17th International Meeting of the **Psychometric Society**

Hong Kong 2011

How Item Response Theory can solve problems of ipsative data

Doctoral student: Anna Brown

Advisor: Prof. Alberto Maydeu-Olivares

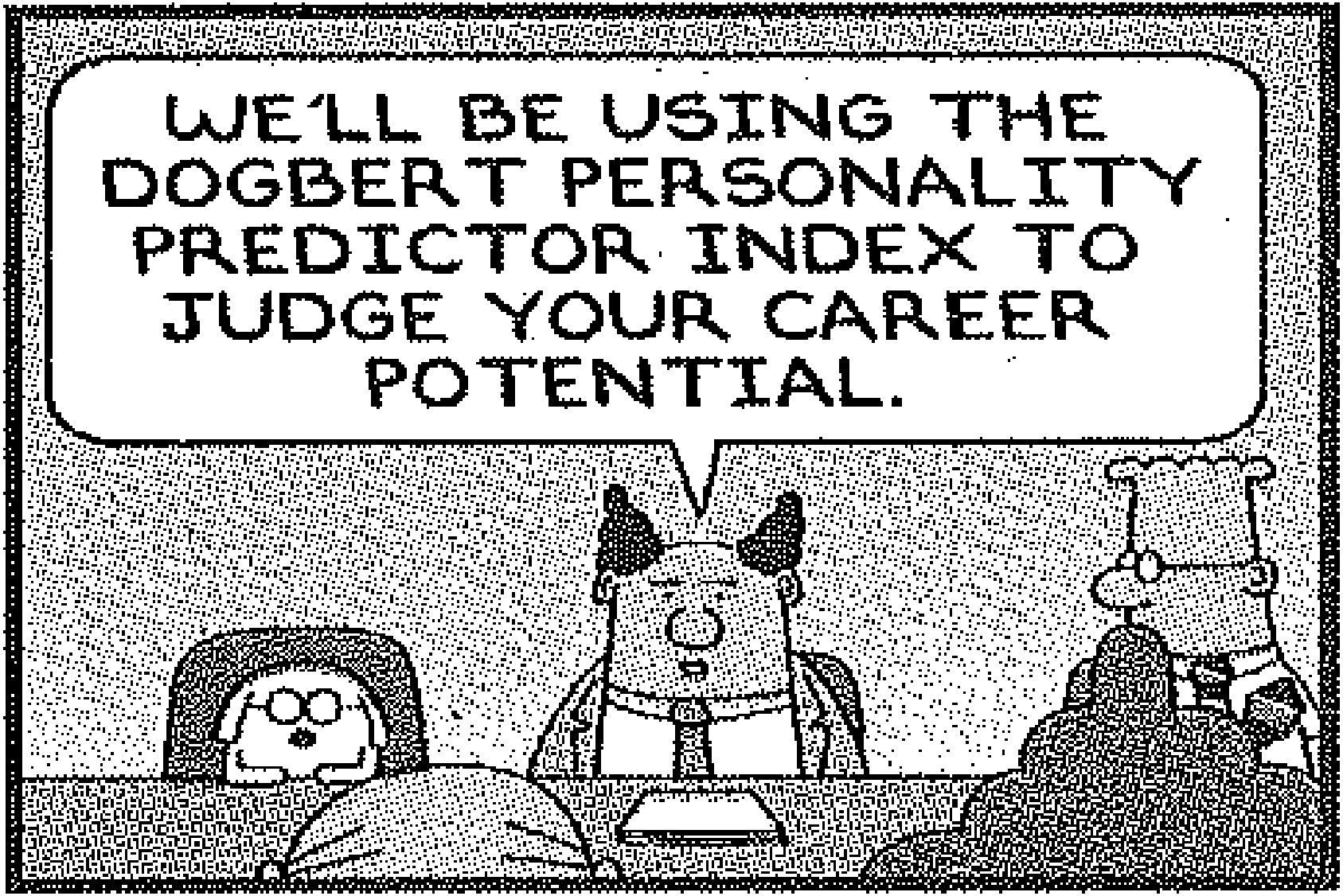
UNIVERSITAT DE BARCELONA

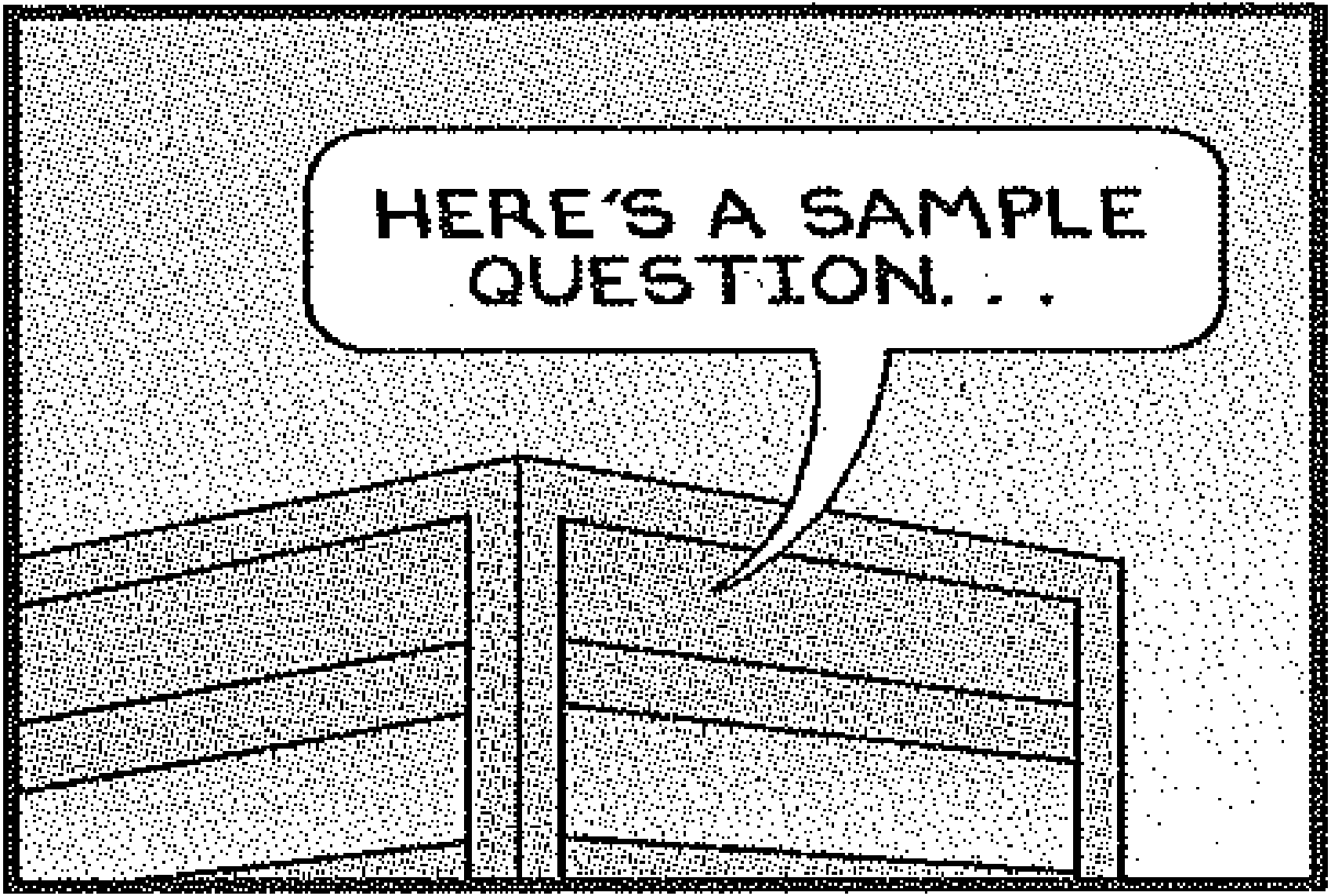
Motivation for this work

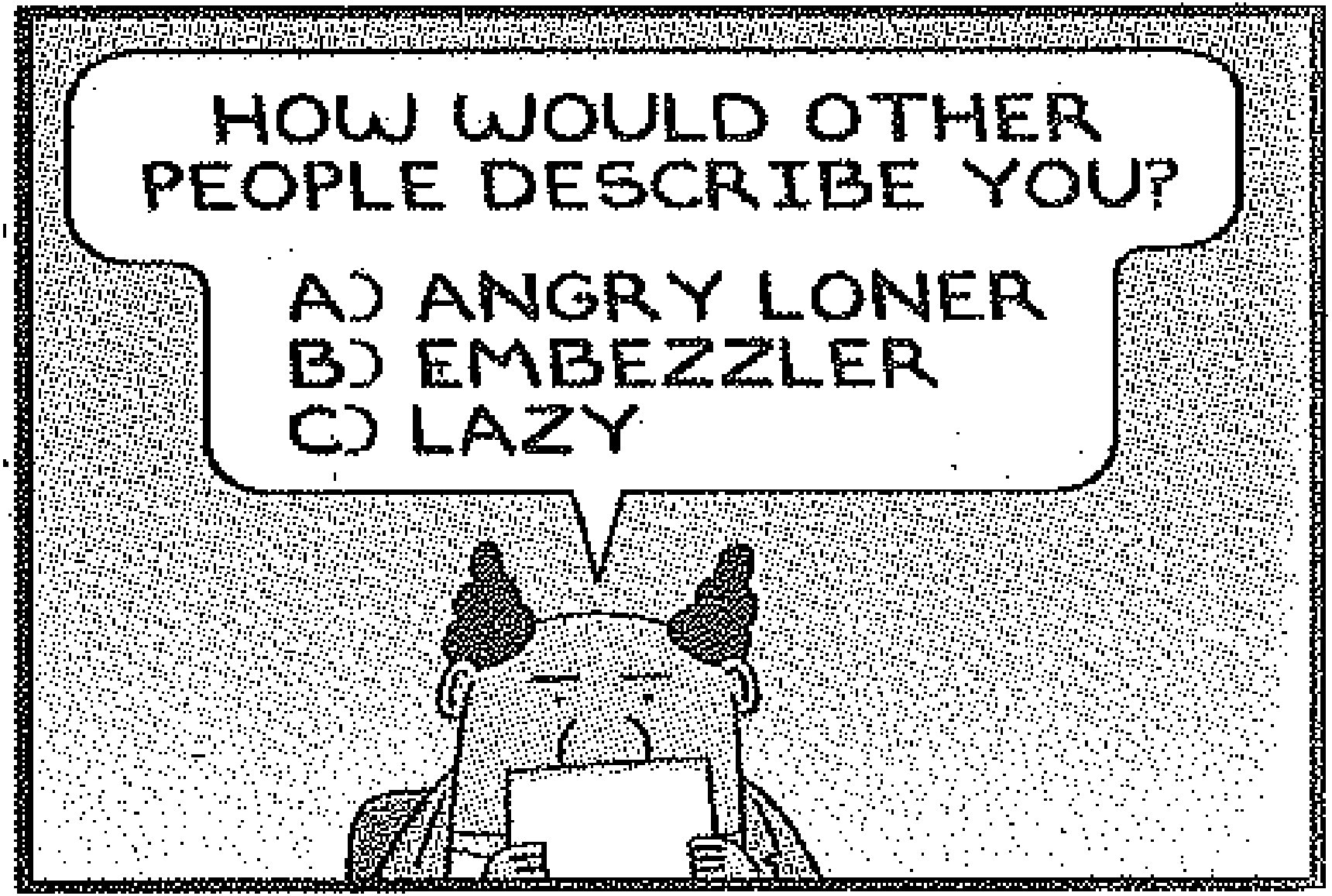
- Practical problem that desperately needed a solution
  - Old
  - Very widespread (mainly workplace selection and assessment tools, millions of administrations per year)
  - Thousands of pages in journals over years have been devoted to the problem
  - Psychometrics with all its sophisticated methodologies had failed to provide a solution

