Solving the problems of ipsative data:

The common framework for proper scaling of comparative response formats

Anna Brown
University of Kent

International Congress of Applied Psychology

29 June 2018

Absolute and comparative judgements

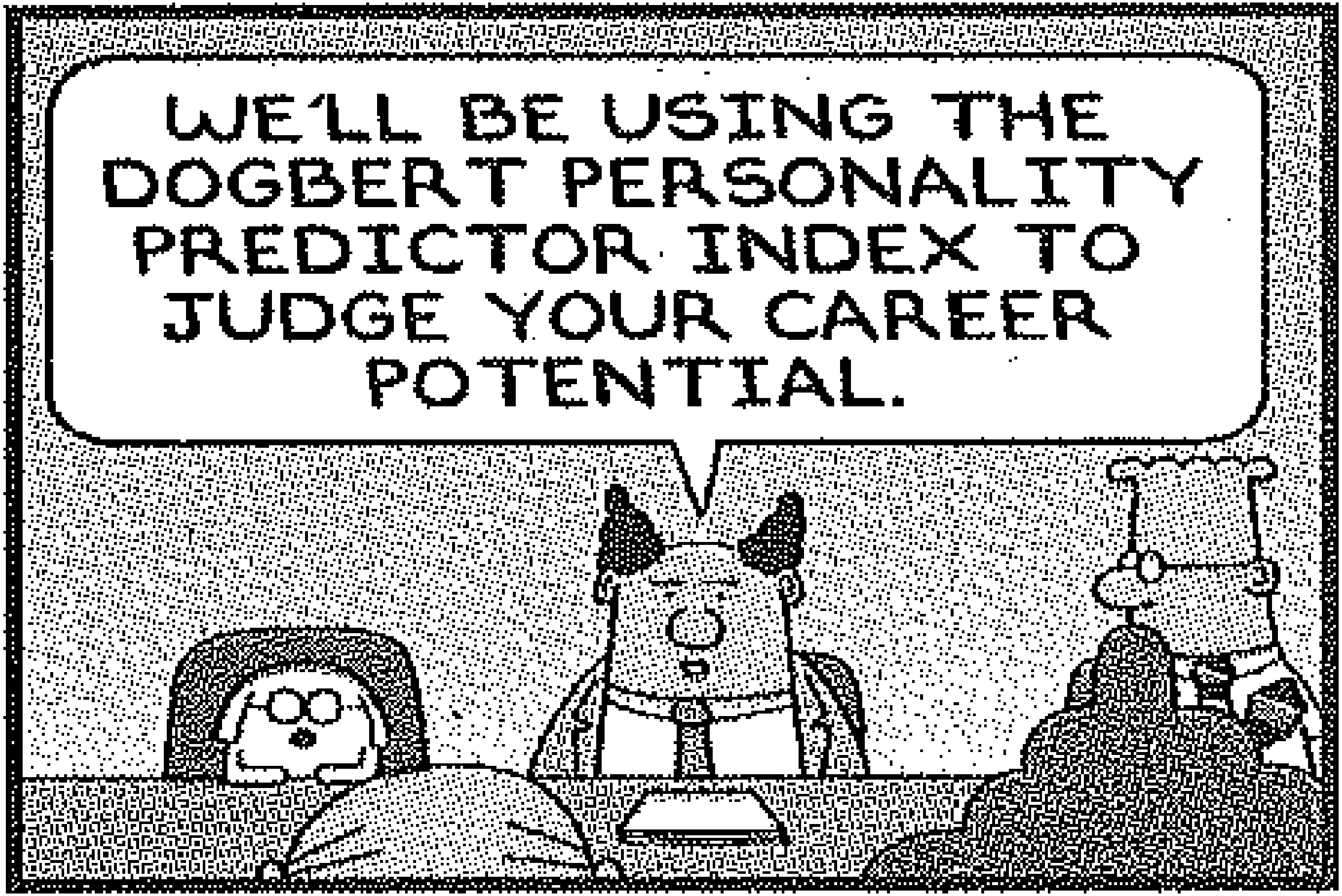
- Rephrasing Coombs (1960), the essential objective of psychological assessments is to associate with each person a **point in a psychological space**
  - Obtain person's absolute position on attributes of interest (personality traits, abilities, etc.)
  - By collecting responses of persons to relevant stimuli
    - "…*basically, all a person can do is to* ***compare*** *stimuli with each other, or against some* ***absolute*** *standard or personal reference point…*"
    - i.e. using either absolute or comparative judgements

