Title
Can Interval-level Scores be Obtained from Binary Responses?

Author
Bentler, Peter M.

Publication Date
2011-03-22

Peer reviewed

CAN INTERVAL-LEVEL SCORES BE OBTAINED FROM BINARY RESPONSES?

Peter M. Bentler

University of California, Los Angeles

Invited Presentation, Western Psychological Association annual convention, April 2011, Los Angeles CA.

In short, YES

(sometimes)

Outline

- Total Scores as Quantifications
- Normalizing Transformations
- Guttman Scales
- Absolute Simplex Theory
- Item Response Theory
- Rasch Models
- Conclusions

Total Scores as Quantifications

Let X_i be a variable, often an *item* on a scale. We concentrate on binary items where X_i takes on only 2 values, such as “correct” vs “incorrect” or “endorsed” vs “not endorsed”. The values 1 and 0 will be used to denote these values

Let X_T be the sum or total score across a set of p items

$$X_T = X_1 + X_2 + ... + X_p$$

Does X_T give the best possible quantification of the item responses? My goal is to describe methods that improve on X_T via $Y=f(X_1,X_2,\ldots,X_p)$

What is a “best possible” quantification?

1. One where differences between two Y values at different magnitudes of Y have the same meaning. This is an “interval” –level scale
2. One that permits intended statistical operations. This may involve linear transformations as well as mean and variance comparisons and correlational analysis

Requiring, achieving, and evaluating #1 is contro-versial (see e.g., Velleman & Wilkinson, 1993)

We will require #2, but aim to achieve #1

A Total Score X_T Can be Acceptable

when X_T has a distribution that is consistent with the theory of the attribute being measured

Usually, this is when X_T is normally distributed. Then relations with other normal variables will all be linear (assuming mv normality) and standard statistical analyses are meaningful

Even if X_T is really ordinal, "if it walks like a duck, swims like a duck, and quacks like a duck..," i.e., acts interval-like, meaningful conclusions are possible

If X_T Is Not Normal: Normalizing Transformations

Two main kinds:

1. Apply an explicit nonlinear function $Y = f(X_T)$ so that the new score Y is normal
2. Refer X_T to a table of the normal (0,1) distribution to get a normal z score

Normalizing Transformations in SAS

Table 32.2: Description of Normalizing Transformations

Transformation	Default Parameter	Name of New Variable	Equation
log(Y+a)	$a = 0$	Log_***Y***	$\log(Y + a), \quad Y + a > 0$
log10(Y+a)	$a = 0$	Log10_***Y***	$\log_{10}(Y + a), \quad Y + a > 0$
sqrt(Y+a)	$a = 0$	Sqrt_***Y***	$\sqrt{Y + a}, \quad Y + a > 0$
exp(Y)		Exp_***Y***	$\exp(Y)$
power(Y;a)	$a = 1$	Pow_***Y***	$Y^a, \quad Y > 0$ if a is not integral
arcsinh(Y)		Arcsinh_***Y***	$\log(Y + \sqrt{Y^2 + 1})$
Box-Cox(Y;a)	MLE	BC_***Y***	See text.

The Box-Cox transformation (Box and Cox 1964) is a one-parameter family of power tr
transformation as a limiting case. For $Y > 0$,

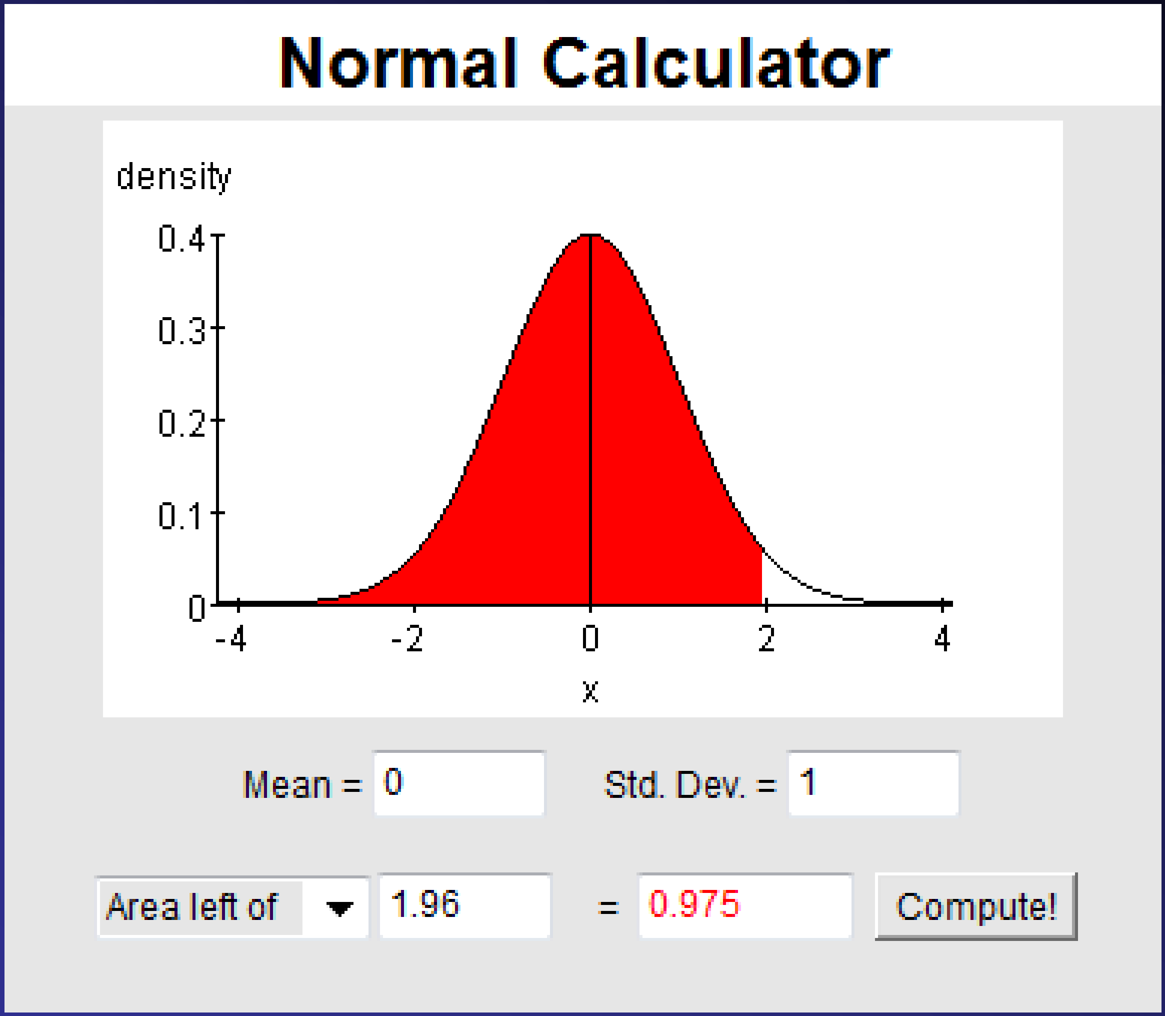
$$\mathrm{BC}(y;\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}$$