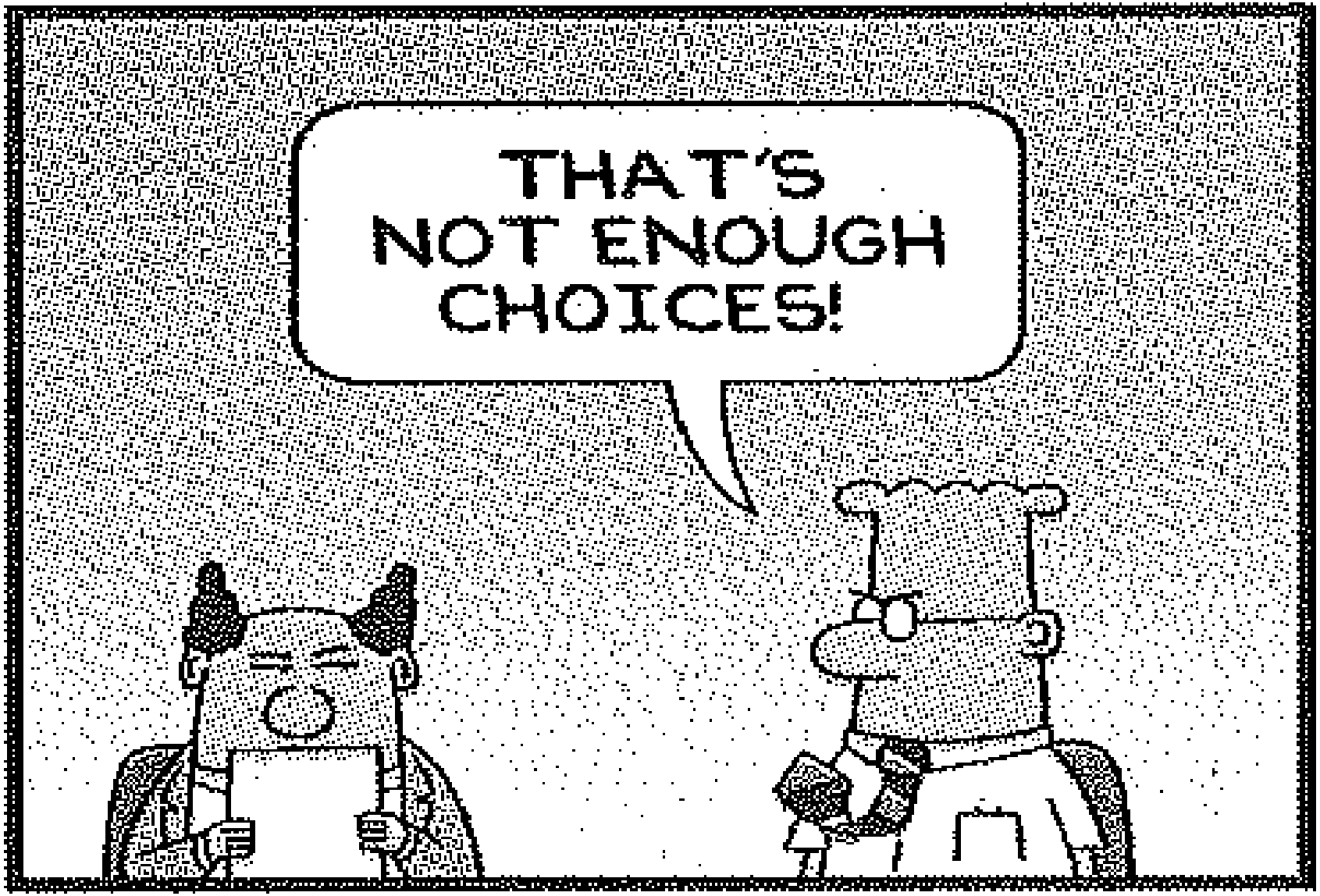
Curse of ipsative data

Forced-choice response format


WE'LL BE USING THE DOGBERT PERSONALITY PREDICTOR INDEX TO JUDGE YOUR CAREER POTENTIAL.


HERE'S A SAMPLE QUESTION. . .


HOW WOULD OTHER PEOPLE DESCRIBE YOU?
A) ANGRY LONER
B) EMBEZZLER
C) LAZY


THAT'S NOT ENOUGH CHOICES!


SAYS
THE
ANGRY
LONER.

