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

$$X_T = X_1 + X_2 + \ldots + X_p$$

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 controversial (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>
<thead>
<tr>
<th>Transformation</th>
<th>Default</th>
<th>Name of</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Y+a)</td>
<td>$a = 0$</td>
<td>Log_Y</td>
<td>$\log(Y + a)$, $Y + a &gt; 0$</td>
</tr>
<tr>
<td>log10(Y+a)</td>
<td>$a = 0$</td>
<td>Log10_Y</td>
<td>$\log_{10}(Y + a)$, $Y + a &gt; 0$</td>
</tr>
<tr>
<td>sqrt(Y+a)</td>
<td>$a = 0$</td>
<td>Sqrt_Y</td>
<td>$\sqrt{Y + a}$, $Y + a &gt; 0$</td>
</tr>
<tr>
<td>exp(Y)</td>
<td></td>
<td>Exp_Y</td>
<td>$\exp(Y)$</td>
</tr>
<tr>
<td>power(Y;a)</td>
<td>$a = 1$</td>
<td>Pow_Y</td>
<td>$Y^a$, $Y &gt; 0$ if $a$ is not integral</td>
</tr>
<tr>
<td>arcsinh(Y)</td>
<td></td>
<td>Arcsinh_Y</td>
<td>$\log(Y + \sqrt{Y^2 + 1})$</td>
</tr>
<tr>
<td>Box-Cox(Y;a)</td>
<td>MLE</td>
<td>BC_Y</td>
<td>See text.</td>
</tr>
</tbody>
</table>

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

$$BC(y; \lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}$$
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 $P(x) = 0.975$

$z$-value is $1.96$

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

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

- 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:

$$S_m = S + \bar{XX}'$$
Under AST, $\Sigma_m$ is a patterned matrix. With $\mu_1 \leq \mu_2 \leq \mu_3 \ldots$

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

- The entire matrix is a function of one parameter per item
- Items can be ordered by this matrix
- Structural equation modeling fits $\Sigma_m$ via $S_m$
For the 10x5 binary data given earlier, $S_m$ is

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

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

The means and SSCP can be fit by 1 par./item
Example Moment Matrix $S_m$:
Male Sexual Behavior (N=175)

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>.891</td>
<td>.771</td>
<td>.697</td>
<td>.583</td>
<td>.566</td>
<td>.497</td>
<td>.394</td>
<td>.377</td>
</tr>
<tr>
<td>V2</td>
<td>.771</td>
<td>.789</td>
<td>.686</td>
<td>.577</td>
<td>.566</td>
<td>.491</td>
<td>.377</td>
<td>.377</td>
</tr>
<tr>
<td>V3</td>
<td>.697</td>
<td>.686</td>
<td>.709</td>
<td>.571</td>
<td>.531</td>
<td>.497</td>
<td>.383</td>
<td>.377</td>
</tr>
<tr>
<td>V4</td>
<td>.583</td>
<td>.577</td>
<td>.571</td>
<td>.594</td>
<td>.526</td>
<td>.463</td>
<td>.371</td>
<td>.383</td>
</tr>
<tr>
<td>V5</td>
<td>.566</td>
<td>.566</td>
<td>.531</td>
<td>.526</td>
<td>.577</td>
<td>.429</td>
<td>.360</td>
<td>.366</td>
</tr>
<tr>
<td>V6</td>
<td>.497</td>
<td>.491</td>
<td>.497</td>
<td>.463</td>
<td>.429</td>
<td>.509</td>
<td>.366</td>
<td>.366</td>
</tr>
</tbody>
</table>
There are many ways to fit the absolute simplex model to data. Here are four:

1. A symmetric matrix with equalities
2. \[ \Sigma_m = TD_{\mu_{\text{diff}}} T' \] where \( T \) is lower triangular with 1’s and \( D_{\mu_{\text{diff}}} = \text{diag}\{\mu_1,...(\mu_i - \mu_{i-1}),...\} \)
3. A simplex model with variances \( D_{\mu_{\text{diff}}} \)
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 \( \rightarrow \) order
- Column SDs \( \rightarrow \) 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 \((\Sigma_m)\) is inflated in quasi-simplex, so order items by off-diagonal entries only, e.g., ignore \(D\) in \(\Sigma_m = TD_{\mu_{\text{diff}}} T' + D\)
Ordered Total Scores Generate the CDF

Motivation: Note that

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Score</th>
<th>$X_T$</th>
<th>% Below</th>
<th>CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000</td>
<td>0</td>
<td></td>
<td>0</td>
<td>.10</td>
</tr>
<tr>
<td>00001</td>
<td>1</td>
<td></td>
<td>10</td>
<td>.20</td>
</tr>
<tr>
<td>00011</td>
<td>2</td>
<td></td>
<td>20</td>
<td>.50</td>
</tr>
<tr>
<td>00111</td>
<td>3</td>
<td></td>
<td>50</td>
<td>.80</td>
</tr>
<tr>
<td>01111</td>
<td>4</td>
<td></td>
<td>80</td>
<td>.90</td>
</tr>
<tr>
<td>11111</td>
<td>5</td>
<td></td>
<td>90</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Person Ordering by $X_T$ is Free of Item Weights – Any Weighted Sum is OK

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

\[
\begin{align*}
11111 & \quad X_T = w_1 + w_2 + w_3 + w_4 + w_5 \\
01111 & \quad = w_2 + w_3 + w_4 + w_5 \\
00111 & \quad = w_3 + w_4 + w_5 \\
00011 & \quad = w_4 + w_5 \\
00001 & \quad = w_5 \\
00000 & \quad = 0
\end{align*}
\]

The % of subjects below a pattern = % below a total score, no matter what the item weighting
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 $\rightarrow$ true CDF as $n$ gets large.
Regression Estimation of Absolute Simplex

Let \( \pi = \text{proportion below} \ (=\%\text{below}/100) \)
\[ x' = (x_1, x_2, \ldots, x_p) \]
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 \( \pi_k \) exactly with \( R^2 = 1.0 \). Adding items with \( \beta_i > 0 \), \( \pi \) becomes continuous as \( p \to \infty \). If also \( n \to \infty \) then \( \pi \) approaches the population trait CDF.
The probabilities $\pi$ are then transformed to a normal $z$-statistic, the interval score of interest. From this viewpoint*, $\pi$ and $z$ are formative measures -- they arise from the item responses.

In contrast, most extant measures are best considered 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 $k=1,\ldots,n$. 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 $X_{1T}, X_{2T}, \ldots$
- The several $X_{1T}, X_{2T}, \ldots$ can yield several proportions below, such as $p_{1T}, p_{2T}, \ldots$
- A latent factor $F$ can be created and a mimic model used in place of regression estimation

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

We have been linearly predicting a limited dependent variable. Everyone recommends logistic regression instead, bounding the DV.
\[ L(X) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} = \frac{\exp(X)}{1 + \exp(X)} \]

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 $\theta$ is a latent trait. We can think of $\theta$ as a factor of factor analysis.
This means that \[ L(X) = \frac{e^X}{1 + e^X} = \frac{\exp(X)}{1 + \exp(X)} \]

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

And thus \[ \ln\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right) = a_i(\theta - b_i) = a_i \theta + d_i \]

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
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.
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 $\theta$ for a set of persons

The model also has to fit real data
There are 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 $\theta$ 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 $\theta$ and $\beta$ (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.