Normalizing Transformation via the Normal Curve

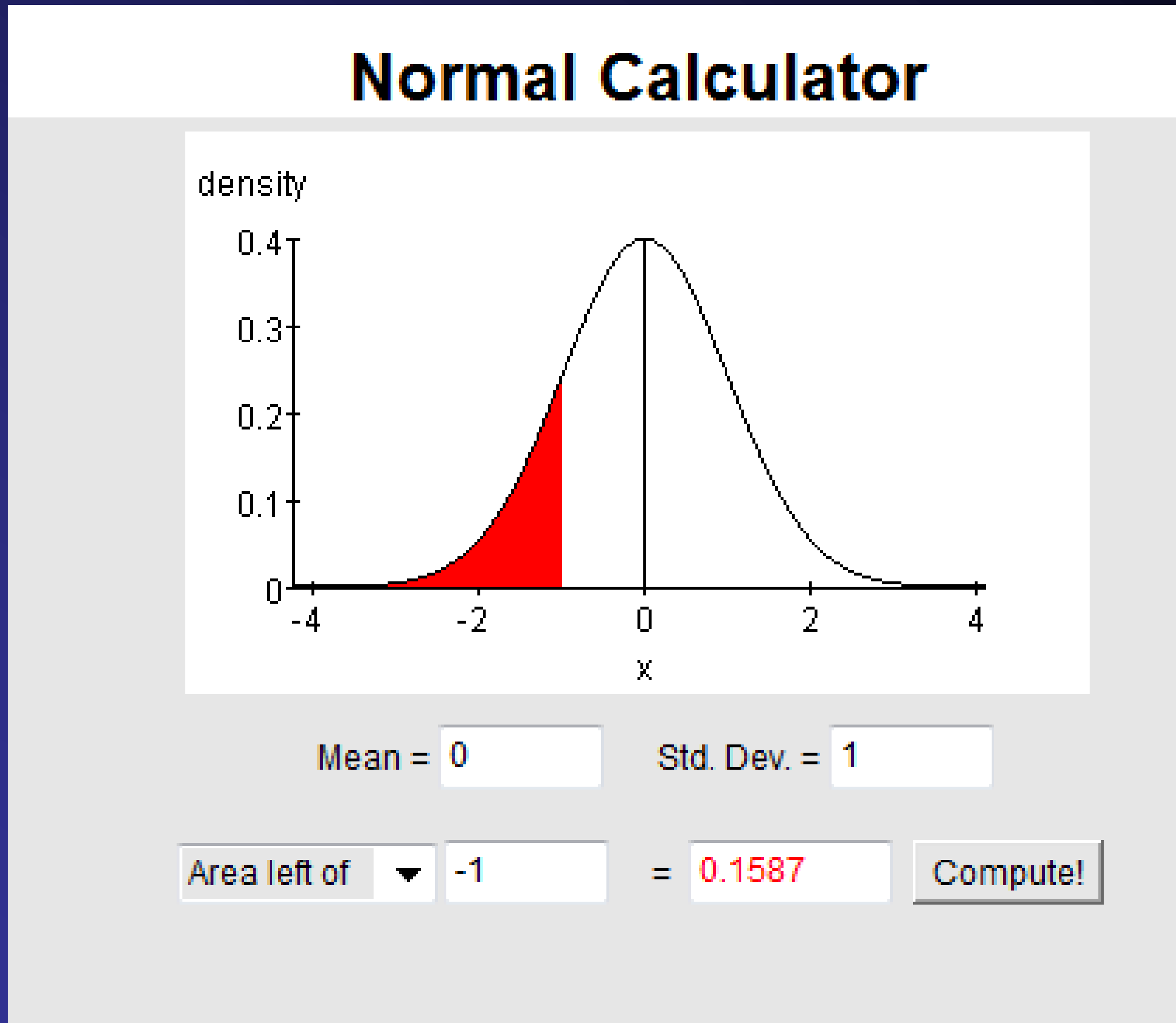
1. Get the frequency distribution of X_T , say f(x)
2. Smooth f(x) if desired
3. Get the cumulative frequency distribution F(x)
4. Find the percentile ranks P(x) of the F(x)
5. Using a table/calculator for N(0,1), do an inverse normal transformation of P(x) to get
   z-scores
6. Do a linear transform of the z-scores to get another mean and SD if desired

Red area is cumulative P(x)=.975

z-value is 1.96



http://www.stat.tamu.edu/~west/applets/normaldemo.html


Normal Calculator
density
0.4
0.3
0.2
0.1
0
-4
-2
0
2
4
x
Mean = 0
Std. Dev. = 1
Area left of
-1
= 0.1587
Compute!