Forced-choice format

- Multidimensional Forced-Choice (MFC) format
  - Rank-order two or more items from different dimensions

A. I manage to relax easily

B. I am careful over detail

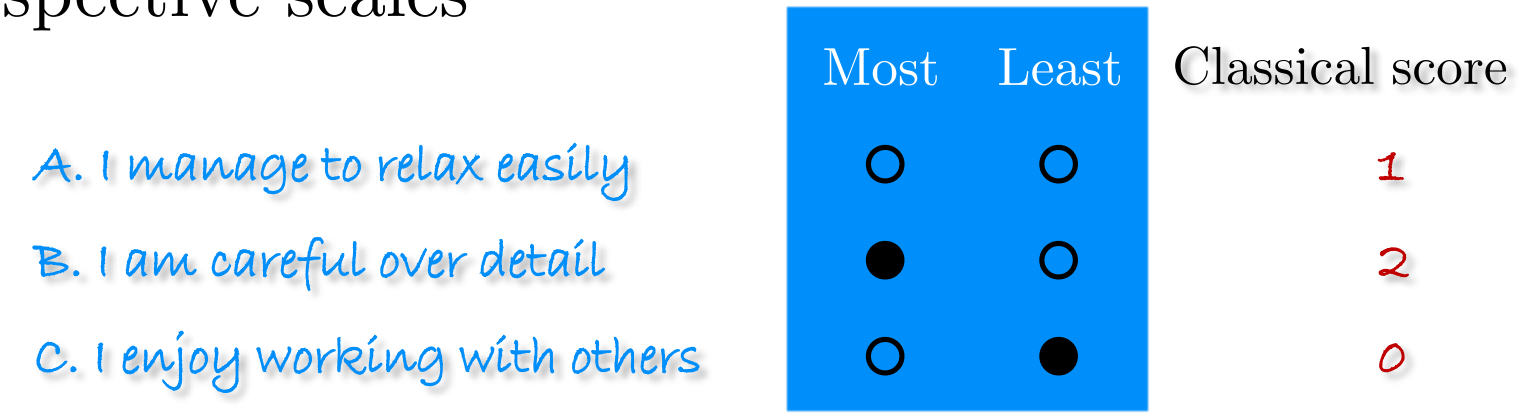
C. I enjoy working with others

Advantages of comparative formats

- Designed to reduce response biases
  - Direct comparison overcomes problems with interpretation of the rating scale
  - Impossible to endorse all items thus:
    - eliminates uniform biases / response sets (acquiescence, extreme/central tendency responding)
    - Reduces halo/horn effects
    - May reduce effects of socially desirable responding

Classical scoring of FC formats

- Inverted rank orders of items are added to their respective scales



  - The same number of points in each block is allocated for any individual
  - The total score on the test is constant for each individual (*ipsative* data)

Problems of ipsative data

1. Scores are relative
   - Impossible to get all high/low scores
   - Intra-individual comparisons are problematic
2. Construct validity is distorted
   - Variance of the total test score is zero
   - Negative average scale inter-correlation $\bar{r} = \frac{-1}{d-1}$
3. Criterion-related validity is distorted
   - Correlations with an external criterion must sum to zero
   - Compensatory correlations
4. Reliability estimates are distorted
   - Basic assumptions are violated (Cronbach's alpha and other coefficients)

Inadequate scoring

- Classical scoring methodology is inadequate for forced-choice items
  - Rankings are treated as ratings (i.e. relative scores are treated as absolute)
  - Items within each block are NOT assessed independently
- Need to radically depart from classical scoring schemes
  - Modelling the psychological process of responding to forced-choice items is the key to making sense of comparative data
- Suitable psychological models for such data have existed for a long time, and they are well known

Modelling decision process behind responding to forced-choice items

Thurstonian IRT model

Psychological value

- Louis Thurstone (1927-1929) introduced the notion of a *psychological value* or *utility*
  - Describes "the affect that the object calls forth";
  - Varies across individuals for the same object, and across objects within the same individual;
  - Can be placed on a *psychological continuum;*
  - Assumed normally distributed across individuals.
- The notion of *utility maximisation*
  - when confronted with choosing between two items, respondents will choose the item with the highest psychological value (utility).

Law of comparative judgement (1927)

- A respondent prefers item i to item k, if her or his utility t_i is larger than t_k

$$y_l = y_l \quad i,k \quad = \begin{cases} 1, & \text{if} \quad t_i \geq t_k \\ 0, & \text{if} \quad t_i < t_k \end{cases}$$

- The difference of two utilities $y_l^* = t_i - t_k$ is normally distributed
- Binary outcome of comparison y_l linked to y_l^* through a threshold process

$$y_l = y_l \quad i,k \quad = \begin{cases} 1, & \text{if} \quad y_l^* \geq 0 \\ 0, & \text{if} \quad y_l^* < 0 \end{cases}$$

