# A Scaled Logistic Quasi-Simplex is a Football and its Stress is not a Function of the Number of Points

Leland Wilkinson and G.V. Ramanathan
University of Illinois at Chicago

November, 1977

**Abstract**

It is shown that logistically distributed responses scaled in two or more dimensions resemble a football, not a horseshoe. Further, the stress of the scaled solution is a function of the slope of the response functions, not the number of responses.

Kendall (1971) and others have noted that a quasi simplex scaled in two or more dimensions frequently resembles a horseshoe because extreme distances are usually truncated in real data. We show here that for data fitting a logistic response model (e.g., Rasch, 1960), squared Euclidean distances between items scale as a football, not a horseshoe. Further, the stress of the scaled model is a function of the slope of the item response functions, not the number of items. We use test theory terminology to motivate the notation; other logistic response models fit with minor modification.

Consider a collection of $N$ subjects divided according to ability into $g$ groups consisting of $n_1, n_2, \ldots, n_g$ subjects, respectively. Assume they are tested on $m$ items arranged in increasing order of difficulty. Let $a_{ijk}$ be a random variable which is the score (0 or 1) of the $i^{th}$ subject $(1 < i < n_j)$ in group $j$ $(1 < j < g)$ on item $k$ $(1 < l < m)$. We assume that the probability that $a_{ijk}$ is 1, denoted by $p_{jk}$, is independent of $i$.

In the logistic response model,

$$p_{j,k} = \frac{e^{b(j-k)}}{1 + e^{b(j-k)}} \tag{1}$$

where $b$ is a positive constant representing the slope of the item characteristic functions. Now, let the squared Euclidean distance between items $r$ and $s$ so defined be

$$d_{r,s} = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (a_{ijr} - a_{ijs})^2 / N \tag{2}$$

In the following theorem, we show that the expected value of $d_{r,s}$ $(r \neq s)$ can be scaled to obtain a directed distance function on the set of items so that they lie on a straight line. We will then compute the variance of these distances to inflate the straight line to a football.

**Theorem.** *Let $a_{ijk}$ be a discrete set of random variables as defined above and let $d_{r,s}$ be defined as in (2). Then the expected value of*

$\delta_{rs} = \tanh[(b/2)(r-s)]d_{rs}$

*is additive in the sense that, for any non-negative integers $r, s, t \leq m$,*

$E(\delta_{rs}) + E(\delta_{st}) = E(\delta_{rt})$.

*Proof.* Using results in Lord and Novick (1968) and Guttman (1969), it can be shown that the expected value of $d_{r,s}$ is

$$E(d_{rs}) = (1/N) \coth[(b/2)(r-s)] \sum_{j=1}^{g} n_j \left( \frac{1}{1+e^{b(s-j)}} - \frac{1}{1+e^{b(r-j)}} \right) \tag{3}$$

And $E(\delta_{rs}) = \tanh[(b/2)(r-s)]E(d_{rs})$ , so

$$E(\delta_{rs}) = (1/N) \sum_{j=1}^{g} n_j \left( \frac{1}{1+e^{b(s-j)}} - \frac{1}{1+e^{b(r-j)}} \right) \tag{4}$$

It follows from (4) that $E(\delta_{rs}) + E(\delta_{st}) = E(\delta_{rt})$. □

We now compute the variance.

$$\sigma^2(d_{rs}) = (1/N^2) \left\{ E\left( \left[ \sum_{j=1}^{g} \sum_{i=1}^{n_j} (a_{ijr} - a_{ijs})^2 \right]^2 \right) - \left[ E\left( \sum_{j=1}^{g} \sum_{i=1}^{n_j} (a_{ijr} - a_{ijs})^2 \right) \right]^2 \right\}$$

Because any two distinct $a_{ijk}$'s are independent,

$$\begin{aligned}
\sigma^2(d_{rs}) &= (1/N^2)\left\{\sum_{j=1}^{g}\sum_{i=1}^{n_j} E\left((a_{ijr}-a_{ijs})^4\right) - \sum_{j=1}^{g}\sum_{i=1}^{n_j}\left[E\left((a_{ijr}-a_{ijs})^2\right)\right]^2\right\} \\
&= (1/N^2)\sum_{j=1}^{g} n_j\left\{(p_{jr}+p_{js}-2p_{jr}p_{js}) - (p_{jr}+p_{js}-2p_{jr}p_{js})^2\right\} \\
&= (1/4N^2)\sum_{j=1}^{g} n_j\left\{1-[1-2(p_{jr}+p_{js}-2p_{jr}p_{js})]^2\right\}
\end{aligned}$$

