-dimensional Euclidean space, S_2. Points along such a curve are not free to vary in any direction; they are restricted to being on that curved line. Mathematically, such curves represent unidimensional variables in a plane. For the types of curves considered in this work, with axes θ_1 and θ_2, each unique point along θ_1 is associated with a unique point along θ_2 and vice versa.

In this way, a set of items may be such that their location parameters lie on a curve in S_2, and that curve defines a unidimensional (curvilinear) subspace of S_2. In general, the same principle can apply to a higher-dimensional space; item location parameters can lie along a unidimensional curve that runs through S_3 or a higher multidimensional space, S_M.

Demonstration Simulation Study

To illustrate this fact in the two-dimensional case, I have defined a hypothetical set of items with parameters displayed in the Appendix.

The first step in locating the items was to specify ω_1, b^*, and a^* for each item, such that the items were located on the curve in S_2 shown in Figure 5. Note that this specification involves using the polar coordinate representation of the exponent in the model; ω_1 represents the direction and b^* the distance of the polar coordinates. These parameters are displayed in the appendix as *Givens*. The a^* parameters were arbitrarily selected because the purpose of this simulation is only to demonstrate my main theme. The c parameters were all set to .2 for the same reason. To show all parameters discussed previously, b_1, b_2, d, a_1, and a_2 are also displayed in the Appendix and they were computed from the equations discussed previously, as follows: Expressions 11 and 15

$$d = -b^* a^*, \tag{23}$$

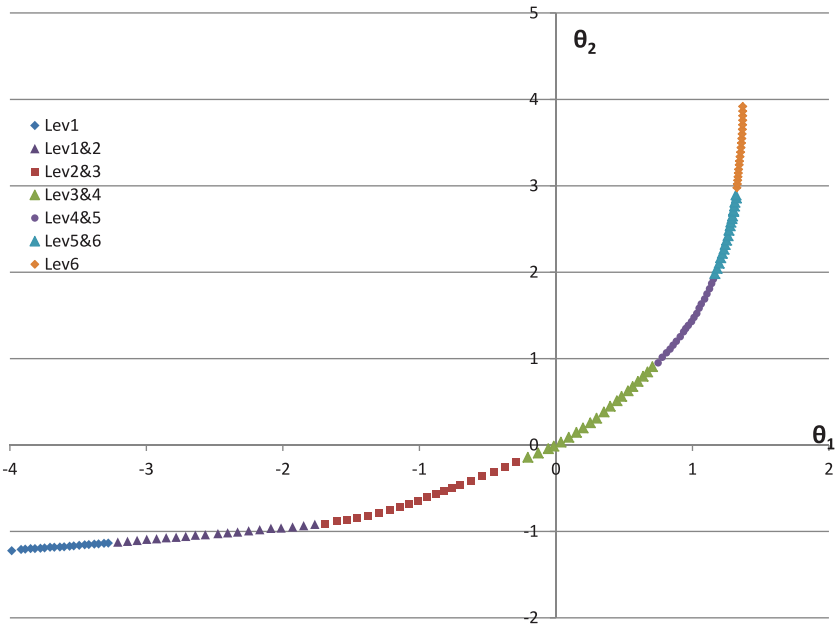

Figure 5 Locations of items in S_2 by levels administered.

Table 1 Design for Simulation: Items by Levels

			Linking items				
Test-taker level	Total items	On-level items	Levs 1 & 2	Levs 2 & 3	Levs 3 & 4	Levs 4 & 5	Levs 5 & 6
Lev1	40	30	10 Lev2				
Lev2	40	20	10 Lev1	10 Lev3			
Lev3	40	20		10 Lev2	10 Lev4		
Lev4	40	20			10 Lev3	10 Lev5	
Lev5	40	20				10 Lev4	10 Lev6
Lev6	40	30					10 Lev5

Expressions 12 and 15

$$a_1 = a^* \cos\left(\omega_1\right), \text{ and}$$
$$a_2 = a^* \cos\left(\omega_2\right) = a^* \sin\left(90^\circ - \omega_2\right) = a^* \sin\left(\omega_1\right), \tag{24}$$

Expression 22

$$b_1 = b^* \cos\left(\omega_1\right), \text{ and}$$
$$b_2 = b^* \sin\left(\omega_1\right). \tag{25}$$

Because the intention is to illustrate how a vertical scale can theoretically be developed across levels such as grades in school, the simulation design is that of a six-level assessment using a nonequivalent group common-item design such as that discussed in Carlson (2011, pp. 60–61) and Kolen & Brennan (2014, pp. 431–432). The item design is displayed in Table 1. For Levels 1 and 6, 30 dichotomously scored on-level items were specified, whereas for Levels 2 through 5, each level had 20 such items specified. Thus test takers were simulated to respond to 40 items each, and the total number of items was 140. In order to link test forms across levels, 20 linking items were shared by each pair of adjacent levels.

Two-dimensional proficiencies were simulated for 2,000 test takers at each level with simulated tests configured as shown in Table 1. All simulations were carried out in Microsoft Excel. The simulated data were scaled using BILOG-MG (du Toit, 2003) and linked using STUIRT (Kim & Kolen, 2004).

Table 2 Parameters of the Two-Dimensional Proficiencies

	Mean		*SD*		*r*
Level	θ_1	θ_2	θ_1	θ_2	
1	−3.228	−1.093	0.496	0.138	0.678
2	−1.528	−0.779	0.496	0.223	0.870
3	−0.265	−0.101	0.507	0.328	0.902
4	0.598	0.801	0.295	0.428	0.928
5	1.100	1.893	0.085	0.383	0.808
6	1.330	3.233	0.069	0.546	0.380

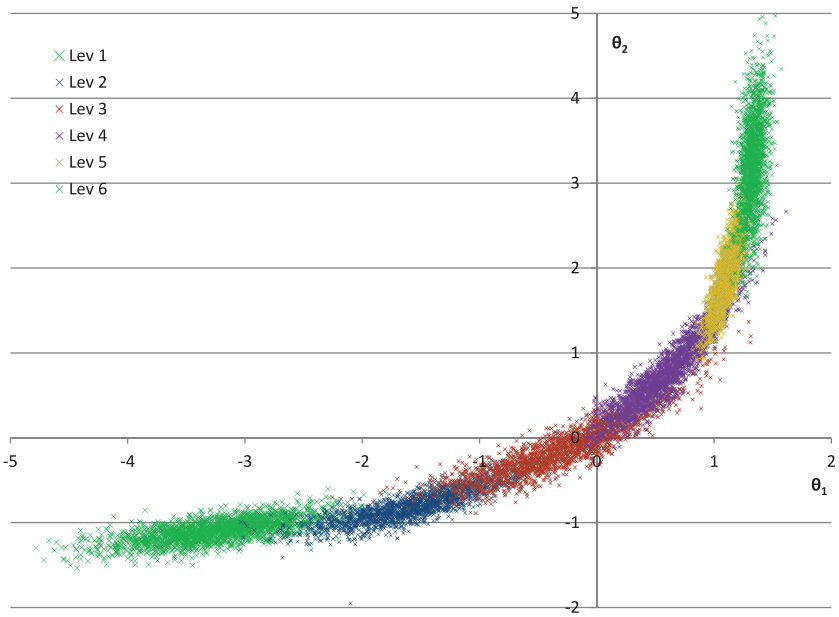

Figure 6 Scatterplot of population proficiencies for the simulation by level.

The thetas specified for the 2,000 simulated test takers at each of the six levels were bivariate normal with means, standard deviations, and correlations shown in Table 2. An Excel spreadsheet, using random generators provided in that software, was used to generate initial values on two bivariate normal uncorrelated variables. Two different linear combinations of these initial variables were then used to define θ_1 and θ_2 separately for each level, such that the test takers were also located along the same curvilinear unidimensional space as the item location parameters. In the case of these proficiencies, they vary somewhat about the line rather than lying exactly on a line as do the location parameters. These transformations were carried out by trial and error so as to achieve distributions appropriate for my demonstration. The final result was the 2,000 simulated test takers at each of the six levels, with the means increasing from Level 1 to Level 6. As may be seen from the standard deviations in Table 2, the lower levels varied more on θ_1 than on θ_2, whereas the higher levels varied more on θ_2. This variation in the standard deviations allows for representing what I intended, a proficiency dimension that changes with respect to two underlying dimensions as the level increases from Level 1 to Level 6, as might occur in a cross-grade educational assessment. Those underlying dimensions are designed to represent two different subject matter subareas within an overall subject area such as mathematics. A scatterplot of the resulting population proficiencies showing relationships to the two initial dimensions is shown in Figure 6.

As may be seen in Figure 6, the Level 1 simulated test takers vary primarily along the θ_1 dimension, whereas at Level 6 they vary primarily along the θ_2 dimension, and as the level increases there is a transition to varying less on θ_1 and more on θ_2. As mentioned above, these changes in variation are designed to simulate a gradual transition of a scale that changes from relying on one subarea of a school subject matter to relying on a different subarea of that same subject as grade level increases.

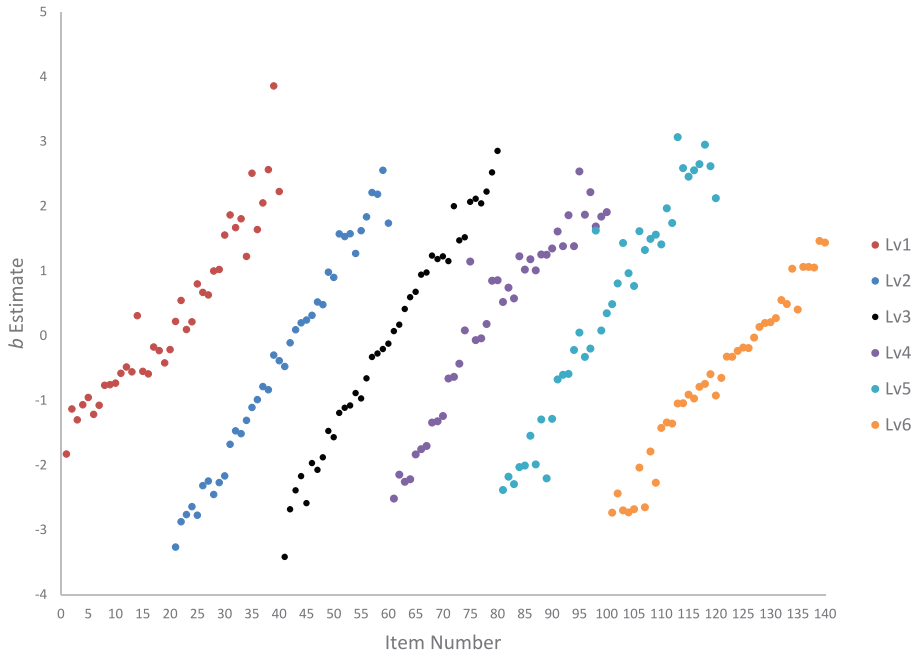

Figure 7 BILOG *b*-parameter estimates by level.