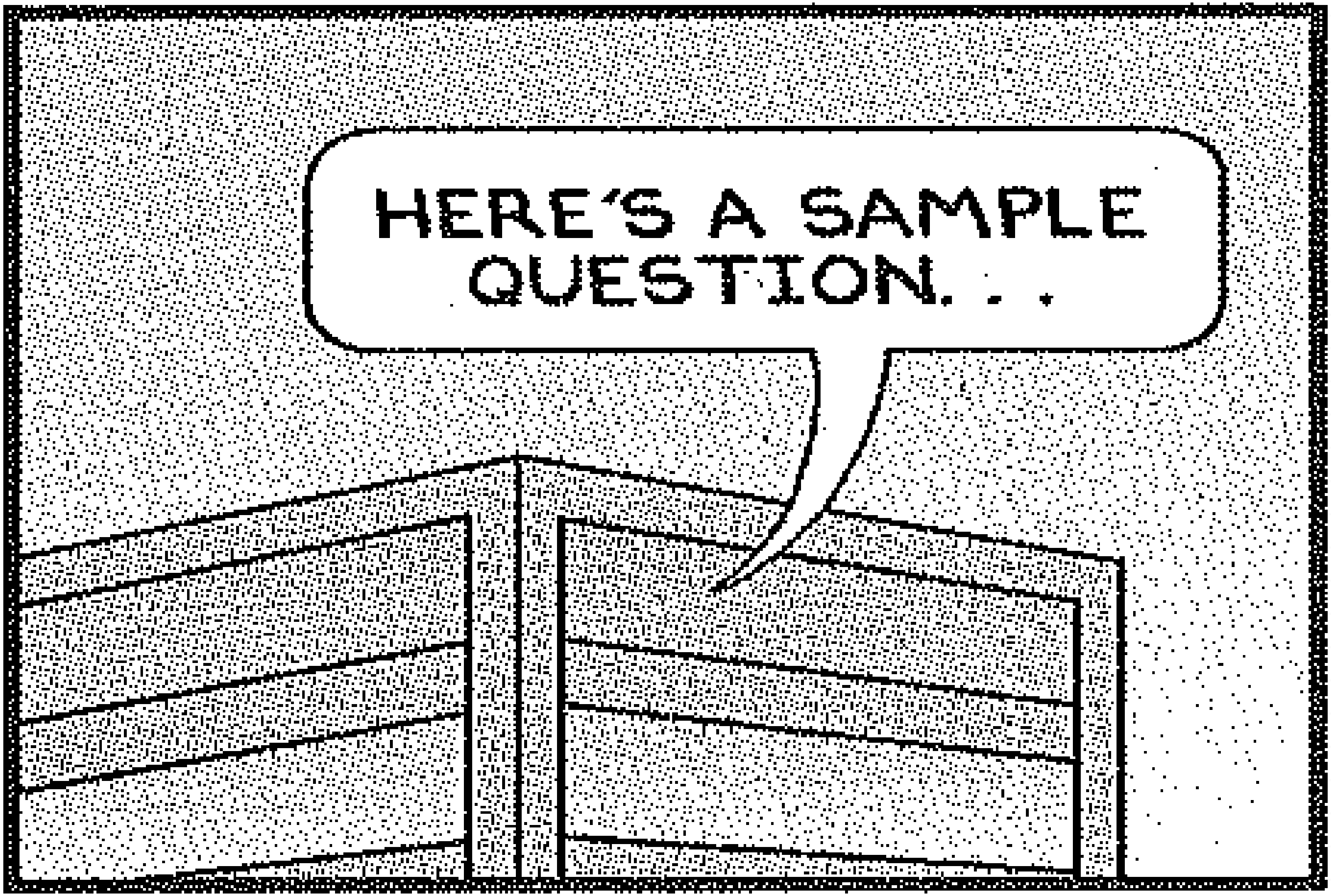
Single stimulus

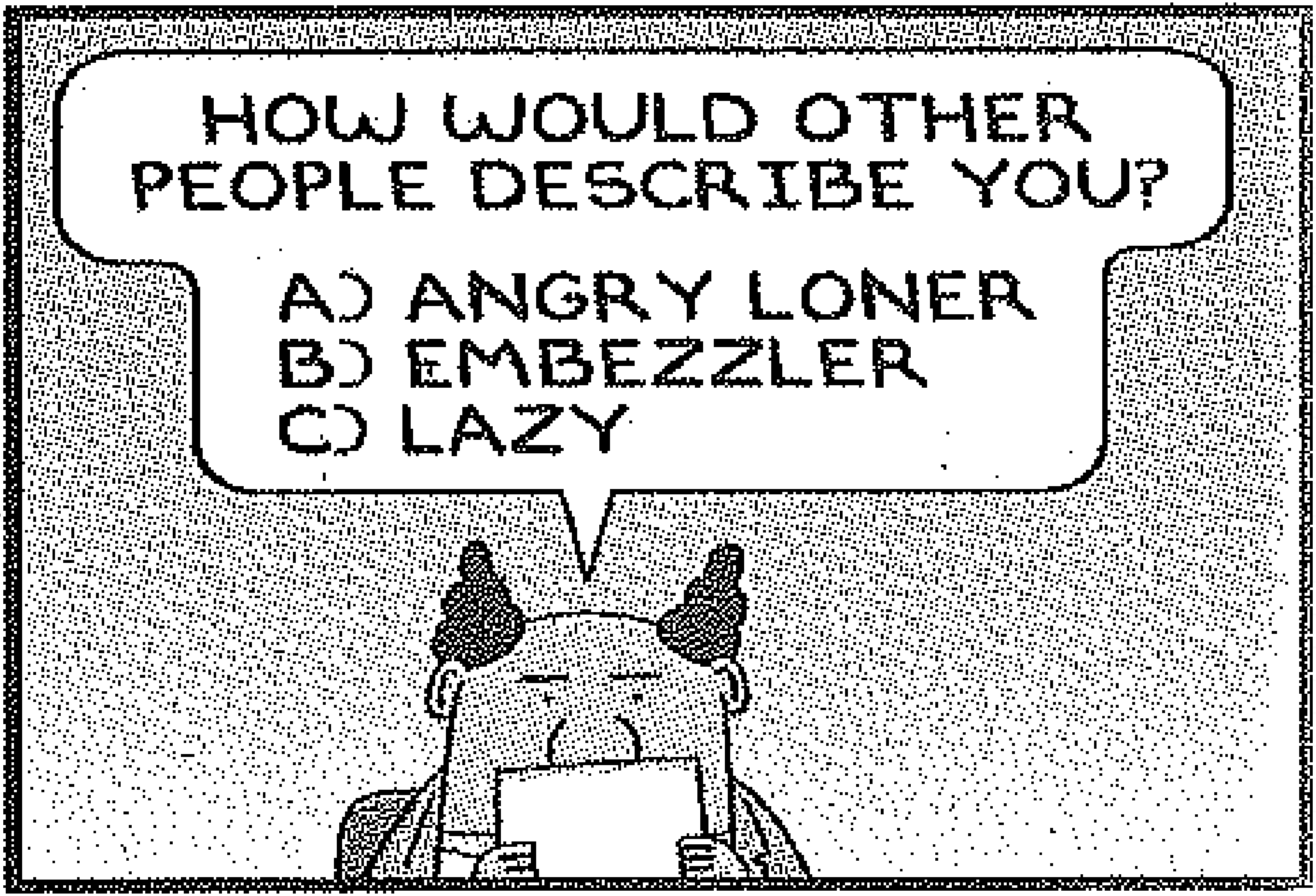
	Not at all like me	Somewhat like me	Quite like me	Completely like me
A. Dependable				X
B. Curious			X	