If we accept that probabilities can be transformed to z-scores

Then we obtain scores that we can treat as interval:

- Linear transformations are allowed – they just change the mean and SD
- A fixed difference between 2 z-scores has the same meaning everywhere along the scale*

*Some theorists may also require empirical verification of equal meaning of differences

How can such a transformation mislead?

If the “true” underlying distribution is not the one we use, e.g., if it is not continuous and/or not normal

- It might be an ordered categorical distribution
  - Piaget’s concept of conservation (e.g. quantity) in children
  - “Stages” of Alzheimer’s progression
- It might be a skewed distribution, e.g. depression in the U.S. population

Application to Binary Data

The data must be unidimensional in some well-defined manner, and the number of items must be large enough

- If data are Bentler-Guttman scalable (defined below), the previous theory can be applied
- If the data fits a unidimensional Item Response Theory (IRT) model, the previous theory can be applied
- If the data fits a Rasch IRT model, the previous theory is not necessary but an interval scale is obtained

Guttman (1944) Scale

00000
00001
00011
00011
00011
00111
00111
00111
01111
11111

- Example of 10 subjects, 5 items
- “1” means correct (keyed) response
- Items ordered from hard to easy

Key point: If person gets a “hard” item right, he/she gets all easier items right

Largely abandoned – no clear statistical estimation and testing machinery

Absolute Simplex Theory (AST)
(Bentler, 1971)

- An *absolute simplex* is an n by p data matrix (n>p) that can be generated completely from one parameter per item
- It is a parameterization and estimation machinery for Guttman and near-Guttman data
- Approach used today is based on recent developments, including structural equation modeling (Bentler, 2009, 2011)

Moment Matrix: Average Sums of Squares and Cross-Products (SSCP Matrix)

Based on means and covariances, in the population this is equivalent to:

$$\Sigma_m = \Sigma + \mu\mu'$$

In a sample it is:

$$S_m = S + \bar{X}\bar{X}'$$

Under AST, Σ_m is a patterned matrix. With $\mu_1 \le \mu_2 \le \mu_3 \ldots$

$$\Sigma_m = \begin{bmatrix} \mu_1 & \mu_1 & \mu_1 & \mu_1 & \mu_1 \\ \mu_1 & \mu_2 & \mu_2 & \mu_2 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 & \mu_3 & \mu_3 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_4 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 \end{bmatrix}$$

- The entire matrix is a function of one parameter per item
- Items can be ordered by this matrix
- Structural equation modeling fits Σ_m via S_m

For the 10x5 binary data given earlier, S_m is

$$S_m = \begin{bmatrix} .1 & .1 & .1 & .1 & .1 \\ .1 & .2 & .2 & .2 & .2 \\ .1 & .2 & .5 & .5 & .5 \\ .1 & .2 & .5 & .8 & .8 \\ .1 & .2 & .5 & .8 & .9 \end{bmatrix}$$

$$\bar{X} = [.1 \quad .2 \quad .5 \quad .8 \quad .9]'$$

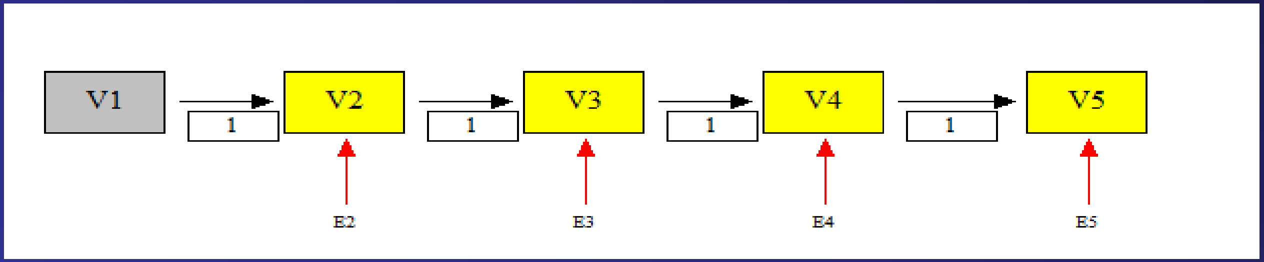
The means and SSCP can be fit by 1 par./item

Example Moment Matrix S_m :
Male Sexual Behavior (N=175)

	V1	V2	V3	V4	V5	V6	V7	V8
V1	.891	.771	.697	.583	.566	.497	.394	.377
V2	.771	.789	.686	.577	.566	.491	.377	.377
V3	.697	.686	.709	.571	.531	.497	.383	.377
V4	.583	.577	.571	.594	.526	.463	.371	.383
V5	.566	.566	.531	.526	.577	.429	.360	.366
V6	.497	.491	.497	.463	.429	.509	.366	.366
V7	.394	.377	.383	.371	.360	.366	.411	.337
V8	.377	.377	.377	.383	.366	.366	.337	.389

There are many ways to fit the absolute simplex model to data. Here are four:

1. A symmetric matrix with equalities
2. $\Sigma_m = TD_{\mu diff}T'$ where T is lower triangular with 1's and $D_{\mu diff} = diag\{\mu_1, \ldots (\mu_i - \mu_{i-1}), \ldots\}$
3. A simplex model with variances $D_{\mu diff}$



4. A regression model (explained later)

Model extensions allow errors, longer lags, etc.