Demonstration Sample Design

To demonstrate that simulated sample data generated using the item parameters in the appendix and the population proficiencies displayed in Figure 6 can actually be scaled in a unidimensional vertical scale, Excel was used to generate the data. Using the item parameters, random dichotomous item response data were generated for the proficiency values displayed in Figure 6 and items by levels, as displayed in Table 1.

Independent Calibration Analyses

In the first analyses, the item response data for each of the six levels were independently calibrated with a 3PL model using the BILOG-MG software. Using the resulting item parameter estimates, the STUIRT software was used to link the resulting scales across the six levels.

Results

Calibration and Item Parameter Estimation Using BILOG-MG

As mentioned previously, the first step in the analyses was to calibrate the simulated item response data independently for each of the six levels. The BILOG estimates of the *b* parameters (locations or difficulties) for all 140 items are displayed by level in Figure 7.

Figure 8 similarly displays the BILOG *a*-parameter estimates (slopes or discriminations) by level. As usual when calibrating item response data, these parameter estimates tend to vary from the parameters more than do the *b* parameters.

Similarly, Figure 9 shows the BILOG *c*-parameter estimates (lower asymptotes) by level. Recall that the *c* parameters were all specified as .2 in the simulation. As shown in Figure 9, most of these estimates lie between .15 and .35, but there are a number of outliers. This is not unexpected when estimating *c* parameters.

Recall that the data were generated using a two-dimensional MIRT model, so the one-dimensional estimates shown in Figures 7, 8, and 9 are not directly estimating the parameters used to generate the data. However, Figure 10 displays plots to show the relationships between the unidimensional BILOG *b*-parameter estimates and the b^* location parameters, representing the location along the DOMs used in the simulation. Clearly the BILOG *b* estimates are highly correlated with the location parameters used in the simulation. All six of the correlations are between .97 and .99.

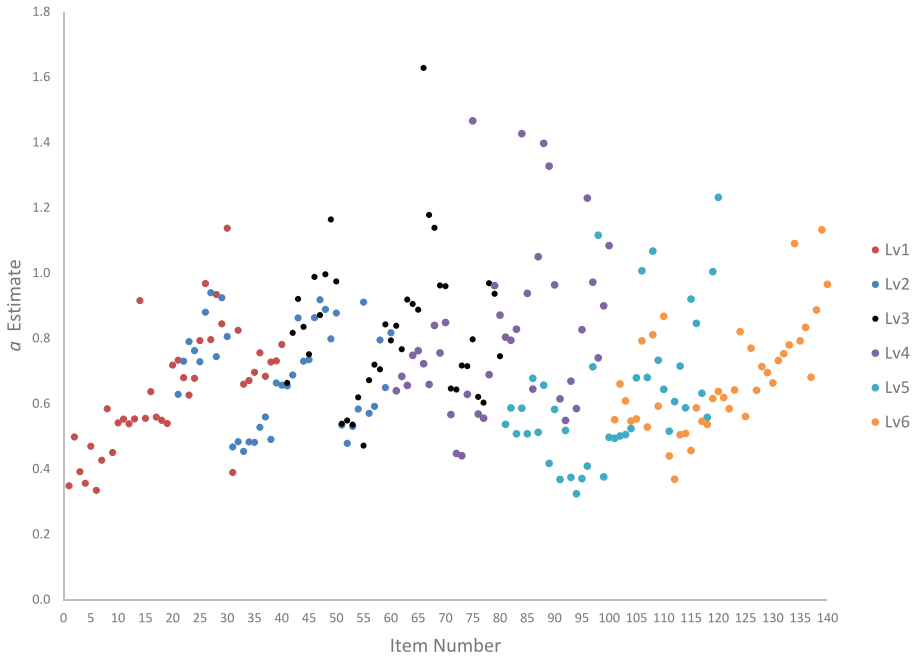

Figure 8 BILOG *a*-parameter estimates by level.

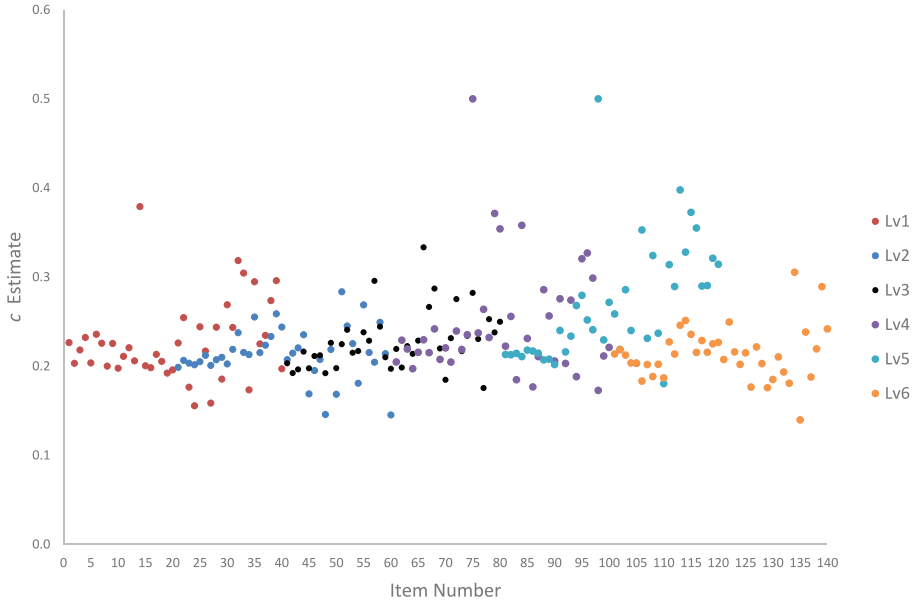

Figure 9 BILOG *c*-parameter estimates by level.

Although it is well known that the *a* parameters are not estimated as well as the *b* parameters, inexplicably, for two Levels, 4 and 5, the correlations of the BILOG *a* estimates with the a^* parameters are very low, .39 and .17, respectively. For the other four levels, those correlations range from .74 to .82, which are quite high for estimation of the slope parameters in IRT.

Proficiency Estimation Using BILOG-MG

The *expected a posteriori* (EAP) method was used to estimate proficiencies in BILOG-MG. Because the data were generated from an MIRT model with two proficiency dimensions, θ_1 and θ_2, whereas the unidimensional BILOG EAP estimation

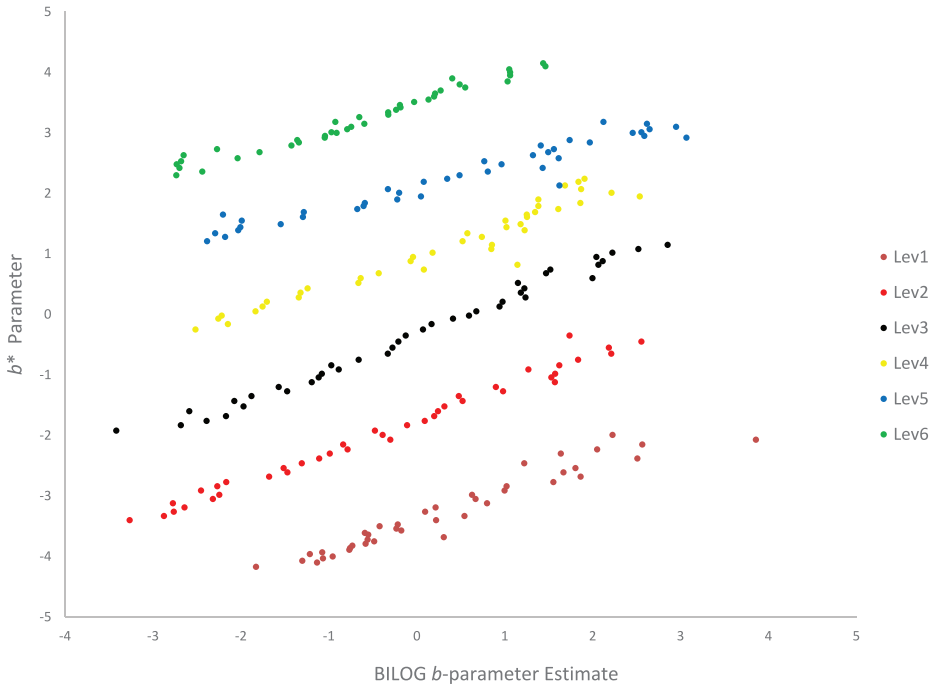

Figure 10 b^* parameters and BILOG b estimates by level.