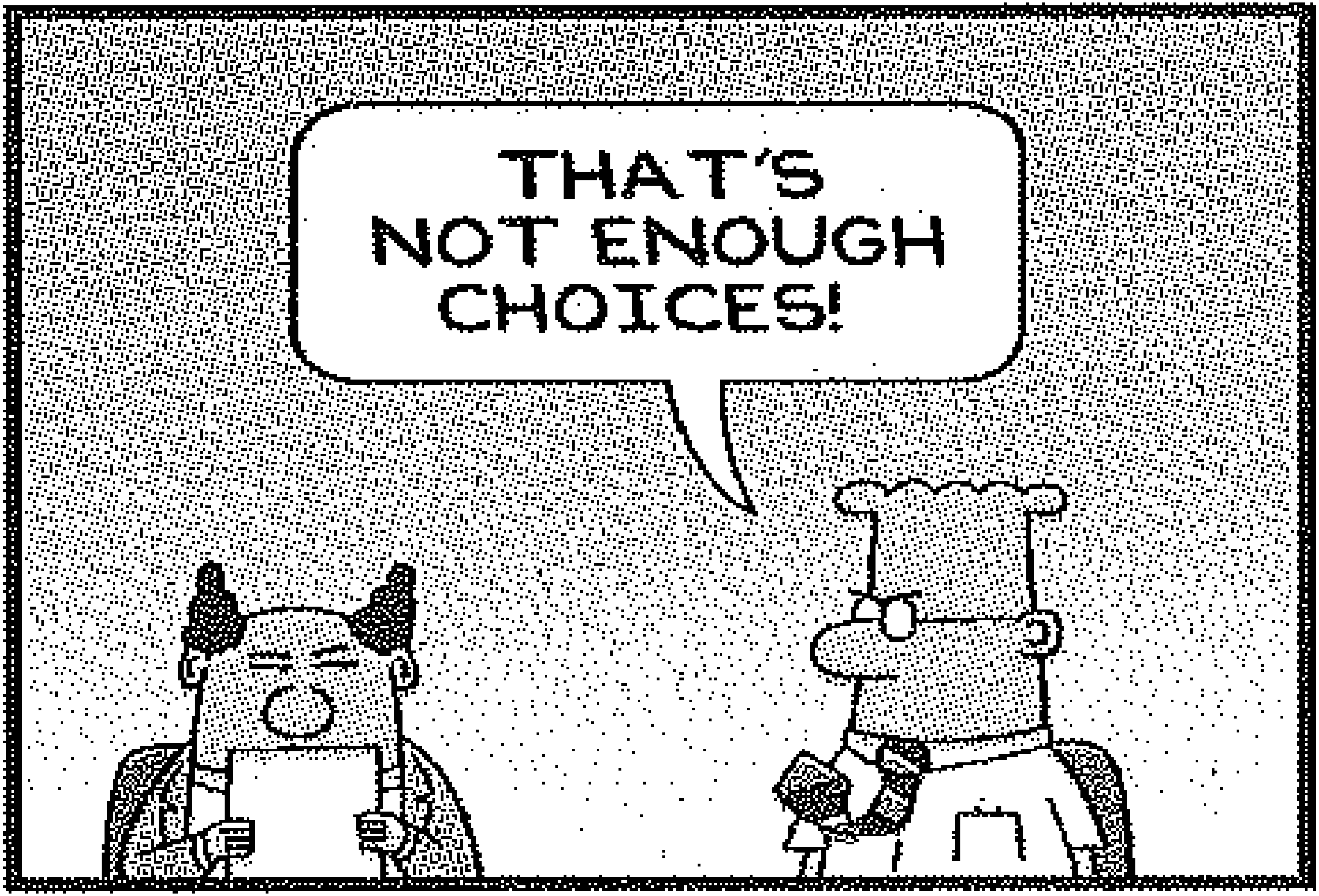
- Person is asked where he/she stands in relation to each stimulus (absolute judgements)

+ Easy to infer absolute positions on relevant **attributes**
- Open to response biases
  - Vulnerable to idiosyncratic uses of the rating options (response styles)
  - Easy to endorse all stimuli (acquiescence, leniency, “halo”)
  - Easy to endorse all desirable stimuli (socially desirable responding)


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

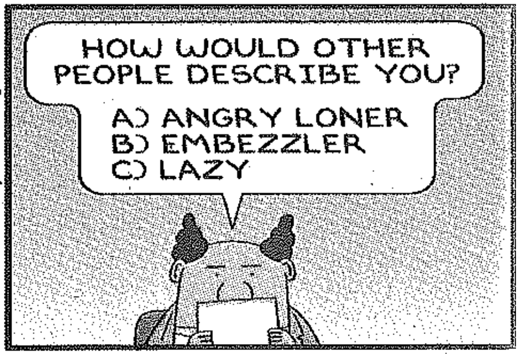

HERE'S A SAMPLE QUESTION. . .


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


THAT'S NOT ENOUGH CHOICES!


SAYS THE ANGRY LONER.