Estimation Requires Items to Be Ordered

$$\Sigma_m = \begin{bmatrix} \mu_1 & \mu_1 & \mu_1 & \mu_1 & \mu_1 \\ \mu_1 & \mu_2 & \mu_2 & \mu_2 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 & \mu_3 & \mu_3 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_4 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 \end{bmatrix}$$

- Column sums →order
- Column SDs → order

- In practice, use column sums and SDs of the sample SSCP (moment) matrix
- Rank these separately, and average
- Use average ranks; break ties using means
- Diag (Σ_m) is inflated in quasi-simplex, so order items by off-diagonal entries only, e.g., ignore D in $\Sigma_m = TD_{\mu diff}T' + D$

Ordered Total Scores Generate the CDF

00000
00001
00011
00011
00011
00111
00111
00111
01111
11111

Motivation: Note that
% of subjects below a pattern
= % of subjects below total score

Pattern	Score X_T	% Below	CDF
11111	5	90	1.00
01111	4	80	.90
00111	3	50	.80
00011	2	20	.50
00001	1	10	.20
00000	0	0	.10

Person Ordering by X_T is Free of Item Weights – Any Weighted Sum is OK

Let $w_i > 0$ in the sum $X_T = w_1 X_1 + w_2 X_2 + \ldots + w_p X_p$

11111	X_T	$= w_1+w_2+w_3+w_4+w_5$
01111		$= w_2+w_3+w_4+w_5$
00111		$= w_3+w_4+w_5$
00011		$= w_4+w_5$
00001		$= w_5$
00000		$= 0$

The % of subjects below a pattern = % below a total score, no matter what the item weighting

Data-based Interval Scale Scores

- Since X_T completely orders the distribution of a unidimensional absolute simplex, it can be used to get the empirical cumulative distribution function (CDF) of the trait
- Given the CDF, we can use the inverse normal distribution function to compute z-scores
- This produces an interval scale if we are correct that the trait is normally distributed
- In real data, this is an approximation. But the empirical CDF → true CDF as n gets large

Regression Estimation of Absolute Simplex

Let π = proportion below (=%below/100)

$x' = (x_1, x_2, \ldots, x_p)$

be a person's item responses to p items

ordered 1,…,p from easy to hard

Then under the model

$$\pi = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

where $\beta_1 = 1 - \mu_1$ and $\beta_i = \mu_{i-1} - \mu_i$ predicts π_k exactly with $R^2 = 1.0$. Adding items with $\beta_i > 0$, π becomes continuous as $p \to \infty$. If also $n \to \infty$ then π approaches the population trait CDF.

The probabilities π are then transformed to a normal z-statistic, the interval score of interest

From this viewpoint*, π and z are *formative* measures -- they arise from the item responses.

In contrast, most extant measures are best con-sidered *reflective* measures, generated by a latent trait or factor.

See Treiblmaier, Bentler, & Mair (2011)

*From an IRT viewpoint, though, Guttman scales can also be considered as reflective.

Absolute Simplex Interval Scores in Practice

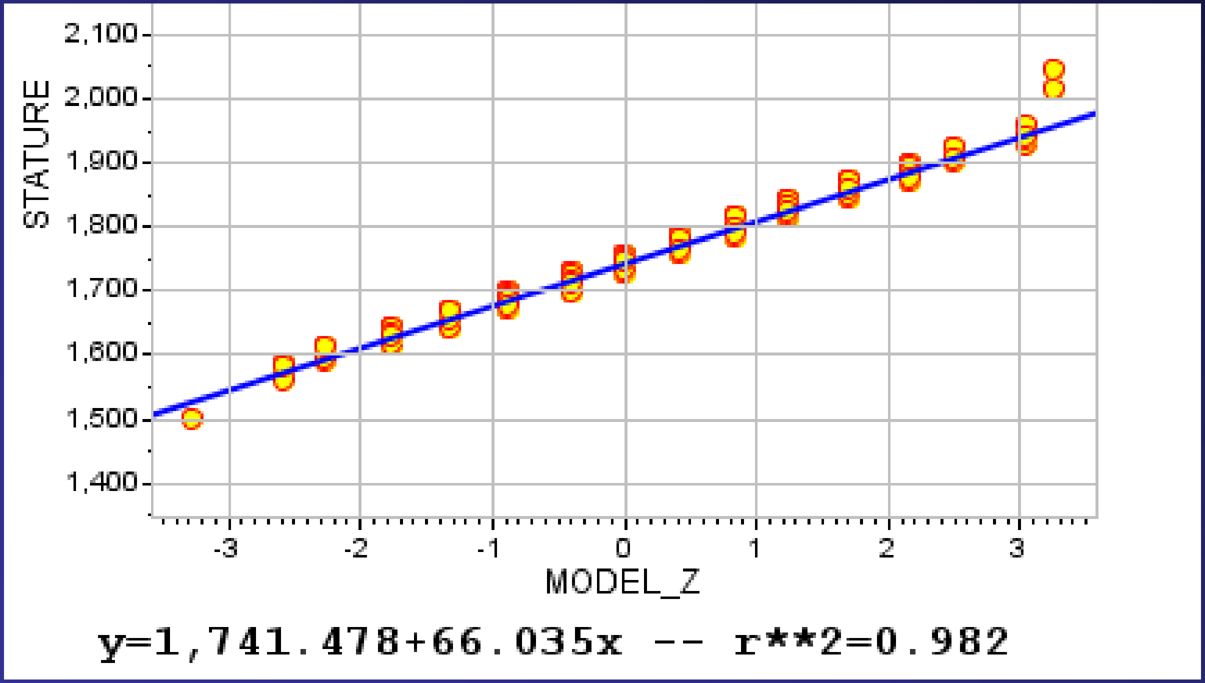
- Order X_T . Compute p_k, the prop. below X_{Tk} , for each person k=1,…,n. Run the regression