Binary coding of ranking data

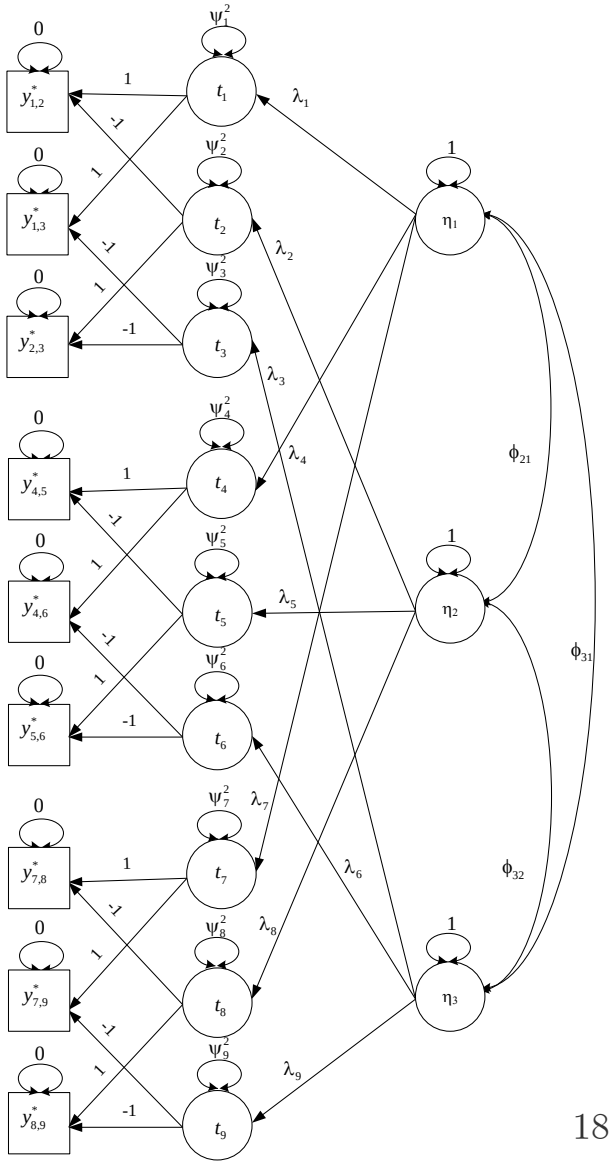
- Any ordering or ranking of n choice alternatives requires $n(n-1)/2$ separate comparison judgements:
  - Rank-ordering of 2 items elicits 1 comparison: {A,B}
  - Rank-ordering of 3 items elicits 3 comparisons: {A,B} {A,C} {B,C}
  - Rank-ordering of 4 items elicits 6 comparisons: {A,B} {A,C} {A,D} {B,C} {B,D} {C,D}
  - Etc.
  - "Most" – "least" format with 4 or more alternatives is partial ranking
- Any comparison {***i***, ***k***} is coded as 1 if ***i*** preferred to ***k***, and 0 otherwise

- Ordering {B, A, C} can be equivalently presented as 3 outcomes

{A, B}	{A, C}	{B, C}
0	1	1

Thurstonian factor models

- Maydeu-Olivares (1999); Maydeu-Olivares & Böckenholt (2005)
- Outcomes of comparisons are determined by the difference in utilities (no error terms)
- Second-order factors (traits) can be modelled
- Identification constraints are needed
  - Fixing uniqueness of one utility per block
  - Factor variances are fixed to 1
- Special identification cases
  - Pairs of items



IRT reparameterization

- Thurstonian second-order models cannot be used directly in person-centric applications
  - we are interested in persons' traits (second-order factors), not the utilities (first-order factors)
  - but the latent traits cannot be estimated
- Re-parameterization as an IRT model (first-order)
- Utilities of items ***i*** and ***k*** are functions of underlying factors (traits) ***a*** and ***b***:

$$t_i = \mu_i + \lambda_i \eta_a + \varepsilon_i \qquad t_k = \mu_k + \lambda_k \eta_b + \varepsilon_k$$

- Latent difference of utilities $y_l^* = t_i - t_k$ is a function of the traits:

$$y_l = t_i - t_k = (\mu_i - \mu_k) + (\lambda_i \eta_a - \lambda_k \eta_b) + (\varepsilon_i - \varepsilon_k) =$$
$$= -\gamma_k + (\lambda_i \eta_a - \lambda_k \eta_b) + (\varepsilon_i - \varepsilon_k)$$

Item response function

- The IRF for the binary outcome variable y_l, which is the result of comparison between items i and k measuring traits η_a and η_b, is

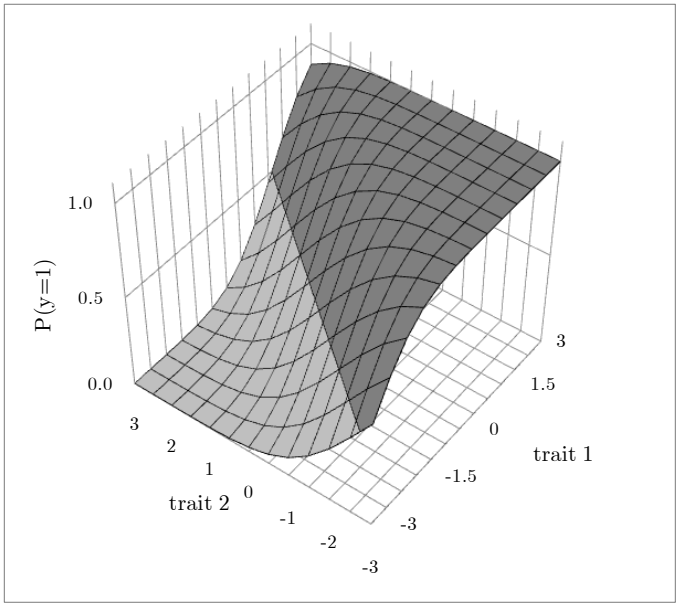
$$\Pr\ y_l = 1 \left| \eta_a, \eta_b \right. = \Phi\left(\frac{-\gamma_l + \lambda_i \eta_a - \lambda_k \eta_b}{\sqrt{\psi_i^2 + \psi_k^2}} \right)$$

- In intercept / slope form:

$$\Pr\ y_l = 1 \left| \eta_a, \eta_b \right. = \Phi\ \alpha_l + \beta_i \eta_a - \beta_k \eta_b$$