Comparative judgements



- Stimuli are compared with each other (comparative judgement)

- Prevents uniform response biases
  - Impossible to endorse all stimuli
  - Can reduce socially desirable responding*
  - Facilitates differentiation beyond absolute judgements

Classical scaling method

	Rank
A. Dependable	3
B. Curious	1
C. Modest	2
D. Calm	4

- Logic is the same in all comparative formats
  - In ranking tasks, assign points to attributes according to stimuli's inverse ranks
- The test score **is constant for everyone** → ipsative data <u>at attribute level</u>

Problems of ipsative data

- The total test score is the same for everyone

1. Attribute scores are relative to the person's mean
   - Interpersonally incomparable
2. Construct validity is distorted
   - Lose one degree of freedom
   - Factor analysis is not possible
   - Negative average scale inter-correlation

$$r_{ave} = \frac{-1}{k-1}$$

3. Criterion-related validity is distorted
   - Correlations with an external criterion sum to zero
4. Assumption of consistent coding violated
   - Alpha is not appropriate measure of reliability

New scaling of comparative data

- The classical scaling approach fails to provide absolute positions on traits
- New models for comparative formats are required
  - Models for choice behaviour in psychology have existed for a long time, and they are well known:
    - Thurstone’s law of comparative judgement (1927)
    - Coombs’s (1950) unfolding preference model
    - Luce’s (1959) choice axioms
    - Tversky’s (1972) “elimination by aspect” theory
    - And others

Law of comparative judgement

- Thurstone (1927)
  - Each item elicits a utility – psychological value, or "*affect that the object calls forth*"
  - Item with the higher utility at the moment of comparison is preferred (utility maximization rule, or UMR)

- In a preferential **choice task**, item 1 is preferred if

$$\text{utilty}_1 \geq \text{utilty}_2,$$

  otherwise item 2 is preferred
- In a **ranking task**, the utilities of items ranked 1, 2 ,…, *n* must be ordered so that

$$\text{utilty}_1 \geq \text{utilty}_2 \geq \ldots \geq \text{utility}_n$$

Response model for choice

	More like me	More like me	
A. Dependable	X		B. Curious

- Person is asked which of two stimuli he/she prefers
- Outcome of comparison $\{i, k\}$ is a binary variable

$$y_{\{i,k\}} = \begin{cases} 1, & \text{if } i \text{ is preferred} \\ 0, & \text{if } k \text{ is preferred} \end{cases}$$

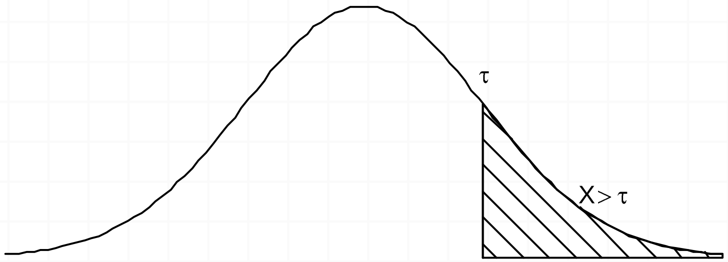
- According to the UMR, the outcome is determined by the relative values of utilities (denoted t)

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

Utilities and binary outcomes

- Consider the difference of utilities $\boxed{y^*_{\{i,k\}} = t_i - t_k}$
- Then the outcome of preferential choice {*i*, *k*}

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



- Threshold process
  - **Unobserved** utility difference y^* is the response tendency for **observed** binary outcome y
  - Assuming utility differences normally distributed, this is an IRT model with the link function = normal ogive

Response model for ranking

		Rank
A.	Dependable	3
B.	Curious	1
C.	Modest	2
D.	Calm	4

- Person is asked to rank several stimuli in order of preference
- Ranking of $\boldsymbol{n}$ stimuli involves $\boldsymbol{n(n-1)/2}$ pairwise preferential choices (binary dummy variables)
- Partial ranking or ranking with ties
  - only top preference; only top and bottom preference; Q-sorts
  - some pairwise outcomes are missing