$$p_k = \sum_{i=1}^{p} \beta_i x_{ki} + \varepsilon_k$$

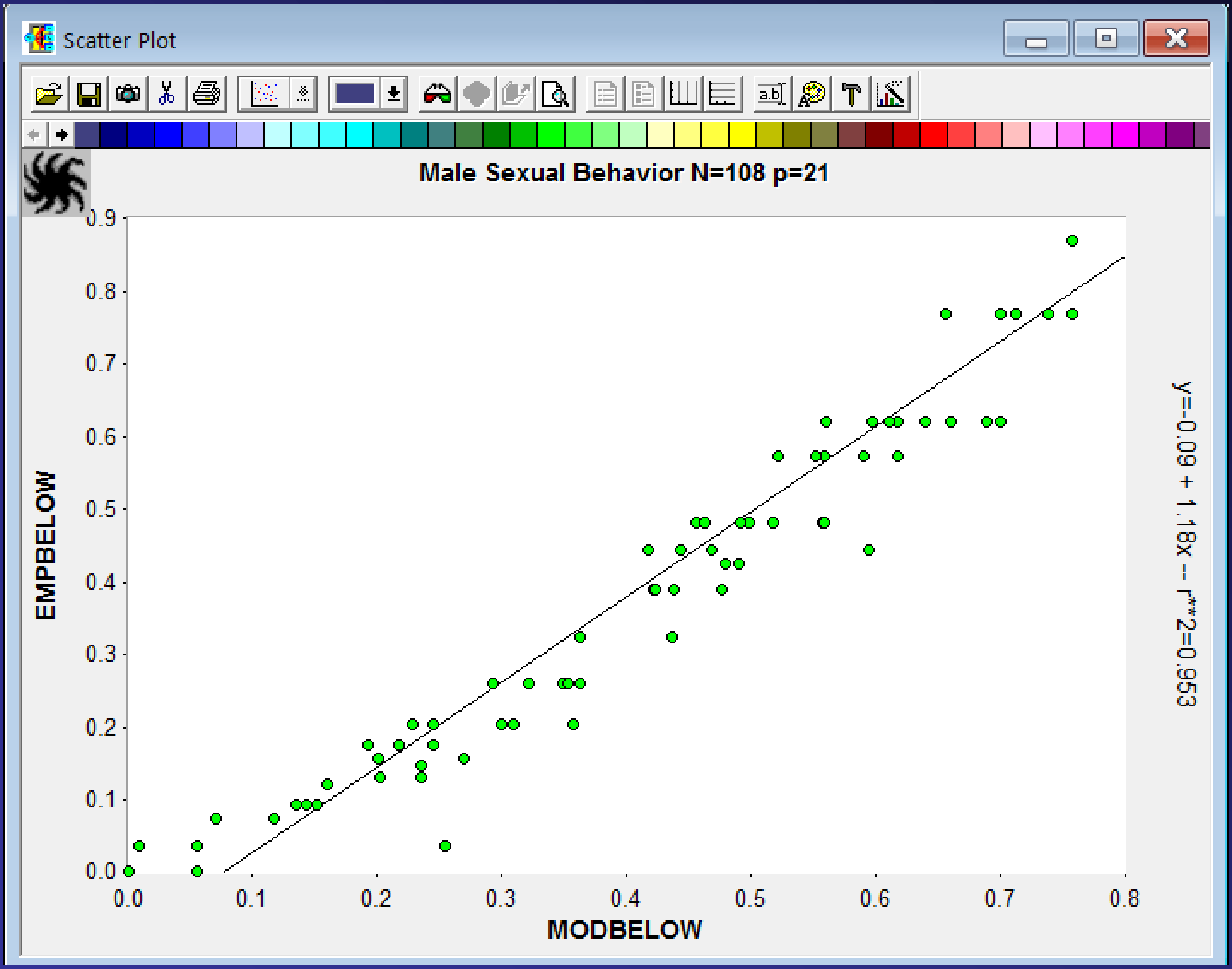
- Possibly, restrict $\hat{\beta}_i \geq 0$ and $\sum_{i=1}^{p} \hat{\beta}_i \leq \tau$. From $\hat{\beta}_i$ compute $\hat{\mu}_i$. The validity of the model is given by $\hat{R}^2$.
- If the model is valid, compute $\hat{\pi}_k = \sum_{i=1}^{p} \hat{\beta}_i x_{ki}$ as the model-based prop. below, get its CDF and obtain $\hat{z}_k$ scores

Example: Male Stature (Height) in cm (n= 1774)

15 artificial Guttman items created from national data.
AST model fitted, z-scores obtained, and height predicted.
Extreme binary data was all 1's, or all 0's – no Bayes



Example: Male Sexual Behavior
21 parameter AST model – Distribution free, no z



Mimic Model Estimation

- With enough items, items can be grouped into sets, each with a full range of item content, item means, and with its own total score X_{1T}, X_{2T}, …
- The several X_{1T}, X_{2T}, … can yield several proportions below, such as p_{1T}, p_{2T}, …
- A latent factor F can be created and a mimic model used in place of regression estimation