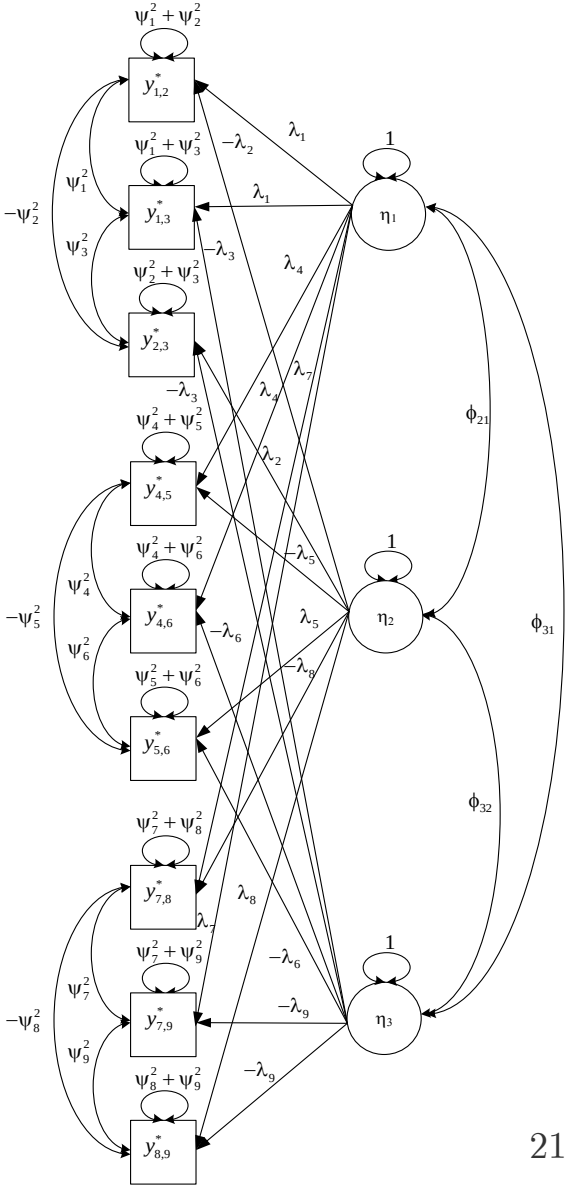
- Special case of same-trait (one-dimensional) comparisons

$$\Pr\ y_l = 1 \left| \eta \right. = \Phi\ \alpha_l + (\beta_i - \beta_k)\eta$$



Thurstonian IRT model

- Outcomes of comparisons are indicators of common factors (traits)
- Special features for blocks of 3 or more items
  - Factor loadings are structured
  - Uniquenesses are structured
  - Structured local dependencies
- Identification constraints are the same as for the second-order model



Estimation and scoring

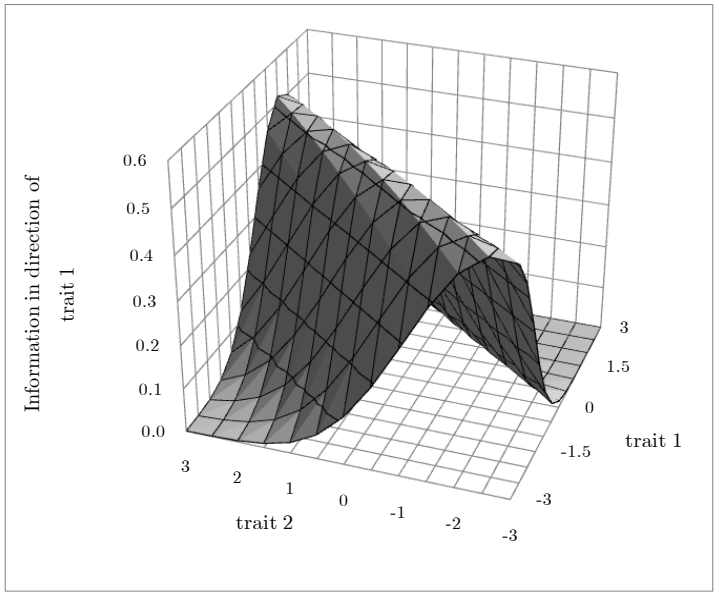
- Estimated with general-purpose SEM software M*plus* (Muthén & Muthén, 1998-2010)
- Limited information methods are the only option for most applications
  - When partial ranking format is used, Bayesian MI are recommended
- Respondents' traits levels are estimated by the MAP method
  - Computationally efficient and unaffected by the number of latent traits

Item information function

- Direction of information is considered
- Information in direction of trait η_a for one binary outcome

$$\mathcal{I}_l^a\ \eta_a, \eta_b\ = \frac{\left[\beta_i - \beta_k \text{corr}\ \eta_a, \eta_b\ \right]^2 \left[\phi\ \alpha_l + \beta_i \eta_a - \beta_k \eta_b\ \right]^2}{P_l\ \eta_a, \eta_b\ \left[1 - P_l\ \eta_a, \eta_b\ \right]}$$

  - Smaller when traits are positively correlated



- One-dimensional case

$$\mathcal{I}_l\ \eta\ = \frac{\left[\beta_i - \beta_k\right]^2 \left[\phi\ \alpha_l + (\beta_i - \beta_k)\eta\ \right]^2}{P_l\ \eta\ \left[1 - P_l\ \eta\ \right]}$$

Test information and reliability

- Test information in direction of trait η_a $\mathcal{I}^a(\boldsymbol{\eta}) = \sum_l \mathcal{I}_l^a(\boldsymbol{\eta})$
- When using posterior latent trait estimator

$$\mathcal{I}_P^a(\boldsymbol{\eta}) = \mathcal{I}^a(\boldsymbol{\eta}) - \frac{\partial^2 \ln \phi(\boldsymbol{\eta})}{\partial^2 \eta_a} = \mathcal{I}^a(\boldsymbol{\eta}) + \varpi_a^a$$