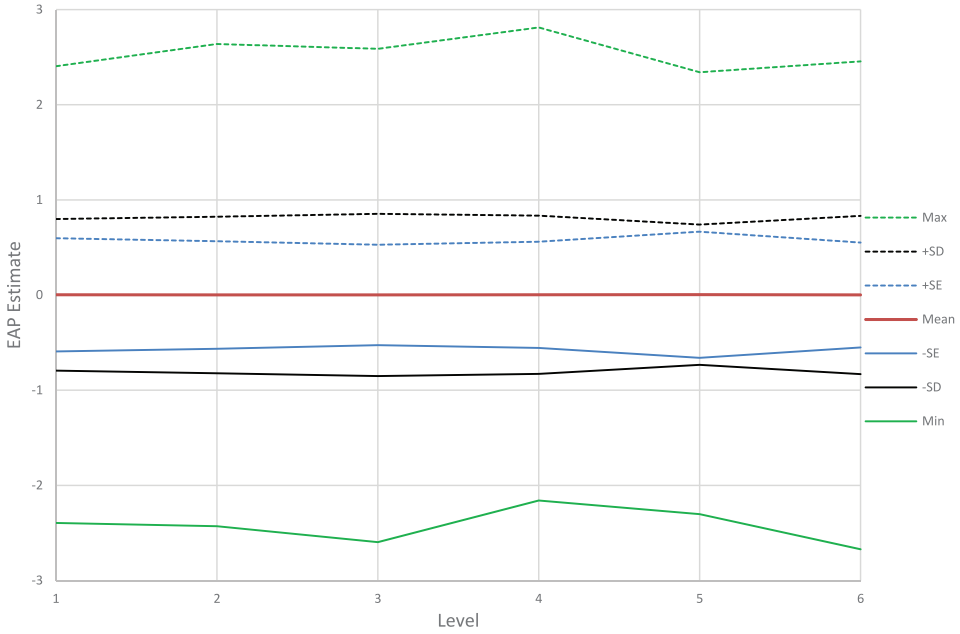

Figure 11 Means and ranges of expected a posteriori proficiency estimates by level.

only yields one proficiency variable, I did not necessarily expect to see high relationships between the generating and estimated proficiency variables.

To provide a first look at the results, I show in Figure 11 the mean unscaled EAP scores by level, the means plus and minus the standard deviations, and the means plus and minus the average estimated standard error of the proficiency estimates. Also displayed are the maxima and minima of the EAP estimates by level. The means, of course, are all near zero as a function of the usual metric arbitrarily selected during estimation. As seen in the figure, the Level 5 data are somewhat anomalous, having a smaller standard deviation and larger average standard error than the other levels.

Table 3 Summary Statistics of BILOG Expected a Posteriori Estimates by Level

Level	Mean	*SD*	Empirical reliability	*r* with NC[a]	Skew	Kurtosis
1	0.003	0.796	0.581	0.979	−0.039	−0.051
2	0.001	0.823	0.666	0.982	−0.024	0.039
3	0.001	0.852	0.729	0.983	−0.003	0.007
4	0.003	0.831	0.688	0.984	0.024	−0.007
5	0.003	0.737	0.583	0.978	0.034	0.030
6	0.001	0.832	0.606	0.986	−0.036	0.012

Note. NC = number-correct.
[a]Correlation with number-correct scores.

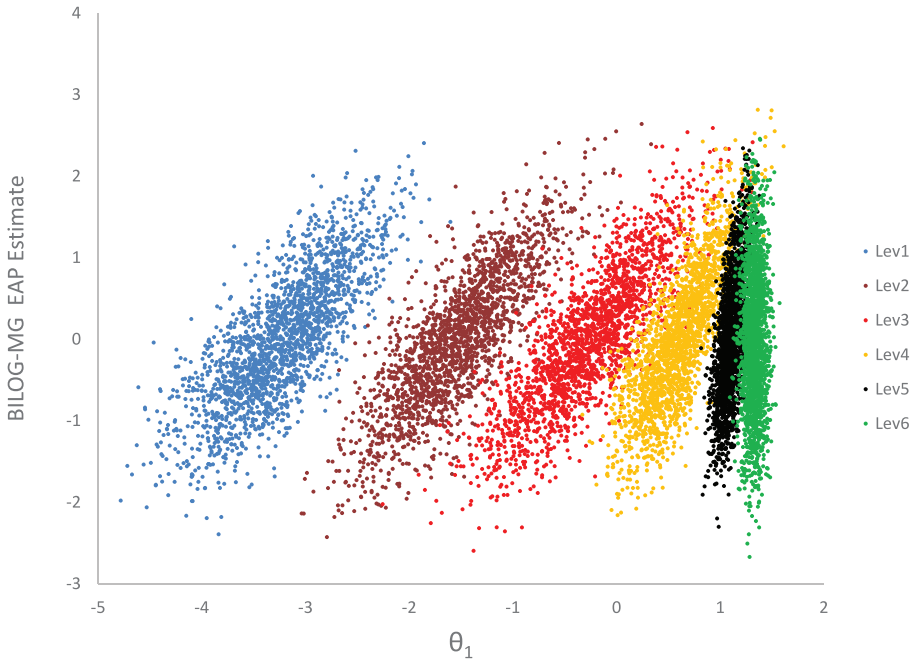

Figure 12 BILOG expected a posteriori estimates and generating values of θ_1 by level.

Interestingly, a phenomenon of decreasing variability, discussed in detail (with literature citations) by Briggs and Weeks (2009) and referred to as scale shrinkage, does not occur in my findings. One of the points made by Briggs and Weeks is that research by Yen in 1985 (cited by Briggs and Weeks, 2009, p. 6) suggested that the scale shrinkage could be related to violation of the IRT assumption of unidimensionality. So the lack of scale shrinkage in my demonstration may be related to my main thesis that one can develop a unidimensional scale in a multidimensional space.

Some descriptive statistics of the unscaled BILOG EAP proficiency estimates are displayed in Table 3. Although the distribution shapes were not constrained during estimation, the distributions are fairly symmetrical, with skewness and kurtosis indices near zero.

To show the relationships of the EAP estimates to the proficiency variables used in generating the data, Figures 12 and 13 display scatter plots of those relationships by level. As expected, given the configuration of generating proficiencies (Figure 6), the lower levels have larger variances on θ_1 than on θ_2, and the higher levels show the reverse of this trend.

Vertical Scaling Using STUIRT: Item Parameter Estimates

The objective of this study was to use multilevel simulated item response data having a configuration of a one-dimensional curvilinear proficiency variable in a two-dimensional space, which is demonstrably scalable unidimensionally. The nonequivalent group common-item design, shown in Table 1, was used to carry out the vertical scaling. The plan was to use the Stocking-Lord (SL; Stocking & Lord, 1983) test characteristic curve (TCC) method to link the parameter

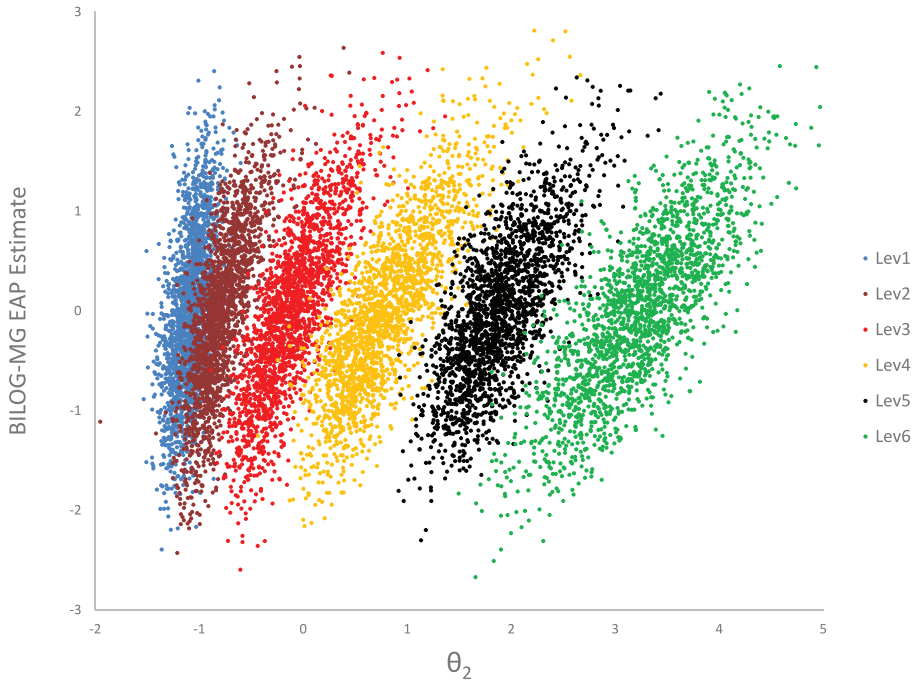

Figure 13 BILOG expected a posteriori estimates and generating values of θ_2 by level.

estimates through the common items simulated to be administered at adjacent levels, using the STUIRT software (Kim & Kolen, 2004). The order of the linking was to first link Level 4 to Level 3, then Level 5 to 4 and 6 to 5. Similarly, Level 2 was linked to Level 3 and then Level 1 to Level 2.

The STUIRT program yields transformation constants for four methods of linking: mean-mean, mean-sigma, Haebara TCC, and SL TCC (Kolen & Brennan, 2014). Because these methods are all linear transformations, there are slope (B) and intercept (A) constants used in the transformations:

$$b^{\mathrm{T}} = A + Bb, \tag{26}$$

$$a^{\mathrm{T}} = a/A, \tag{27}$$

and

$$\widehat{\theta}^{\mathrm{T}} = A + B\widehat{\theta}. \tag{28}$$

In these expressions, the T superscripts indicate the transformed parameter and proficiency estimates. Hence the $\widehat{\theta}^{\mathrm{T}}$ values are the estimates for each test taker on the vertically scaled proficiency.

The transformation constants are shown in Table 4.

For all but one of these linking steps, the resulting SL scaling constants were very reasonable and worked well, as were those of the other TCC method, the Haebara. However, in using the SL method to link Level 6 to Level 5, the results were larger than reasonable, as were the Haebara method constants (A and B values shaded in Table 4). Following a suggestion from the first author of the STUIRT program (personal communication from Seonghoon Kim), I tried rerunning the program using the mean-mean results as starting values for the TCC methods. This technique, however, did not yield better results (Haebara results unchanged; SL results $A = .034$, $B = 2.050$). I therefore decided to use the mean-sigma transformation constants for the link of Level 6 to Level 5 in the vertical scaling.

Because the objective of vertical scaling is to link the tests across levels in a unidimensional scale, the important results are how the TCCs, scaled item parameter estimates (primarily the locations, estimates of the *b*s), and scaled test-taker proficiency estimates line up. The scaled TCC plots are displayed in Figure 14, the scaled *b* estimates in Figure 15, and the scaled proficiency estimates in Figures 16 and 17.