But,

$$\begin{aligned}
1-2(p_{jr}+p_{js}-2p_{jr}p_{js}) &= 1-2[p_{jr}(1-p_{js})+p_{js}(1-p_{jr})] \\
&= 1-\frac{2[e^{b(j-r)}+e^{b(j-s)}]}{[1+e^{b(j-r)}][1+e^{b(j-s)}]} \\
&= \frac{[1-e^{b(j-r)}][1-e^{b(j-s)}]}{[1+e^{b(j-r)}][1+e^{b(j-s)}]} \\
&= \tanh[(b/2)(j-r)]\tanh[(b/2)(j-s)]
\end{aligned}$$

Thus, the variance of $d_{rs}$ can be written as

$$\sigma^2(d_{rs}) = (1/4N^2)\sum_{j=1}^{g} n_j\left(1-\tanh^2[(b/2)(r-j)]\tanh^2[(b/2)(s-j)]\right) \quad (5)$$

The variance of the scaled distance is therefore

$$\sigma^2(\delta_{rs}) = \tanh^2[(b/2)(r-s)]\sigma^2(d_{rs}) \quad (6)$$

Finally, we show that for fixed $r-s$, as $r$ is moved from 1 to $k$ and when $r$ becomes sufficiently small or large compared to $g$, the variance decreases. We assume for this demonstration that $n_j$ is a constant (say, $n$). Then

$$\begin{aligned}
\sigma^2(\delta_{r+1,s+1}) &= n(\tanh^2[(b/2)(r-s)]/4N^2) \\
&\quad \sum_{j=1}^{g}\left(1-\tanh^2[(b/2)(r+1-j)]\tanh^2[(b/2)(s+1-j)]\right) \\
&= n(\tanh^2[(b/2)(r-s)]/4N^2) \\
&\quad \sum_{j=0}^{g-1}\left(1-\tanh^2[(b/2)(r-j)]\tanh^2[(b/2)(s-j)]\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}\sigma^2(\delta_{r+1,s+1}) - \sigma^2(\delta_{r,s}) &= n(\tanh^2[(b/2)(r-s)]/4N^2) \\ &\quad \tanh^2[(b/2)(r-g)]\tanh^2[(b/2)(s-g)] \\ &\quad -\tanh^2[(b/2)r]\tanh^2[(b/2)s]\end{aligned}$$

If $r, s < g/2$, this is positive and if $r, s > g/2$ it is negative. If the interval between $r$ and $s$ overlaps $g/2$, it can be of either sign. The football is now inflated.

A consequence of these results is that the stress of scaled squared Euclidean distances between logistically distributed items does not depend on the number of items. The term in the last brackets in (5) is positive and less than 1, so $\sigma^2(d_{rs})$ is always less than $1/4N$; the variance in (6) is likewise bounded by $1/4N$. With $N$ and $b$ fixed, the variances of the squared distances between a new item and the others are a function of its location on the unidimensional scale. Thus, tables from Monte Carlo studies of the null distribution of a stress statistic cannot be used to test unidimensionality of scaled logistic items. Tables for this purpose should be conditioned on $N$ and some estimate of the slope parameter $b$ or a function of it, such as the average biserial correlation among items. Finally, since the logistic and normal CDF's are so similar, these results should apply in the latter case. For nonmetric scaling, they may apply to an even wider class of monotonic characteristic functions.

# References


[1] Guttman, L. (1969). Review of F.M. Lord and M.R. Novick, Statistical Theory of Mental Test Scores. *Psychometrika, 34*, 398-404.

[2] Kendall, D.G. (1971). Seriation from abundance matrices. In: F.R. Hodson, D.G. Kendall, and P. Tautu (Eds) *Mathematics in the Archaeological and Historical Sciences*. Edinburgh: Edinburgh University Press.

[3] Lord, F.M. and Novick, M.R. (1968). *Statistical Theory of Mental Test Scores*. Reading, MA: Addison-Wesley.

[4] Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedogogiske Institute.