- Standard error for trait η_a $SE(\hat{\boldsymbol{\eta}}_a) = \frac{1}{\sqrt{\mathcal{I}_P^a(\boldsymbol{\eta})}}$
- Empirical reliability (for estimated scores in a sample)

$$\bar{\sigma}^2_{error}(\hat{\boldsymbol{\eta}}) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{\mathcal{I}_P^a(\hat{\boldsymbol{\eta}}_j)} \qquad \rho = \frac{\sigma_P^2 - \bar{\sigma}^2_{error}}{\sigma_P^2}$$

Application – CCSQ

CCSQ instrument and sample

- Customer Contact Styles Questionnaire (CCSQ)
  - Measures 16 work-related traits, used in assessment for customer service roles (published by SHL)
  - Forced-choice format
    - **128 items** grouped into **32** **quads** (7 to 10 items per scale)
  - All items are also administered with a 5-point rating scale

- Sample
  - N=610
  - Paper & Pencil UK standardisation sample from 2001
  - 39% female
  - Half of the sample applicants, half job incumbents

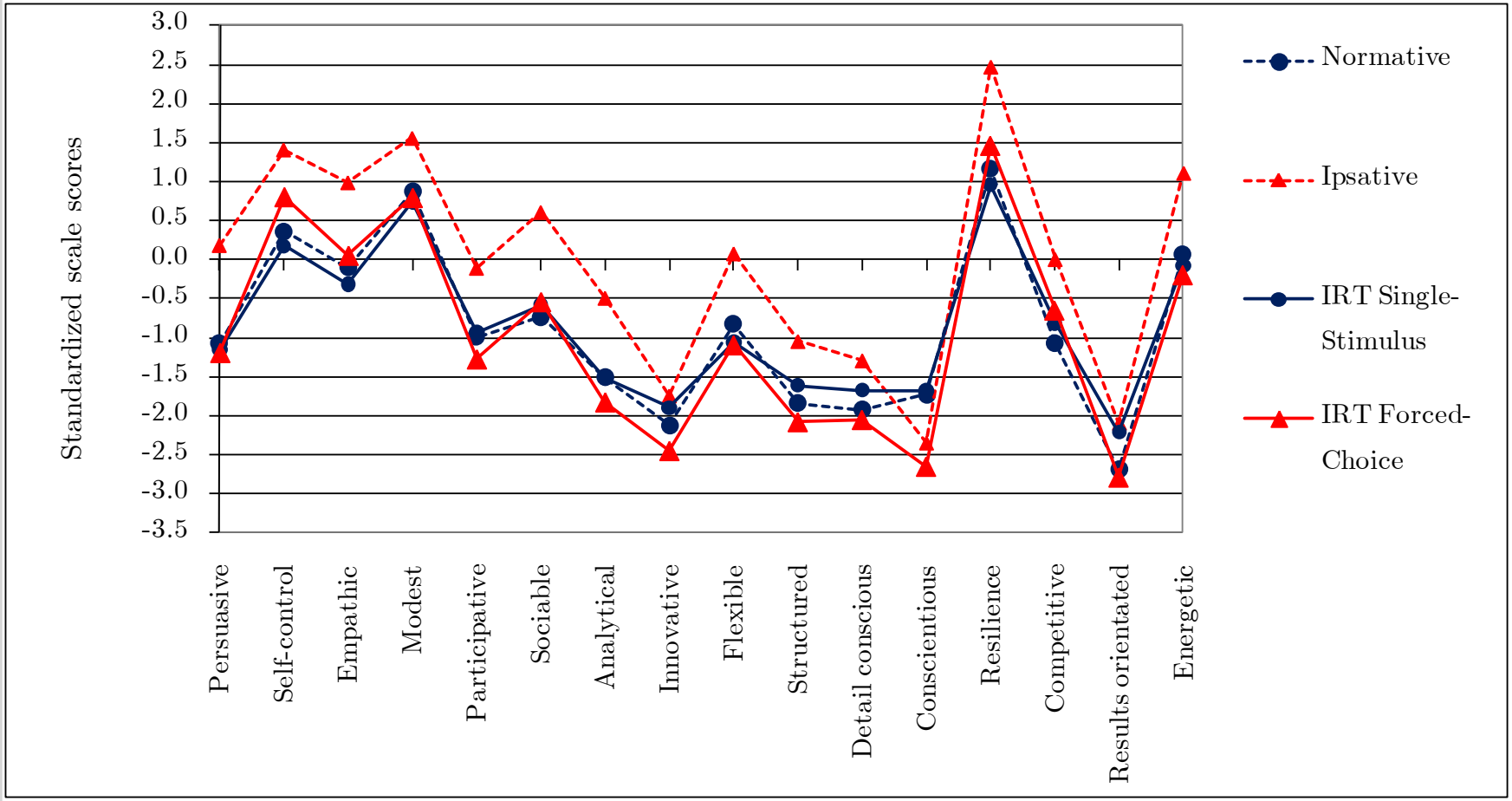
CCSQ model estimation

- Missing data problem arises

Partial ranking				Binary Outcomes					
A	B	C	D	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
	most	least		0	1	.	1	1	0