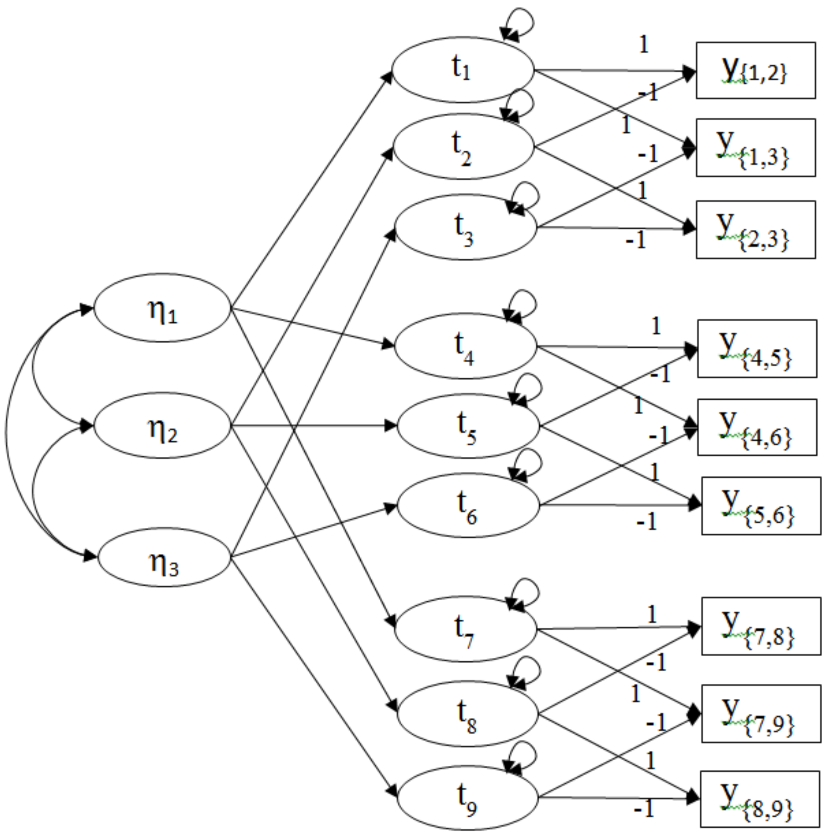
Measurement model for utilities

- Preferential choices (*observed* variables) are determined by utility judgements (*latent* variables)
- Utility judgements depend on underlying psychological attributes that we want to measure
  - Measurement model is needed to link utilities to the attributes (latent factors)
    - Linear Factor Analysis models (LFA)
    - Ideal Point models (IP)
  - We use LFA models; for example, factorially pure utility

$$t_i = \mu_i + \lambda_{ia}\eta_a + \varepsilon_i$$

Thurstonian factor model for ranking blocks

- Second-order factor model with binary outcomes
- Model estimation using
  - Tetrachoric correlations
  - ULS or DWLS
- Responses are transitive; no pairwise errors



Graded preference

	Much more like me	Slightly more like me	Slightly more like me	Much more like me	
A. Dependable		X			B. Curious

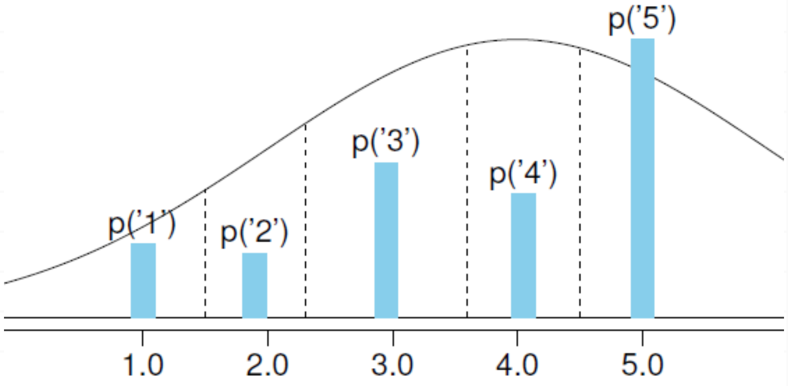
- Person is asked to what extent he/she prefers one or the other stimulus, using graded categories
- Outcome of comparison $\{i, k\}$ is an ordinal variable

$$y_{\{i,k\}} = \begin{cases} 4, & \text{if item } i \text{ is preferred "much more"} \\ 3, & \text{if item } i \text{ is preferred "slightly more"} \\ 2, & \text{if item } k \text{ is preferred "slightly more"} \\ 1, & \text{if item } k \text{ is preferred "much more"} \end{cases}$$