As can be seen in Figure 14, the vertical scale for these data shows an excellent configuration of the unidimensional scale across the six levels. The TCCs of the levels are ordered as expected and separated as one would expect for a vertical

Table 4 STUIRT Estimates of Linking Constants

Method	Estimate	Link				
		L1 to L2	L2 to L3	L4 to L3	L5 to L4	L6 to L5
Mn-mn	A	1.130	0.938	0.826	0.417	0.358
	B	−6.014	−2.449	2.077	3.462	5.001
Mn-sigma	A	0.931	1.070	0.691	0.344	0.430
	B	−5.734	−2.589	1.962	3.389	5.120
Haebara	A	1.030	0.985	0.782	0.509	3.679
	B	−5.591	−2.513	2.105	3.786	23.067
Stocking-Lord	A	1.153	0.973	0.720	0.441	4.994
	B	−6.014	−2.519	2.027	3.600	28.175

Note: Shaded cells highlight unreasonably high transformation constants.

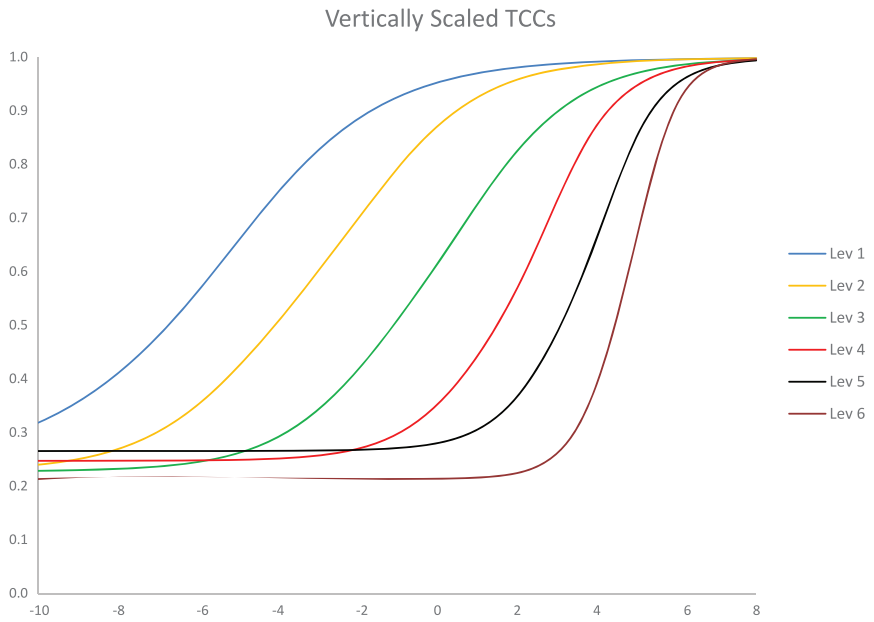

Figure 14 Vertically scaled test characteristic curves by level.

scale. The one TCC that is somewhat different is that for Level 6, having a steeper slope than the others; recall that the TCC linking methods did not work well for that level, so I decided to use the mean-sigma transformation for linking Level 6 to Level 5.

As shown in Figure 15, the vertical scaling lined up the location parameters of the items across levels in a very clearly unidimensional configuration.

In Figure 16, I show the differences between the scaled *b*-parameter estimates of the linking items on the adjacent levels. As may be seen, most of those differences are in the range of −0.5 to +0.5, with only nine points out of 100 outside of this range, including one in the link of Level 2 to Level 1 that is very much an outlier at −1.65.

The next three figures show how the simulated test-taker mean scores lined up in the unidimensional scale. Figure 17 shows how the means lined up across levels and also shows the variation (in terms of standard deviations and range).

Figure 17 shows a very regular pattern of how the vertically scaled mean scores lined up across levels. The previously mentioned smaller variation at Level 5 is also obvious in this figure; Level 6 also shows a similar smaller spread of scores, undoubtedly due to that level being linked to Level 5.

Similarly, Figure 18 shows how the quartile points on the scale line up in a configuration that supports the unidimensional scale across levels. Again, the smaller variation at Levels 5 and 6 is clear compared to the other levels.

An interesting comparison of the plots in Figures 18 and 19 is that with Figure 3 of Briggs and Weeks (2009). The shapes of the curves in my figures are very similar to those of Briggs and Weeks: decreasing slope with level (grade level

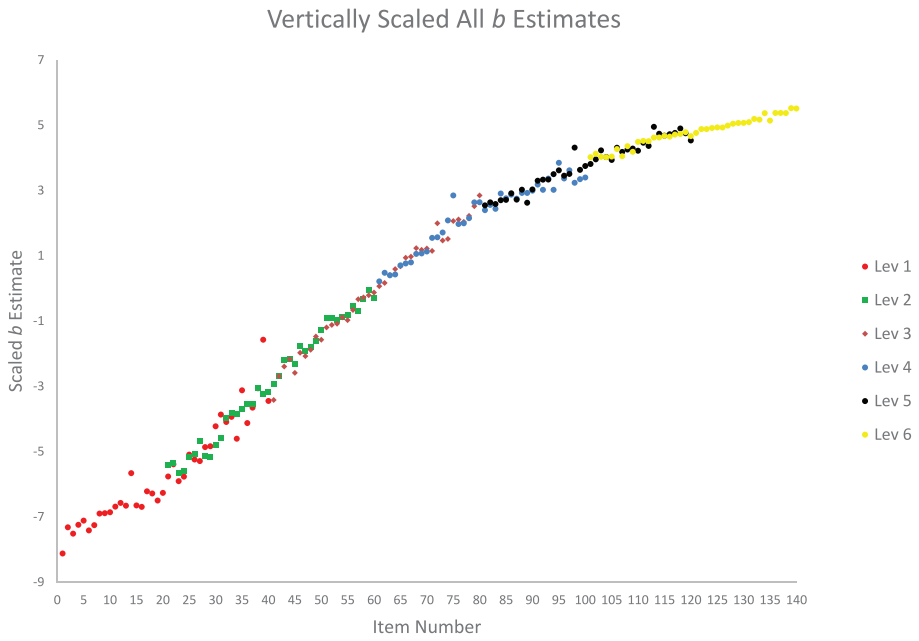

Figure 15 Vertically scaled *b*-parameter estimates by level.

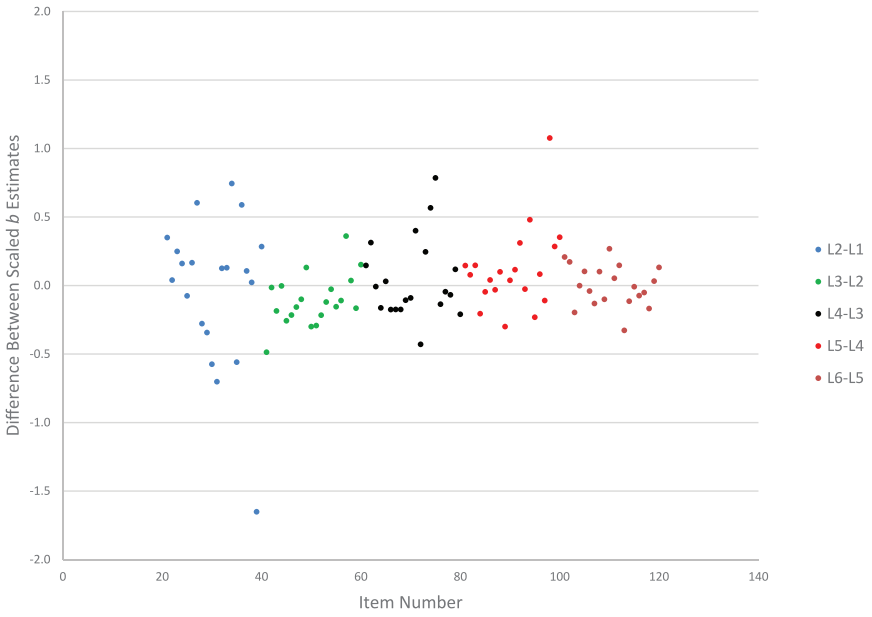

Figure 16 Differences between linking item *b*-parameter estimates for adjacent levels.

in the Briggs and Weeks case). These authors pointed out that "the growth in score means from grade to grade appears somewhat nonlinear, and decelerating over time" (p. 8) and also that "this is consistent with previous findings of Kolen (2006)" (p. 8).

Figures 19 and 20 show how the BILOG EAP estimates, after being vertically scaled, relate to the generating proficiency variables, θ_1 and θ_2. Again the lower levels have more spread in the direction of the first dimension and less spread in the direction of the second, and the reverse is true of the higher levels, as expected.

These two figures provide further evidence that, as conjectured, a vertical scale following a curvilinear unidimensional configuration in the context of a multidimensional space can be very accurately recovered with a unidimensional IRT model analysis of item response data.

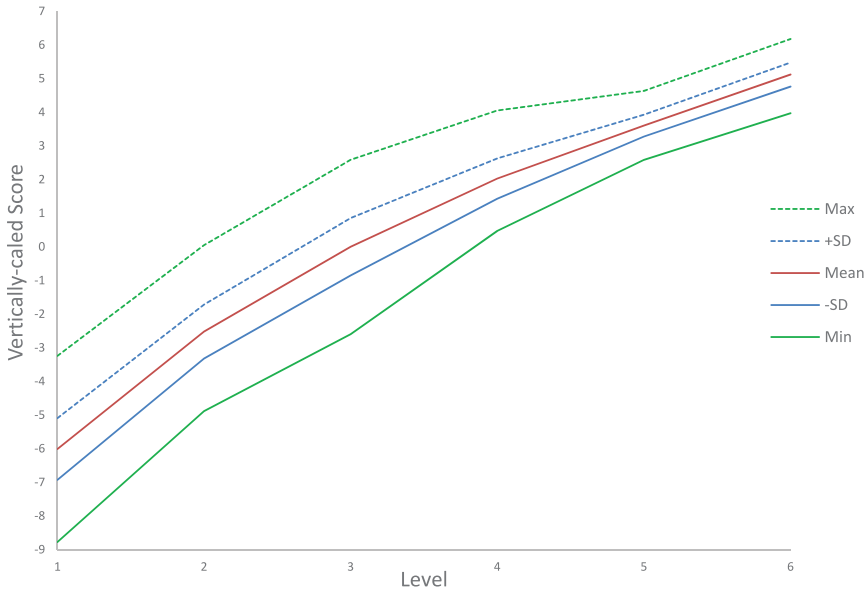

Figure 17 Means and variation of vertically scaled scores, across levels.