$\rightarrow p_{1T}$

$\rightarrow p_{2T}$

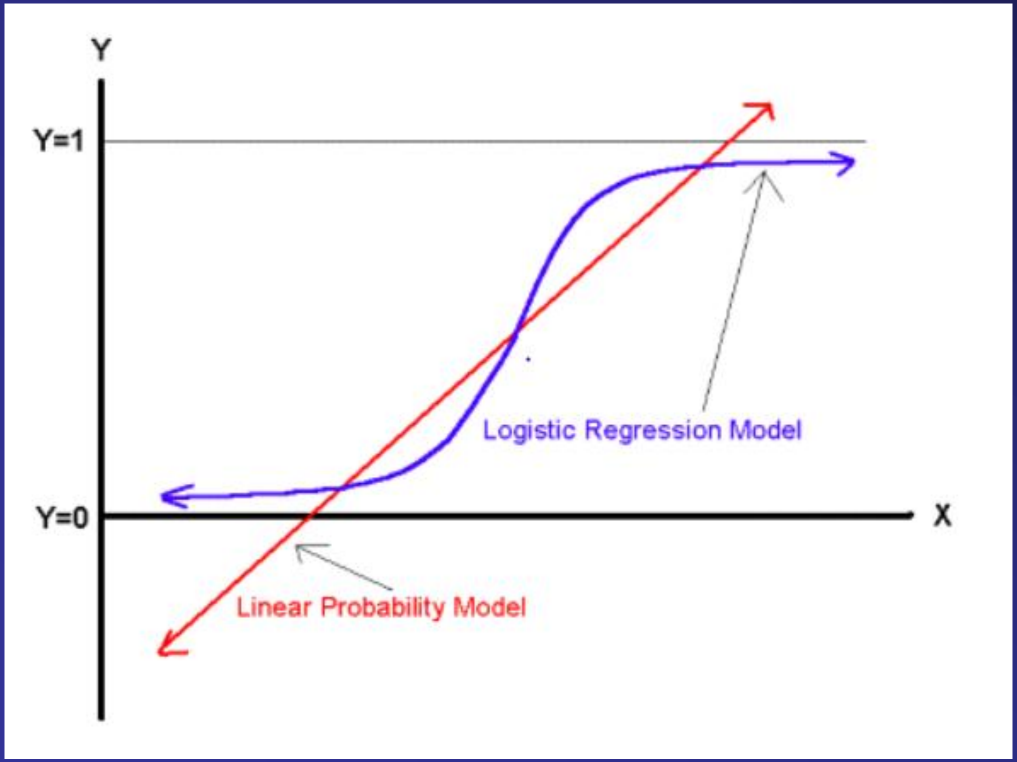
$\rightarrow p_{3T}$

…

$$F = \sum_{i=1}^{p} \beta_i x_{ki} + \varepsilon_k$$

Logistic Regression

We have been linearly predicting a limited dependent variable. Everyone recommends logistic regression instead, bounding the DV.



www.appstate.edu/~whiteheadjc/service/logit/logit.gif

$$L(X) = \frac{1}{1+e^{-X}} = \frac{e^{X}}{1+e^{X}} = \frac{\exp(X)}{1+\exp(X)}$$

Reminder: If $y = e^{x},\ x = \ln(y)$

Such a function is used in item response theory (IRT, Embretson & Reise, 2000). One curve is given for each item, say item i. It is usually called an item response function or item characteristic curve.

X relates to an underlying latent trait and L(X) is probability of a “1” (correct, yes, or other keyed response). (It is not a CDF, as before)

In IRT, X is often taken as a linear function of more basic parameters of item i so that

$$L(X) = P_i = (P_i = 1 \mid \text{other parameters})$$

where P_i is the probability of a person getting item i correct (response “1” vs. “0”).

We use a simple model for X, based on 2 parameters

$$X = a_i(\theta - b_i) = a_i\theta + d_i$$

a_i is a discrimination parameter, b_i is a difficulty parameter, and θ is a latent trait. We can think of θ as a factor of factor analysis.

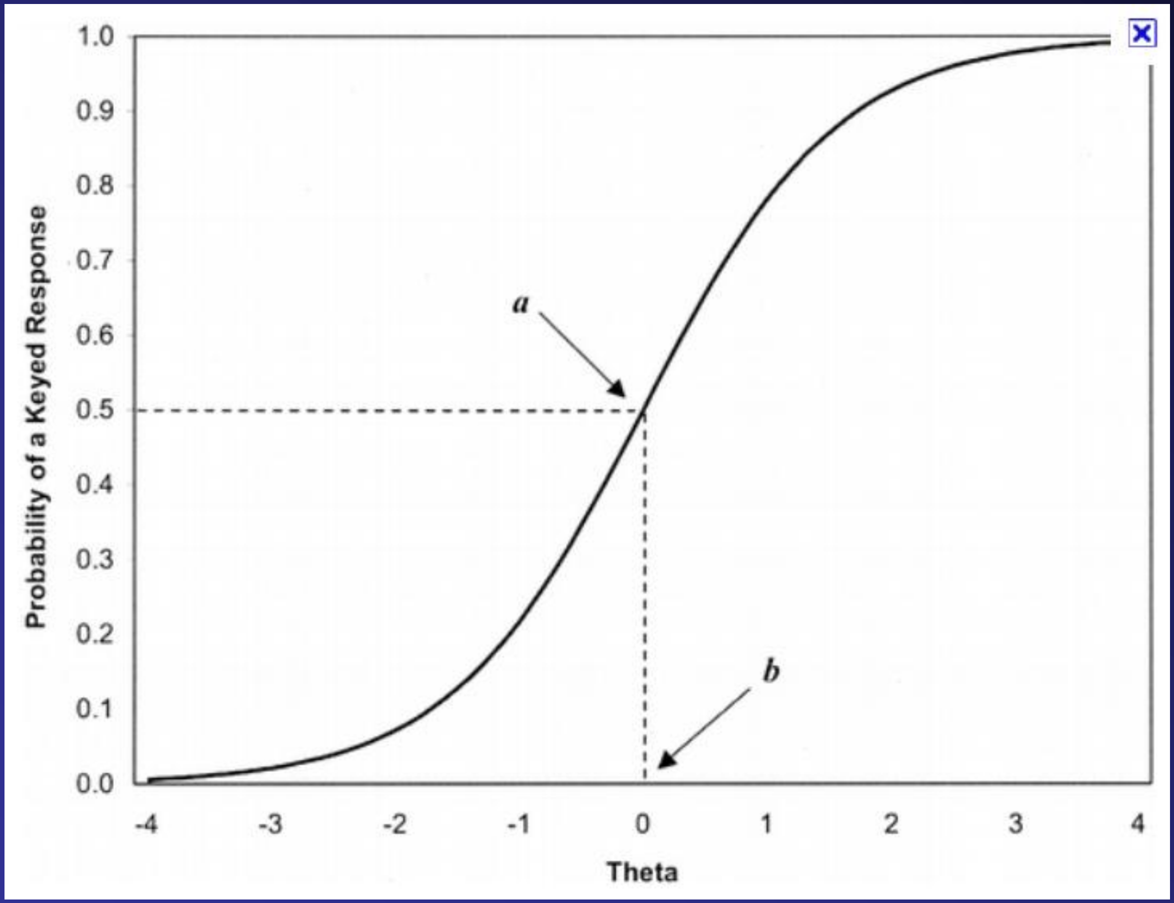
This means that $L(X) = \dfrac{e^X}{1+e^X} = \dfrac{\exp(X)}{1+\exp(X)}$

becomes $P_i(\theta) = \dfrac{\exp[a_i(\theta - b_i)]}{1+\exp[a_i(\theta - b_i)]}$