Utilities and ordinal outcomes

- UMR applied to graded preference decisions

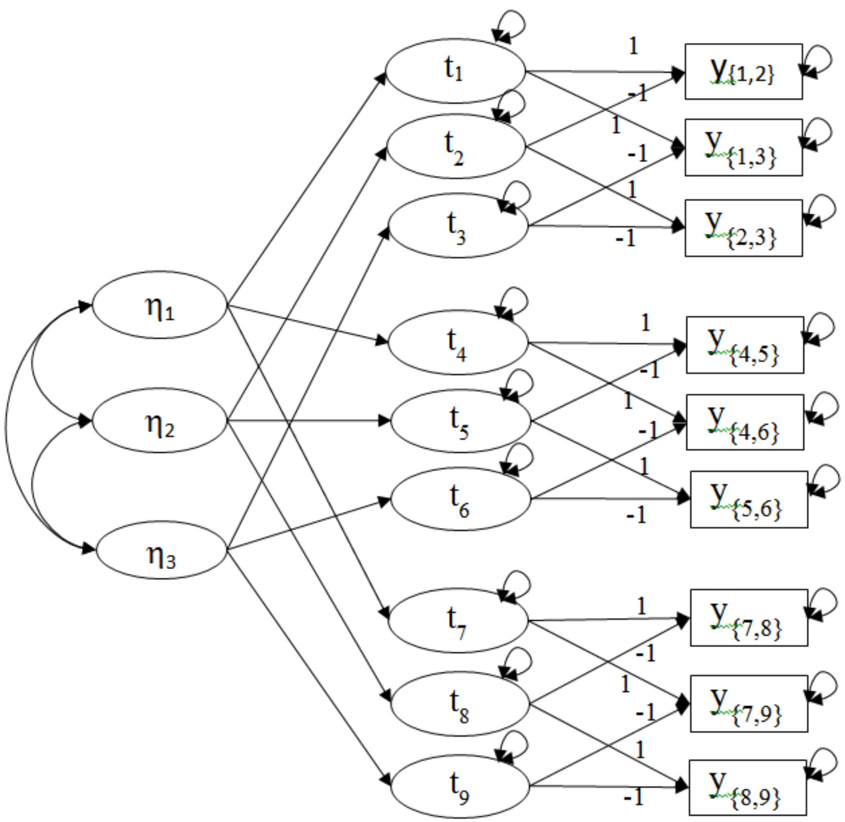
$$y_{j\{i,k\}} = \begin{cases} C, & \text{if} \quad y^*_{j\{i,k\}} \geq \tau_{\{i,k\}C-1} \\ C-1, & \text{if} \quad \tau_{\{i,k\}C-2} \leq y^*_{j\{i,k\}} < \tau_{\{i,k\}C-1} \\ \quad \dots & \\ 2, & \text{if} \quad \tau_{\{i,k\}1} \leq y^*_{j\{i,k\}} < \tau_{\{i,k\}2} \\ 1, & \text{if} \quad y^*_{j\{i,k\}} < \tau_{\{i,k\}1} \end{cases}$$



- Ordinal outcomes are **categorised** response tendencies
  - Threshold process (C–1 ordered thresholds)
  - IRT model with link function = normal ogive

Thurstonian factor model for graded blocks

- Second-order factor model with ordinal outcomes
- Model estimation using
  - Polychoric correlations
  - ULS or DWLS
- Responses may be **in**transitive; pairwise errors



Proportion-of-total preference ("composition")

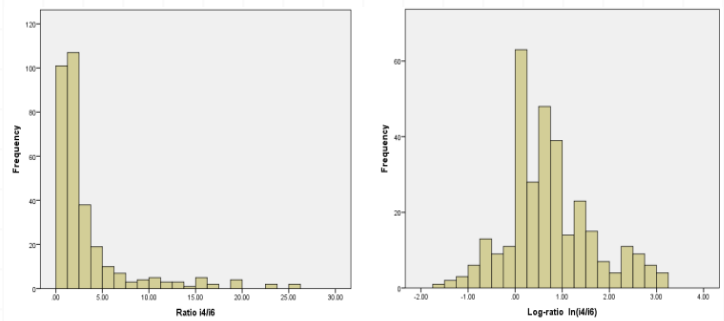
	% like me	% like me	
A. Dependable	60	40	B. Curious

- Person is asked to express preference for one or the other stimulus as proportion of total
- Points given to the stimuli y_i and y_k are ratio variables
  - Ratio of points y_i/y_k is the obvious outcome variable
  - Preserves the **ratio** of psychological values felt for the stimuli

$$\frac{y_i}{y_k} = \frac{C v_i \Big/ \sum_{q=1}^{n} v_q}{C v_k \Big/ \sum_{q=1}^{n} v_q} = \frac{v_i}{v_k}$$