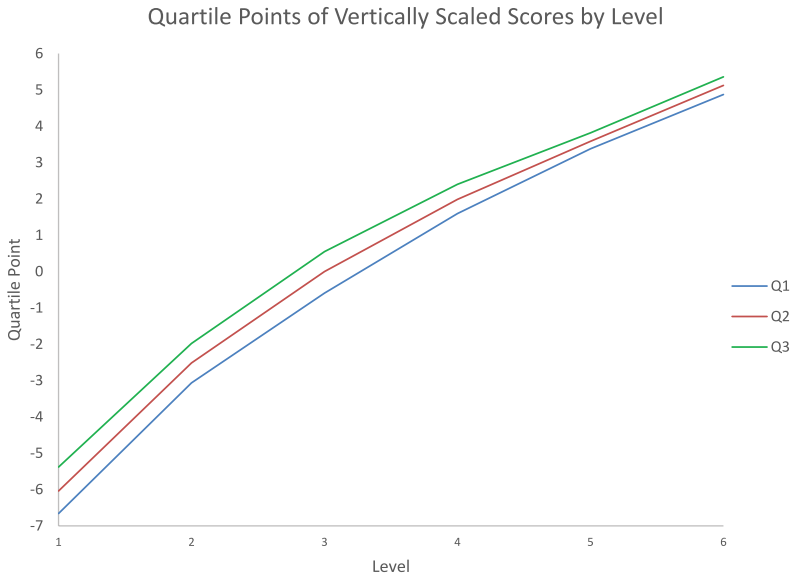

Figure 18 Quartile points of the vertically scaled scores, across levels.

Concurrent Calibration Procedures

Partly because of the issue of linking Level 6 to Level 5 by the TCC methods, I decided to try concurrently calibrating the same data to investigate how that type of procedure would work. I started by trying a concurrent calibration of the data from all six levels. As will be seen below, the results from using this procedure were very unsatisfactory, so I tried a different method: three concurrent calibrations of adjacent level data, followed by linking. This method provided more satisfactory results.

Concurrent Calibration of All Six Levels

First I tried concurrently calibrating all six levels with those items not administered to simulated test takers in each level specified as not administered. This procedure, surprisingly to me, did not result in reasonable estimates of the item

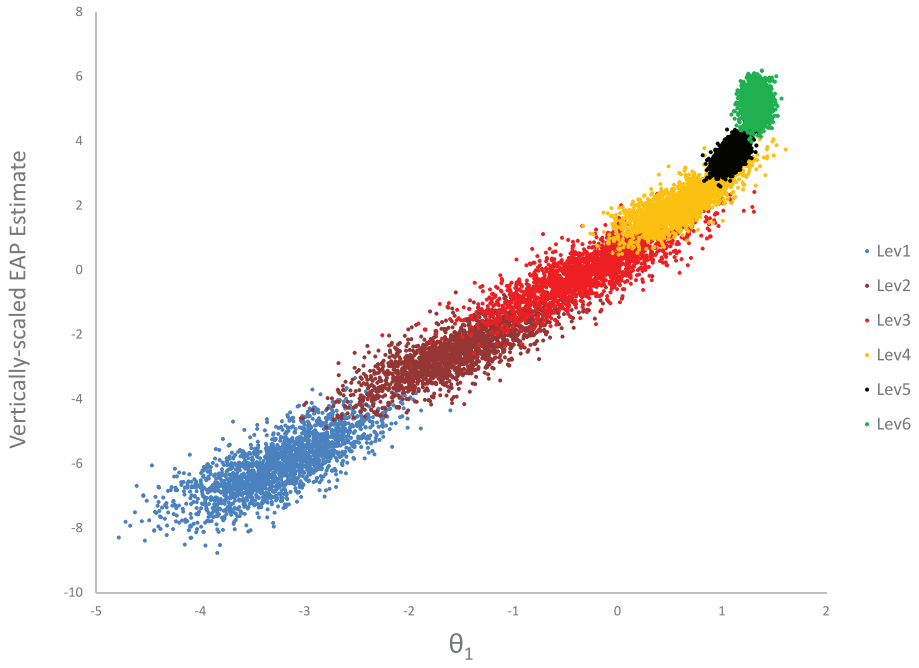

Figure 19 Relationships between vertically scaled expected a posteriori estimates and θ_1.

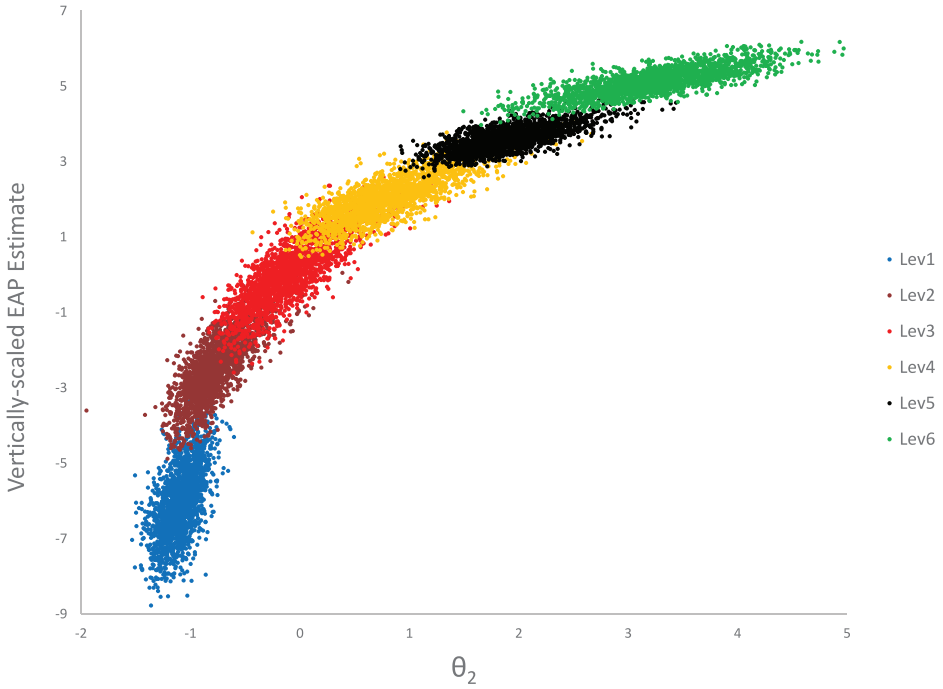

Figure 20 Relationships between vertically scaled expected a posteriori estimates and θ_2.

parameters. I tried using both BILOG-MG, as in the previously discussed study and results, and MULTILOG (du Toit, 2003) to do these calibrations; neither produced reasonable results. BILOG-MG yielded *a*-parameter estimates falling between .014 and .022 for the 20 anchor items administered at Levels 4 and 5. MULTILOG yielded 17 *a*-parameter estimates between 5.00 and 13.20, and all 20 *b*-parameter estimates fell between 16.41 and 32.93 for the 20 anchor items used to link Level 2 to Level 3. To verify these types of unreasonable estimates, I generated a new dataset with the same generating parameters, with similar poor results.

Table 5 Statistical Summary of Parameter Estimates From Concurrent Calibration of Adjacent Levels

Levels	Statistic	*a*	*b*	*c*
1 & 2	Mean	1.013	0.094	0.222
	SD	0.381	1.256	0.042
	Min	0.496	−1.704	0.123
	Max	2.006	2.699	0.340
3 & 4	Mean	0.979	0.003	0.223
	SD	0.244	1.640	0.041
	Min	0.562	−3.795	0.144
	Max	1.488	2.812	0.357
5 & 6	Mean	0.903	−0.189	0.225
	SD	0.325	1.216	0.045
	Min	0.338	−3.328	0.138
	Max	1.831	1.724	0.364

Concurrent Calibration of Adjacent Levels

Reasoning that the poor results from concurrent calibration of all six levels could be related to the sparsity of the overall item response data matrix, I decided to try a different concurrent calibration procedure using less sparse data matrices. This involved three concurrent calibrating runs involving Levels 1 and 2, Levels 3 and 4, and Levels 5 and 6, respectively. This procedure is somewhat similar to the "hybrid" method used by Briggs and Weeks (2009). However, their design involved linking across grades within years and also across test administration years within grades. They used linking with separate calibrations across grades within years of test administration and concurrent calibration within grades across administration years. Following these concurrent calibrations, I again used the STUIRT SL method to link the parameters and EAP estimates from the three calibrations. This exercise resulted in far more satisfactory results than the concurrent calibration of all six levels. It also seems somewhat better than the original analysis of six independent calibration followed by five vertical linking steps.

Table 5 displays the means, standard deviations, and minimum and maximum parameter estimates for the three calibration runs. All of these estimates seem perfectly reasonable.

After linking the three sets of calibration results using the SL TCC method in STUIRT, the *b*-parameter estimates lined up very well, as shown in Figure 21. Comparing Figures 15 and 21, for the individually calibrated and linked procedure of the six levels with these results from three concurrent calibrations and TCC linking, the results are fairly similar.

Figure 22 displays the relationships, by level, between the vertically scaled *b*-parameter estimates from the six separate calibrations and linking procedure and those from the three concurrent calibrations and linking procedure. As shown in the figure, these location estimates are very similar for the two procedures. The correlations between the two sets of estimates range from .93 to .98.

In addition, the differences between the *b*-parameter estimates for the anchor items all fell between -.61 and .36, as shown in Figure 23. Comparing Figures 16 and 23, these results appear to be better than those of the method of separate calibration of each of the six levels, followed by linking.

The results for the EAP estimates from these analyses were equally satisfying. Table 6 shows summary statistics for both the untransformed and the vertically scaled results. These data are also plotted in Figure 24, which shows a reasonable progression of the vertically scaled means and variations in the scaled scores. The quartile points are similarly displayed in Figure 25. As in the individually scaled and linked results in the previous section, these results show less variation in Levels 5 and 6 than in the lower levels.

An interesting feature of these results is that the curvilinearity displayed in both my analyses with separate calibration and those of Briggs and Weeks (2009) is not evident in these results.

Finally, Figure 26 is a scatterplot showing the relationships, by level, between the two vertical scaling methods. The correlations between these proficiency estimates of the simulated test takers across the six levels range from .57 to .71.