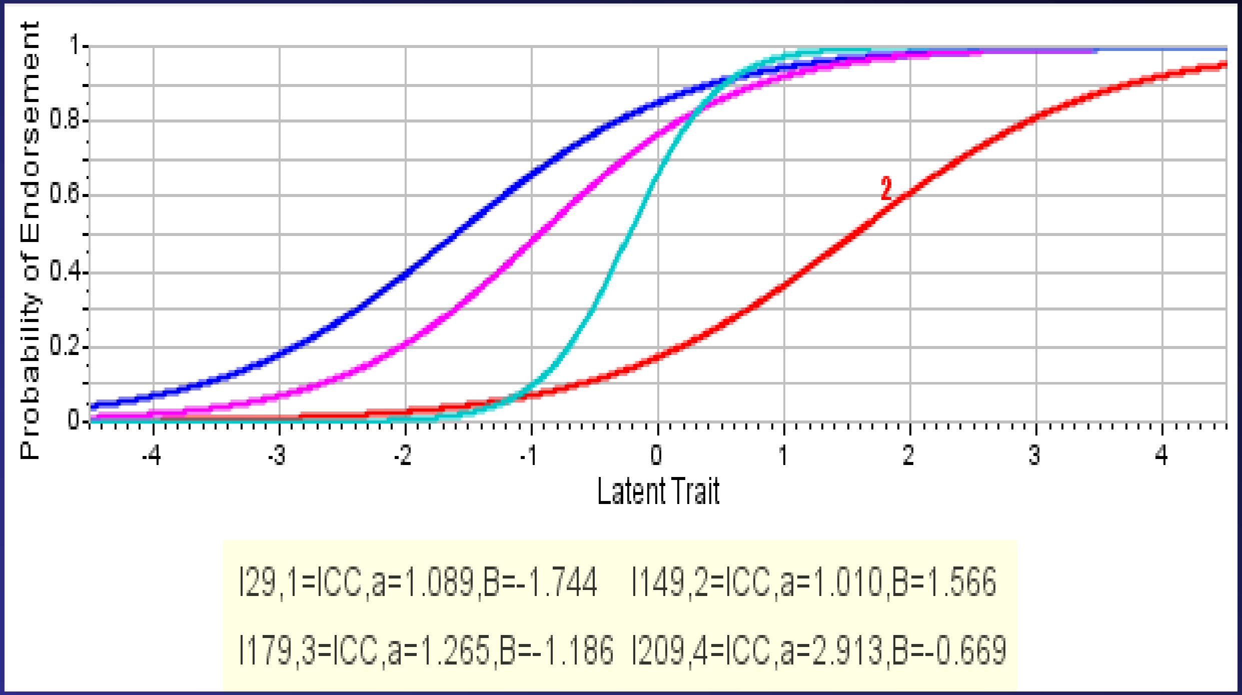
And thus $\ln\left(\dfrac{P_i(\theta)}{1-P_i(\theta)}\right) = a_i(\theta - b_i) = a_i\theta + d_i$

The *logit* (log-odds) of getting item i correct is a linear function of the trait level. It also depends on the 2 item features of discrimination and difficulty – a 2PL model.

An item response function for one 2PL item, showing where a and b parameters are read



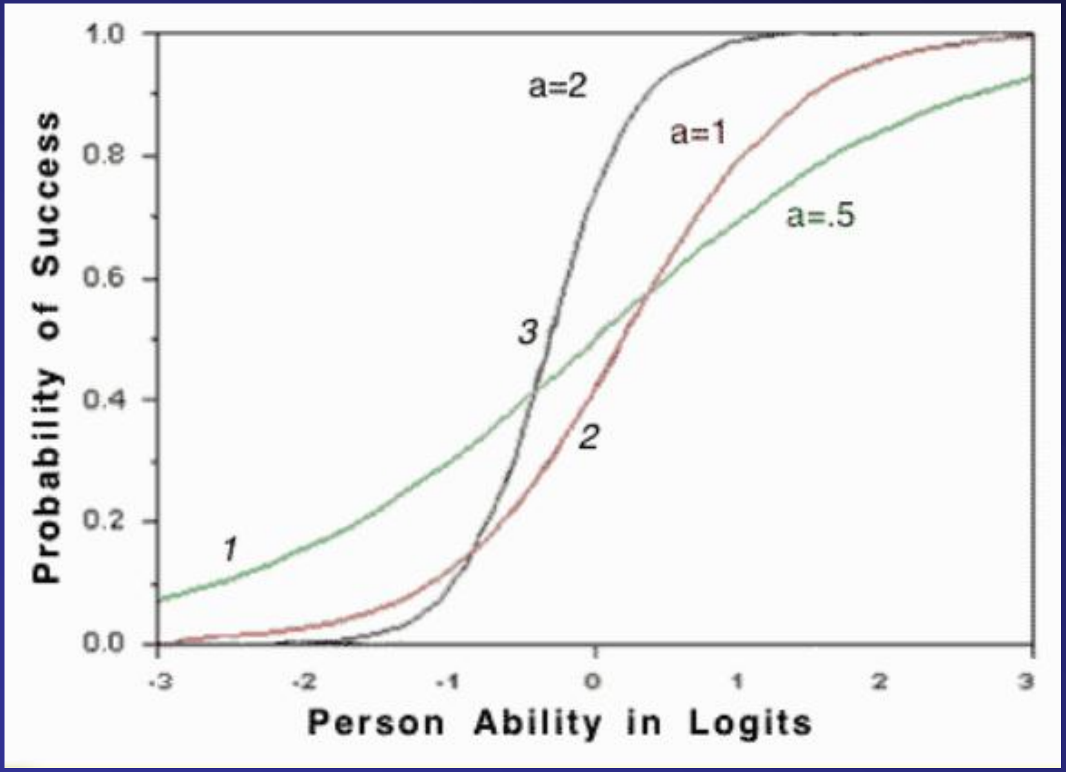
Ainsworth: www.csun.edu/~ata20315/psy427/Topic08_IntroIRT.ppt