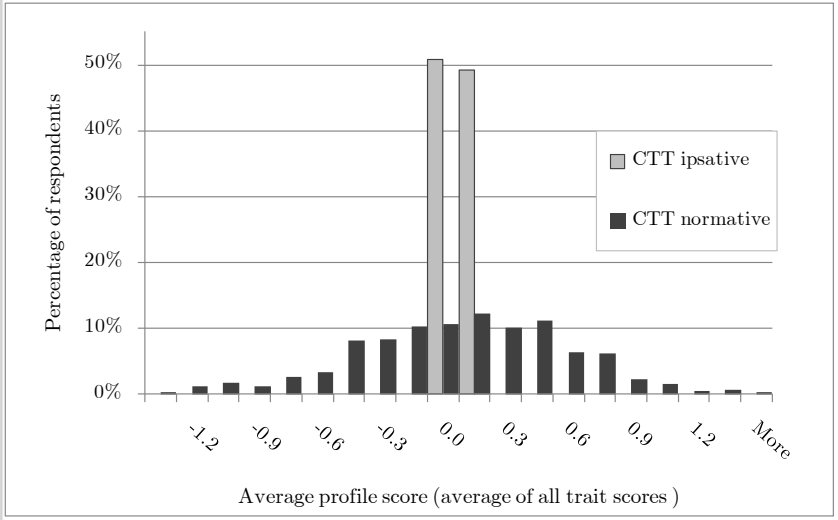
  - MAR but not MCAR
  - Limited information methods would produce distorted parameter estimates (Asparauhov & Muthen, 2010)
- Multiple imputations in Mplus are performed
- Model parameters are estimated on the 10 imputed datasets (ULS estimator)
- Person parameters are estimated on the original dataset (MAP estimator)

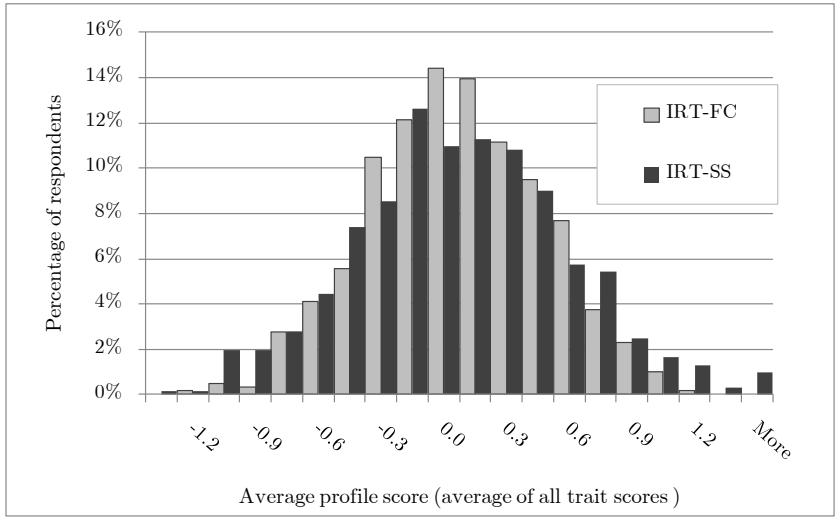
Example profile



No constraint on overall test score

- Can have all high/low scores
  - Scores are directly interpretable for comparison between individuals





Scores are no longer ipsative

- Absolute trait locations are recovered well
  - Mahalanobis distances between rating and ranking scores are much smaller with IRT method
- Construct validity undistorted
  - Average scale inter-correlation for ratings $r = \mathbf{0.21}$
  - For rankings scored with IRT $r = \mathbf{0.12}$
  - And for classically scored rankings (ipsative) $r = \mathbf{-.07}$
  - Identical factors extracted from rankings and ratings
- Reliabilities can now be estimated as well as SE for each individual set of scores

Simulation studies

Optimal forced-choice designs

Simulation designs

- Conditions crossed (1000 replications with 1000 cases):

Number of traits	2	5
Correlations between traits	*0, +0.5, -0.5*	*as in FFM*
Number of items	*12 and 24 items per trait*	*12 items per trait*
Block sizes	*pairs*	*pairs, triplets, quads*
Keyed direction of items	*positively worded / positively and negatively worded*	*positively worded / positively and negatively worded*

- For all reasonable designs
  - Good parameter recovery, including correlations between traits
  - Good latent trait recovery
  - Empirical chi-square rejection rates for more complex models are too high
  - Empirical reliabilities are sufficiently close to actual reliabilities

Forced-choice design rules

- Given sufficient number good quality items, the following are important factors:
- **Positively and negatively keyed items**

  With approximately the same number of binary outcomes comes from comparing items keyed in the same and opposite directions, the trait recovery can be good with any number of traits, and any trait correlations

- **Number of traits assessed**

  When the number of traits is large, and traits are not strongly positively correlated overall, any forced-choice designs will reliably locate trait scores

- **Correlations between traits**

  Comparing items keyed in the same direction is more effective the lower correlations between the latent traits

- **Block size**

  Same items provide more information if combined in larger blocks

Conclusions

Ipsative problem solved

Thurstonian IRT model in perspective

- Other IRT models have been suggested for creating new FC questionnaires
  - McCloy et al.(2005)
  - Stark, Chernyshenko & Drasgow (2005)
  - These models do not provide a solution for the existing forced-choice tests
- Thurstonian IRT model can be readily applied to any forced-choice data, with the objective of estimating
  - item parameters,
  - relationships between the latent traits,
  - and persons' parameters.
- It works with any existing tests using ranking format
  - Any number of items per block
  - Any number of traits
  - Multi- and one-dimensional comparisons

Growing area of research

- Benefits of the forced-choice format can be enjoyed without the disadvantages of ipsative data
  - Reducing halo effects in research
  - Cross-cultural research free of response sets
  - Exploration of factor structures without method factors
  - Etc.
- Embedding the model in an SEM framework allows further latent variable modeling
- There should be no more ipsative data – the problem of ipsative data has been effectively solved

Acknowledgements

- This work would not have been possible without
  - the extraordinary support of my advisor Alberto Maydeu-Olivares
  - moral support given to me by Simon

- Thanks to my former employer SHL Group and my research director Dave Bartram
- Thanks to the SMEP for the Dissertation support award

- Thanks to the Psychometric Society for the 2011 Dissertation Award
- Thank YOU FOR LISTENING!