Discussion

In this paper, I set out to demonstrate the fact that a unidimensional proficiency scale can exist and item response data can be vertically scaled with a unidimensional IRT model and TCC linking under certain conditions: when the items and the

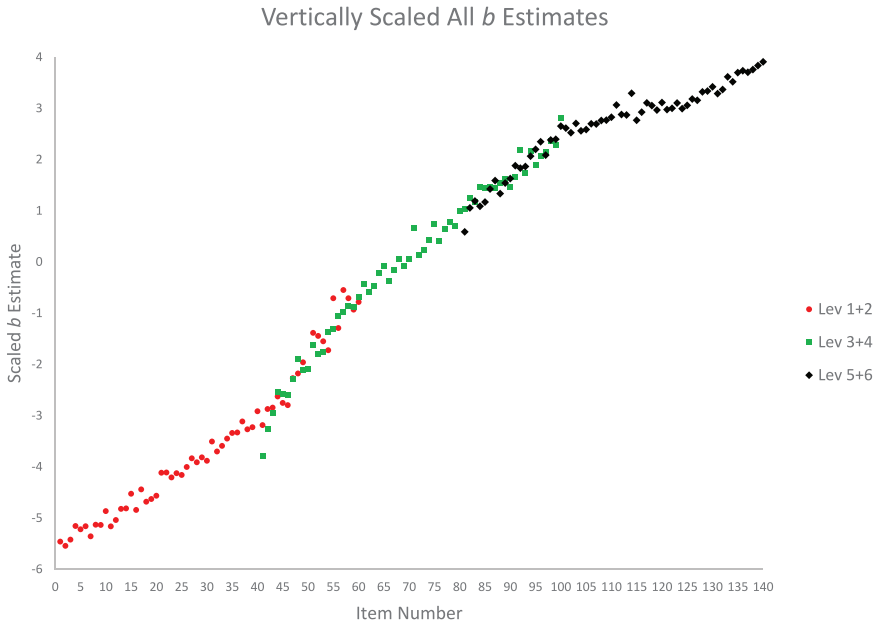

Figure 21 Scaled difficulty estimates based on three concurrent calibrations and Stocking-Lord linking.

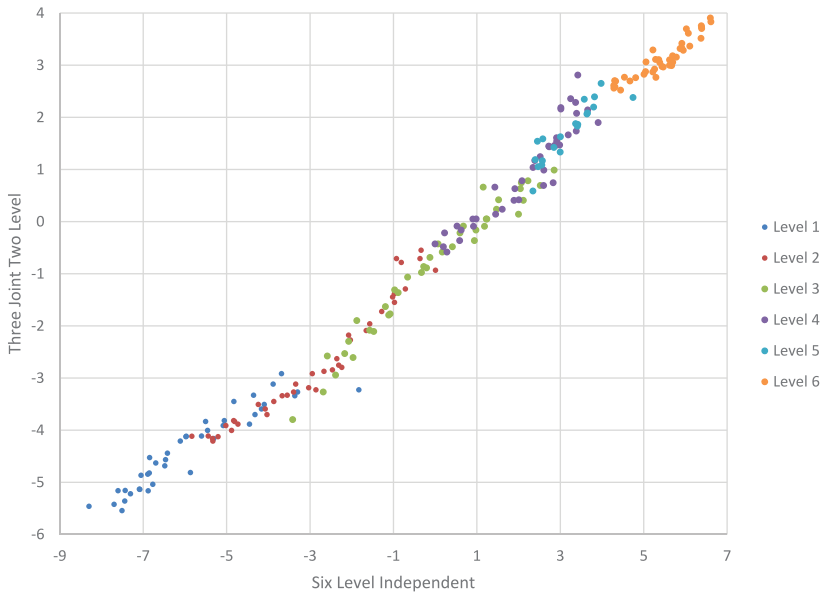

Figure 22 Vertically scaled *b*-parameter estimates for the two different analysis procedures.

test-taking population members, at different levels such as school grade levels, are administered assessment instruments that depend on multiple dimensions, as, for example, related to different subareas of a subject matter field. Although my demonstration was limited to a two-dimensional configuration of data and dichotomously scored items, these results can be inferred to also exist in a higher-dimensional situation and with polytomously scored items.

One very interesting finding is that independently calibrating each level and linking the results produced very similar scales to the alternative method of concurrently calibrating pairs of adjacent levels and then linking these concurrently calibrated results. The latter procedure, involving many fewer steps, is attractive. There are fewer steps at which mistakes could be made, and careful checking of results is necessary to ensure accurately analyzed, and hence interpreted, scales.

Kolen and Brennan (2014) presented some information about separate versus concurrent calibration. In a brief summary, they pointed out that separate calibration is more time consuming and stated that "concurrent estimation is expected

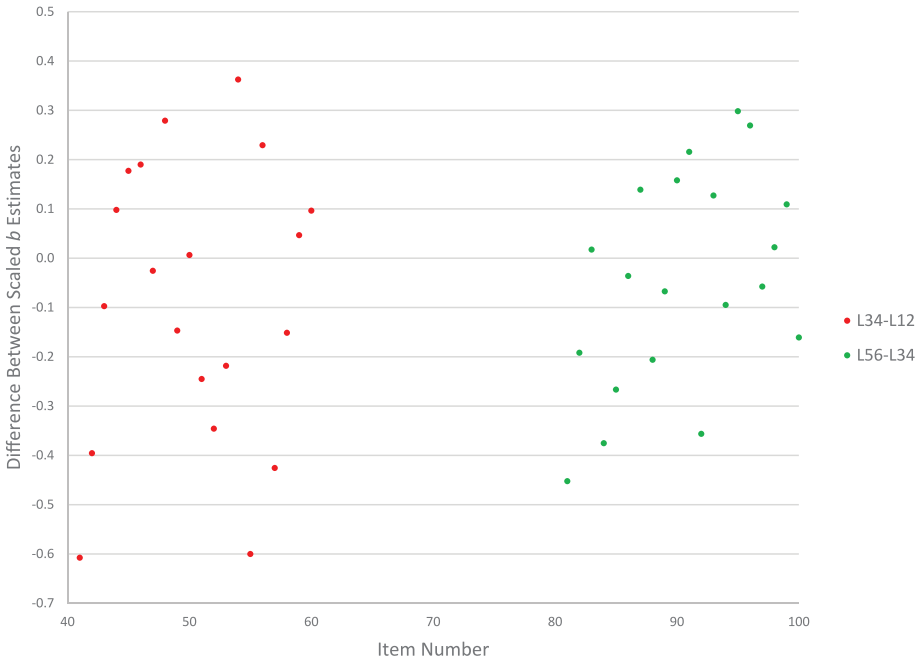

Figure 23 Differences between anchor item estimates of *b* parameters.

Table 6 Statistics of Expected A Posteriori Estimates for Concurrent Adjacent Level Calibration and Scaling

Untransformed expected a posteriori estimates				
Level	Mean	*SD*	Min	Max
1	−0.704	0.596	−2.672	1.238
2	0.705	0.611	−1.123	2.924
3	−0.577	0.714	−3.150	2.017
4	0.580	0.693	−1.419	3.088
5	−0.625	0.605	−2.641	1.358
6	0.628	0.698	−1.671	2.706
Vertically scaled expected a posteriori estimates				
Level	Mean	*SD*	Min	Max
1	−4.409	0.676	−6.641	−2.206
2	−2.811	0.693	−4.884	−0.293
3	−0.577	0.714	−3.150	2.017
4	0.580	0.693	−1.419	3.088
5	2.364	0.397	1.039	3.667
6	3.187	0.458	1.677	4.552

to produce more stable results because it makes use of all of the available information for parameter estimation" (p. 444). The results of the current study for the concurrent calibration of the six levels are not in accord with expectation. However, the results for the analyses in which I concurrently calibrated only the adjacent levels and then linked those results might be seen to be somewhat in agreement with that expectation. But I would caution that we need more research on the issue of separate versus concurrent calibration before we can make any definitive conclusions. Kolen and Brennan (2014) also stated that

> with concurrent estimation, violation of the unidimensionality assumption might be quite severe. This assumption requires that a single ability be measured across all grades, which seems unlikely with achievement tests. Violation of the unidimensionality assumption might cause problems with concurrent estimation. (p. 444)

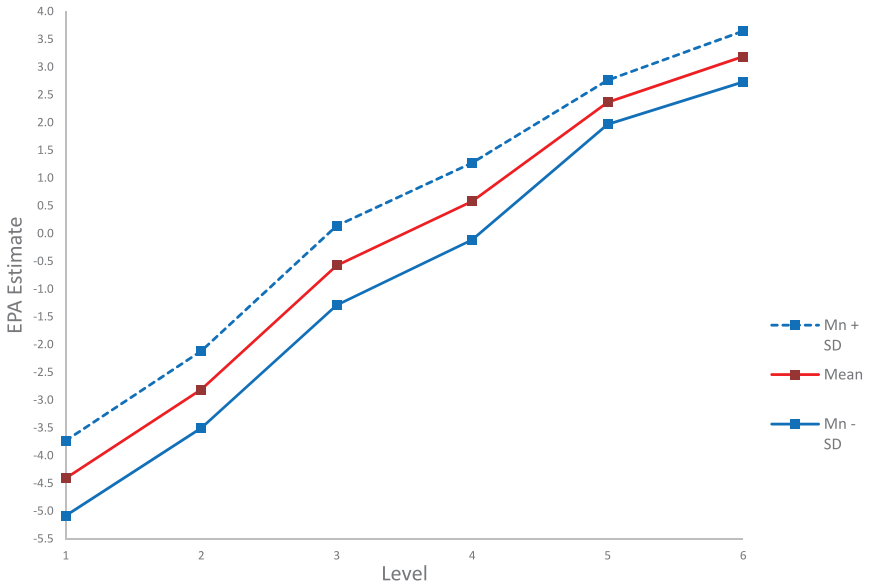

Figure 24 Means and variation for concurrent adjacent level calibration and scaling.