from EQSIRT

Four 2PL item curves. The orange item is hardest (largest "b" parameter). The aqua item has the largest slope ("a" parameter)

Here is an example of 3 items with 3 different “a” parameters. The green item is easiest at low ability, but it is hardest at high ability – a critique of 2PL

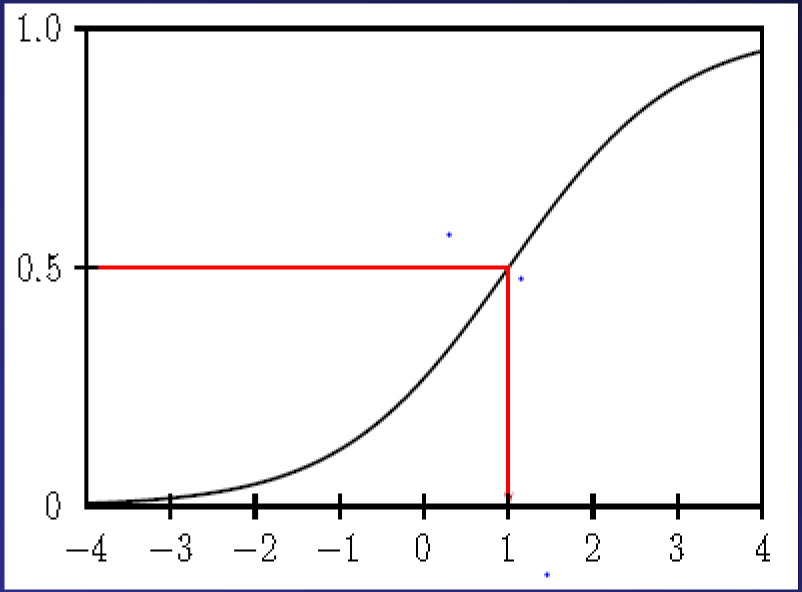


http://jalt.org/test/sic_5.htm

If a_i = 1, we get the 1PL or the Rasch (1960) Model

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}$$

If a person has $\theta > b_i$, their probability of a “1” (keyed direction) is greater than .5



Partchev: VisualIRT.pdf

So, IRT takes 0-1 item responses and explains those responses in terms of a latent trait (“math ability”, “extraversion,” etc.) and some item features. Of course there are more complicated models, e.g., they may have a “guessing” parameter, or deal with multicategory ordinal items, etc.

In practice, the parameters of the model have to be estimated, and perhaps also the latent trait scores θ for a set of persons

The model also has to fit real data

There a many additional features to IRT, such as item and test information. Those aspects exceed our limited goal here, which is to ask:

Does IRT yield Interval Scores?

If the model is valid, and (if necessary) the trait has the assumed distribution, I would say “yes”. Linear transformations make sense, and a fixed difference between two values of θ has the same meaning along the continuum

Others disagree.

In the 2PL, “with discrimination varying from item to item, the very meaning of the construct changes from point to point on the dimension… Measurement in its true sense has not been achieved” (Salzberger, 2002)

The Rasch model does not have these problems

Theorems exist for the Rasch model that prove interval scale status:

"…the parameters θ and β (b) are unique up to positive linear transformations with a common multiplicative constant i.e., they have interval scale properties with a common unit of measurement" (Fischer, 1995, p. 21)

No assumption on the distribution of the trait needs to be made (but is made for a typical estimation method)

Conclusions

- If an absolute simplex is a relevant model for binary data; or
- If assumptions such as unidimensionality, local independence (not reviewed here), etc. of item response theory or Rasch are met; and
- If the chosen model fits empirical data (by tests with high power; by fit indices)

Then it seems to me that interval-level scores ***can*** be obtained from binary responses.

Bentler, P. M. (1971). An implicit metric for ordinal scales: Implications for assessment of cognitive growth. In D. R. Green, M. P. Ford, & G. B. Flamer (Eds.), *Measurement and Piaget* (pp. 34-63). New York: McGraw Hill.

Bentler, P. M. (2009). Estimation, tests, and extensions of a 1-parameter item scaling model. Paper presented at SMEP, Salishan, OR

Bentler, P. M. (2011). SEM and regression estimation in the absolute simplex (Bentler-Guttman Scale). Paper presented at IMPS, Hong Kong.

Bentler, P. M., & Wu, E. (in prep). *EQSIRT*. Encino: Multivariate Software.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch Models* (pp. 15-38). New York: Springer.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139-150.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Salzberger, T. (2002). The illusion of measurement: Rasch versus 2-PL. *Rasch Measurement Transactions*, *16*, p. 882.

Treiblmaier, H., Bentler, P. M., & Mair, P. (2011). Formative constructs implemented via common factors. *Structural Equation Modeling*, *18*, 1-17.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, *47*, 65-72.