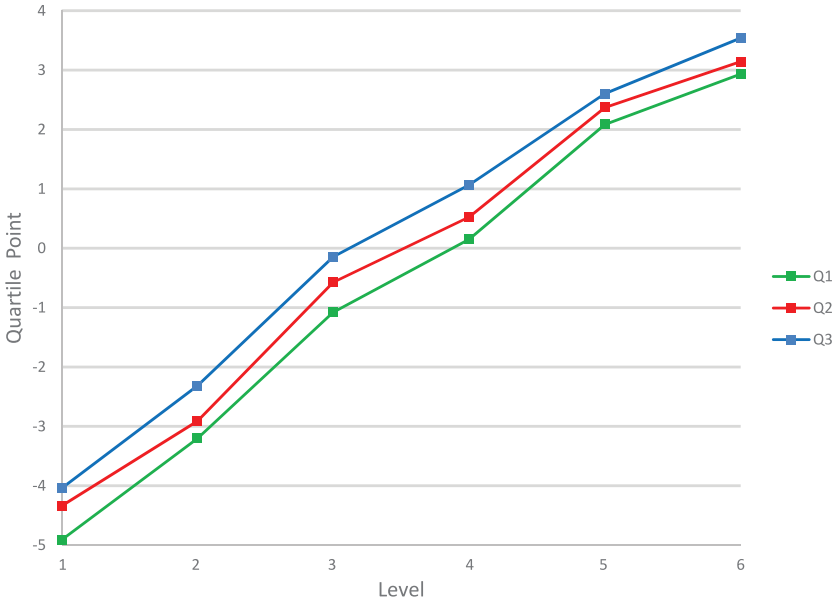

Figure 25 Quartile points for concurrent adjacent level calibration and scaling.

Considering this statement in light of comments cited earlier is interesting. As cited earlier, Patz and Yao (2007, p. 260) stated, "When calibrating items from multiple test forms for the purpose of measuring students across a range of grade levels, the IRT assumption of unidimensionality would appear implausible." The conclusion that I reached from my demonstration is that this assumption is not necessarily implausible. Also as cited earlier, Kolen and Brennan (2014, p. 469) stated, "One of the most challenging aspects of applying IRT to vertical scaling is the assumption that the same unidimensional ability is assessed across grade." I believe that my demonstration here shows that this assumption may not always be as challenging as they stated. Also, as I pointed out earlier, my results do not show the decreasing variability mentioned by Briggs and Weeks (2009), which they pointed out could be associated with violation of the IRT assumption of unidimensionality. In my opinion, the fact that my demonstration does not show this decreasing variability may well be related to the fact that, under condition like those of my demonstration, a curvilinear unidimensional scale *can* exist and be developed in a multidimensional space.

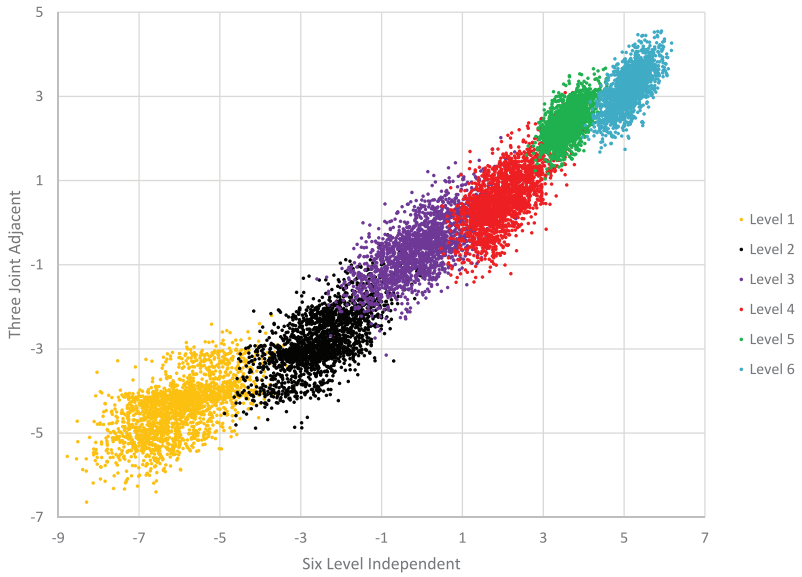

Figure 26 Relationships by level of the vertically scaled expected a posteriori estimates by level.

As I have demonstrated in this and my previous research (Carlson, 2001), a unidimensional scale can exist, and be derived, under the condition that the items on the scale are located on a curved line (or perhaps very close to it; this has not been investigated in this study) in the multidimensional space and, of course, that the populations of test takers' proficiencies on the underlying dimensions are closely aligned with that curve. Furthermore, unidimensional scaling and linking of such data can yield a very reasonable scale that should be interpretable as, for example, growth in an academic assessment subject matter area in which the focus on instruction in various subareas varies across grade level.

In my view, the results speak for themselves without further discussion. There is one caveat that must, however, be presented. When measurement professionals are dealing with real educational assessment data across, for example, grade levels, the item response data cannot be analyzed in a way that will reveal the existence of the type of curvilinear multidimensionality discussed in this report. We can, of course, use MIRT models and factor analysis to study the dimensionality of most datasets. But, as I have shown previously (Carlson, 2011), relying on the eigenvalue–eigenvector structure of the data is an unreliable method of inferring dimensionality; that methodology overestimates the number of dimensions. The other side of this dimensionality coin is, however, that in this study I have demonstrated that it is possible to scale multilevel data in the presence of multidimensionality of a certain type. Hence, when data from an assessment program are vertically scaled across grades, if the item response data fit a unidimensional IRT model, we can interpret the data as having an underlying unidimensional scale. We just do not truly know strictly from IRT analysis results whether or how the derived scale relates to underlying multiple dimensions. To interpret that scale, psychometricians will have to work closely with assessment development professionals to come up with reasonable ways in which to interpret the resulting scales. Such cooperative efforts can bring into play the cognitive theoretic aspects of the subject matter and how students learn subject matter in a discipline (e.g., school subjects) to explain what differences in performance at different grade levels on a vertical scale mean. Explaining the meaning of scores and score differences (between test takers with different scores, across grades within years, or across years within test takers) requires such cooperative efforts of a variety of testing professionals. One purpose of the demonstration in this report is to encourage such cooperation.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*(4), 3–14.

Carlson, J. E. (2001, April). *Curvilinear dimensions of tests and items*. Paper presented at the meeting of the American Educational Research Association, Seattle, WA.

Carlson, J. E. (2011). Statistical models for vertical linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 59–70). New York, NY: Springer.

du Toit, M. (Ed.) (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models (Version 1.0)*. Iowa City: Iowa Testing Programs, The University of Iowa.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: American Council on Education and Praeger.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253–272). New York, NY: Springer.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D. (1986, April). *The discriminating power of items that measure more than one dimension*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D. (1989). *The interpretation and application of multidimensional item response theory models; and computerized testing in the instructional environment* (Office of Naval Research Report ONR 89-2). Iowa City, IA: ACT.

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models*. Paper presented at the meeting of the American Educational Research Association, Montreal, Canada.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361–373.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Appendix

		Givens: polar coordinates & a^*			Computed				
Item	Level admin.	ω_1	b^*	a^*	b_1	b_2	d	a_1	a_2
1	One only	17.06	−4.17	0.80	−3.987	−1.223	3.336	0.765	0.235
2		17.14	−4.10	0.80	−3.918	−1.208	3.280	0.764	0.236
3		17.22	−4.07	0.83	−3.888	−1.205	3.378	0.793	0.246
4		17.30	−4.03	0.85	−3.848	−1.198	3.426	0.812	0.253
5		17.43	−4.00	0.87	−3.816	−1.198	3.480	0.830	0.261
6		17.55	−3.96	0.90	−3.776	−1.194	3.564	0.858	0.271
7		17.63	−3.93	0.93	−3.745	−1.190	3.655	0.886	0.282
8		17.70	−3.89	0.95	−3.706	−1.183	3.696	0.905	0.289
9		17.85	−3.86	0.97	−3.674	−1.183	3.744	0.923	0.297
10		18.00	−3.82	1.00	−3.633	−1.180	3.820	0.951	0.309
11		18.10	−3.79	1.03	−3.602	−1.177	3.904	0.979	0.320
12		18.20	−3.75	1.05	−3.562	−1.171	3.938	0.997	0.328
13		18.30	−3.72	1.07	−3.532	−1.168	3.980	1.016	0.336
14		18.40	−3.68	1.10	−3.492	−1.162	4.048	1.044	0.347
15		18.50	−3.64	1.12	−3.452	−1.155	4.077	1.062	0.355
16		18.60	−3.61	1.15	−3.421	−1.151	4.152	1.090	0.367
17		18.73	−3.57	1.17	−3.381	−1.146	4.177	1.108	0.376
18		18.85	−3.54	1.20	−3.350	−1.144	4.248	1.136	0.388
19		18.98	−3.50	1.22	−3.310	−1.138	4.270	1.154	0.397
20		19.10	−3.47	1.25	−3.279	−1.135	4.338	1.181	0.409
21	One & two	19.30	−3.40	1.30	−3.209	−1.124	4.420	1.227	0.430
22		19.55	−3.33	1.35	−3.138	−1.114	4.496	1.272	0.452
23		19.80	−3.26	1.40	−3.067	−1.104	4.564	1.317	0.474
24		20.00	−3.19	1.45	−2.998	−1.091	4.626	1.363	0.496
25		20.35	−3.12	1.50	−2.925	−1.085	4.680	1.406	0.522
26		20.60	−3.05	1.55	−2.855	−1.073	4.728	1.451	0.545
27		21.00	−2.98	1.60	−2.782	−1.068	4.768	1.494	0.573
28		21.30	−2.91	1.65	−2.711	−1.057	4.802	1.537	0.599

Continued

Item	Level admin.	Givens: polar coordinates & a^*			Computed				
		ω_1	b^*	a^*	b_1	b_2	d	a_1	a_2
29		21.55	−2.84	1.70	−2.641	−1.043	4.828	1.581	0.624
30		22.00	−2.77	1.75	−2.568	−1.038	4.848	1.623	0.656
31		22.50	−2.68	0.80	−2.476	−1.026	2.144	0.739	0.306
32		22.90	−2.61	0.85	−2.404	−1.016	2.219	0.783	0.331
33		23.35	−2.54	0.90	−2.332	−1.007	2.286	0.826	0.357
34		23.80	−2.46	0.95	−2.251	−0.993	2.337	0.869	0.383
35		24.30	−2.38	1.00	−2.169	−0.979	2.380	0.911	0.412
36		24.80	−2.30	1.05	−2.088	−0.965	2.415	0.953	0.440
37		25.50	−2.23	1.10	−2.013	−0.960	2.453	0.993	0.474
38		26.20	−2.15	1.15	−1.929	−0.949	2.473	1.032	0.508
39		26.80	−2.07	1.20	−1.848	−0.933	2.484	1.071	0.541
40		27.45	−1.99	1.25	−1.766	−0.917	2.488	1.109	0.576
41	Two & three	28.35	−1.92	1.30	−1.690	−0.912	2.496	1.144	0.617
42		28.65	−1.83	1.35	−1.606	−0.877	2.471	1.185	0.647
43		29.40	−1.76	1.40	−1.533	−0.864	2.464	1.220	0.687
44		30.00	−1.68	1.45	−1.455	−0.840	2.436	1.256	0.725
45		30.80	−1.60	1.50	−1.374	−0.819	2.400	1.288	0.768
46		31.40	−1.52	1.55	−1.297	−0.792	2.356	1.323	0.808
47		31.80	−1.43	1.60	−1.215	−0.754	2.288	1.360	0.843
48		32.00	−1.35	1.65	−1.145	−0.715	2.228	1.399	0.874
49		32.20	−1.27	1.70	−1.075	−0.677	2.159	1.439	0.906
50		32.40	−1.20	1.75	−1.013	−0.643	2.100	1.478	0.938
51		32.50	−1.12	0.80	−0.945	−0.602	0.896	0.675	0.430
52		32.70	−1.04	0.85	−0.875	−0.562	0.884	0.715	0.459
53		32.90	−0.98	0.90	−0.823	−0.532	0.882	0.756	0.489
54		33.10	−0.91	0.95	−0.762	−0.497	0.865	0.796	0.519
55		33.30	−0.84	1.00	−0.702	−0.461	0.840	0.836	0.549
56		33.50	−0.75	1.05	−0.625	−0.414	0.788	0.876	0.580
57		33.70	−0.65	1.10	−0.541	−0.361	0.715	0.915	0.610
58		33.90	−0.55	1.15	−0.457	−0.307	0.633	0.955	0.641
59		33.91	−0.45	1.20	−0.373	−0.251	0.540	0.996	0.669
60		34.00	−0.35	1.25	−0.290	−0.196	0.438	1.036	0.699
61	Three & four	34.50	−0.25	1.30	−0.206	−0.142	0.325	1.071	0.736
62		35.50	−0.16	1.35	−0.130	−0.093	0.216	1.099	0.784
63		36.50	−0.07	1.40	−0.056	−0.042	0.098	1.125	0.833
64		37.50	−0.02	1.45	−0.016	−0.012	0.029	1.150	0.883
65		43.50	0.05	1.50	0.036	0.034	−0.075	1.088	1.033
66		44.00	0.13	1.55	0.094	0.090	−0.202	1.115	1.077
67		44.50	0.21	1.60	0.150	0.147	−0.336	1.141	1.121
68		45.00	0.28	1.65	0.198	0.198	−0.462	1.167	1.167
69		45.70	0.36	1.70	0.251	0.258	−0.612	1.187	1.217
70		46.30	0.43	1.75	0.297	0.311	−0.753	1.209	1.265
71		47.50	0.52	0.80	0.351	0.383	−0.416	0.540	0.590
72		48.50	0.60	0.85	0.398	0.449	−0.510	0.563	0.637
73		49.00	0.68	0.90	0.446	0.513	−0.612	0.590	0.679
74		49.50	0.74	0.95	0.481	0.563	−0.703	0.617	0.722
75		50.00	0.82	1.00	0.527	0.628	−0.820	0.643	0.766
76		50.30	0.88	1.05	0.562	0.677	−0.924	0.671	0.808
77		50.80	0.95	1.10	0.600	0.736	−1.045	0.695	0.852
78		51.30	1.02	1.15	0.638	0.796	−1.173	0.719	0.897
79		51.60	1.08	1.20	0.671	0.846	−1.296	0.745	0.940
80		52.10	1.15	1.25	0.706	0.907	−1.438	0.768	0.986
81	Four & five	51.80	1.21	1.30	0.748	0.951	−1.573	0.804	1.022
82		52.50	1.28	1.35	0.779	1.015	−1.728	0.822	1.071
83		52.80	1.34	1.40	0.810	1.067	−1.876	0.846	1.115
84		53.00	1.39	1.45	0.837	1.110	−2.016	0.873	1.158
85		53.40	1.44	1.50	0.859	1.156	−2.160	0.894	1.204
86		53.70	1.49	1.55	0.882	1.201	−2.310	0.918	1.249

Continued

Item	Level admin.	Givens: polar coordinates & a^*			Computed				
		ω_1	b^*	a^*	b_1	b_2	d	a_1	a_2
87		54.00	1.55	1.60	0.911	1.254	−2.480	0.940	1.294
88		54.50	1.61	1.65	0.935	1.311	−2.657	0.958	1.343
89		54.80	1.65	1.70	0.951	1.348	−2.805	0.980	1.389
90		55.00	1.69	1.75	0.969	1.384	−2.958	1.004	1.434
91		55.20	1.74	0.80	0.993	1.429	−1.392	0.457	0.657
92		55.60	1.79	0.85	1.011	1.477	−1.522	0.480	0.701
93		55.90	1.84	0.90	1.032	1.524	−1.656	0.505	0.745
94		56.50	1.90	0.95	1.049	1.584	−1.805	0.524	0.792
95		56.90	1.95	1.00	1.065	1.634	−1.950	0.546	0.838
96		57.70	2.07	1.05	1.106	1.750	−2.174	0.561	0.888
97		57.20	2.01	1.10	1.089	1.690	−2.211	0.596	0.925
98		58.10	2.13	1.15	1.126	1.808	−2.450	0.608	0.976
99		58.60	2.19	1.20	1.141	1.869	−2.628	0.625	1.024
100		59.00	2.24	1.25	1.154	1.920	−2.800	0.644	1.071
101	Five & six	59.60	2.30	1.30	1.164	1.984	−2.990	0.658	1.121
102		60.00	2.36	1.35	1.180	2.044	−3.186	0.675	1.169
103		60.40	2.42	1.40	1.195	2.104	−3.388	0.692	1.217
104		60.90	2.48	1.45	1.206	2.167	−3.596	0.705	1.267
105		61.20	2.53	1.50	1.219	2.217	−3.795	0.723	1.314
106		61.50	2.58	1.55	1.231	2.267	−3.999	0.740	1.362
107		61.90	2.63	1.60	1.239	2.320	−4.208	0.754	1.411
108		62.20	2.68	1.65	1.250	2.371	−4.422	0.770	1.460
109		62.50	2.73	1.70	1.261	2.422	−4.641	0.785	1.508
110		63.00	2.79	1.75	1.267	2.486	−4.883	0.794	1.559
111		63.30	2.84	0.80	1.276	2.537	−2.272	0.359	0.715
112		63.60	2.88	0.85	1.281	2.580	−2.448	0.378	0.761
113		63.80	2.92	0.90	1.289	2.620	−2.628	0.397	0.808
114		64.00	2.95	0.95	1.293	2.651	−2.803	0.416	0.854
115		64.30	3.00	1.00	1.301	2.703	−3.000	0.434	0.901
116		64.40	3.01	1.05	1.301	2.715	−3.161	0.454	0.947
117		64.70	3.06	1.10	1.308	2.766	−3.366	0.470	0.994
118		65.00	3.10	1.15	1.310	2.810	−3.565	0.486	1.042
119		65.20	3.15	1.20	1.321	2.859	−3.780	0.503	1.089
120		65.50	3.18	1.25	1.319	2.894	−3.975	0.518	1.137
121	Six only	66.00	3.26	1.30	1.326	2.978	−4.238	0.529	1.188
122		66.25	3.30	1.32	1.329	3.021	−4.356	0.532	1.208
123		66.50	3.34	1.35	1.332	3.063	−4.509	0.538	1.238
124		66.75	3.38	1.37	1.334	3.106	−4.631	0.541	1.259
125		67.00	3.42	1.40	1.336	3.148	−4.788	0.547	1.289
126		67.25	3.46	1.43	1.338	3.191	−4.948	0.553	1.319
127		67.50	3.51	1.45	1.343	3.243	−5.090	0.555	1.340
128		67.75	3.55	1.48	1.344	3.286	−5.254	0.560	1.370
129		68.00	3.60	1.50	1.349	3.338	−5.400	0.562	1.391
130		68.25	3.65	1.52	1.353	3.390	−5.548	0.563	1.412
131		68.50	3.70	1.55	1.356	3.443	−5.735	0.568	1.442
132		68.75	3.75	1.57	1.359	3.495	−5.888	0.569	1.463
133		69.00	3.80	1.60	1.362	3.548	−6.080	0.573	1.494
134		69.25	3.85	1.62	1.364	3.600	−6.237	0.574	1.515
135		69.50	3.90	1.65	1.366	3.653	−6.435	0.578	1.546
136		69.75	3.95	1.67	1.367	3.706	−6.597	0.578	1.567
137		70.00	4.00	1.70	1.368	3.759	−6.800	0.581	1.597
138		70.25	4.05	1.72	1.369	3.812	−6.966	0.581	1.619
139		70.50	4.10	1.75	1.369	3.865	−7.175	0.584	1.650
140		70.75	4.15	1.77	1.368	3.918	−7.346	0.584	1.671
Mean		43.561	0.140	1.275	−0.462	0.698	−0.430	0.856	0.854
SD		18.521	2.630	0.291	1.845	1.683	3.419	0.298	0.398
Min		17.060	−4.170	0.800	−3.987	−1.223	−7.346	0.359	0.235
Max		70.750	4.150	1.770	1.369	3.918	4.848	1.623	1.671

Suggested citation:

Carlson, J. C. (2017). *Unidimensional vertical scaling in multidimensional space* (Research Report No. RR-17-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12157

Action Editor: Matthias von Davier

Reviewers: Lixiong Gu and Rebecca Zwick