Utilities and ratio outcomes

- Ratio of observed points is log-normal



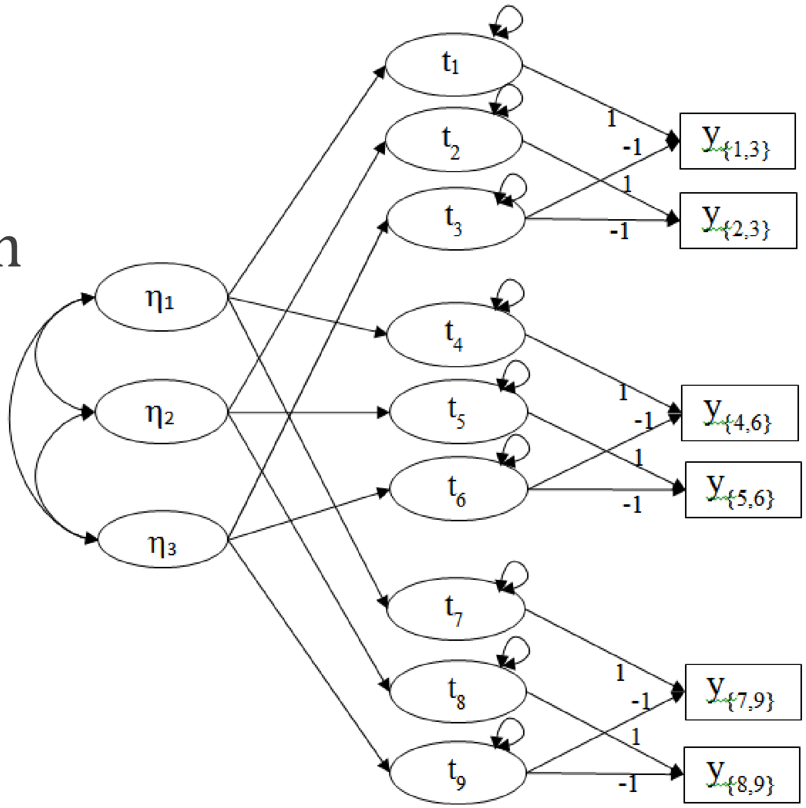
- Log-transformed ratios of points is the observed outcome (normal)

$$y_{\{i,k\}} = \ln\left(\frac{y_i}{y_k}\right) = \ln\left(\frac{v_i}{v_k}\right) = \ln(v_i) - \ln(v_i) = t_i - t_k$$

  - The utility difference is actually **observed**
  - Interval level of measurement, linear model

Thurstonian factor model for compositional blocks

- Compositional blocks of size=***n*** are described by ***n*–1** contrasts with a referent item
- Second-order factor model with continuous outcomes
- Model estimation using
  - Pearson's correlations
  - Maximum Likelihood



Estimating persons' positions on attributes

- Once model parameters are known, factor scores may be estimated
  - For binary and ordinal outcomes (IRT models), a combination of scores is found, for which the observed response pattern is most likely
    - Maximises the mode of the posterior likelihood (MAP)
    - In **ranking** blocks, pairwise error is 0 and factor scores cannot be estimated. Second-order factor model is parameterised as 1st-order model (TIRT model)
  - For continuous outcomes, regression with correlated factors is used
- <u>Person scores on attributes are no longer ipsative!</u>

But how can relative information ever become absolute???

- Positions of **utilities are** relative

$$(t_i + c) - (t_k + c) = t_i - t_k$$

- But we wanted absolute positions of **attributes** (second-order factors), not utilities!
  - From unidimensional comparison, the scale of attribute can be identified, **unless** the factor loadings are equal

$$\overline{t}_i - \overline{t}_k = \mu_i - \mu_k + (\lambda_i - \lambda_k)\eta$$

  - From multidimensional comparisons, the scales of attributes can be identified, **unless** the factor loadings are linear combinations of each other

$$\overline{t}_i - \overline{t}_k = \mu_i - \mu_k + \lambda_{ia}\eta_a - \lambda_{kb}\eta_b$$
$$\overline{t}_q - \overline{t}_r = \mu_q - \mu_r + \lambda_{qa}\eta_a - \lambda_{rb}\eta_b$$

The new rules of comparative measurement

- Recommendations for good comparative questionnaire designs are available
  - How to maximize information (Brown & Maydeu-Olivares, 2011)
  - How to ensure identification of attribute scales? (Brown, 2016)
  - What kind of items? (Brown & Maydeu-Olivares, 2010)
- Other considerations
  - Cognitive complexity increases as the block size increases
  - For good control of social desirability, careful matching of items within blocks is needed

Summary: Comparative data

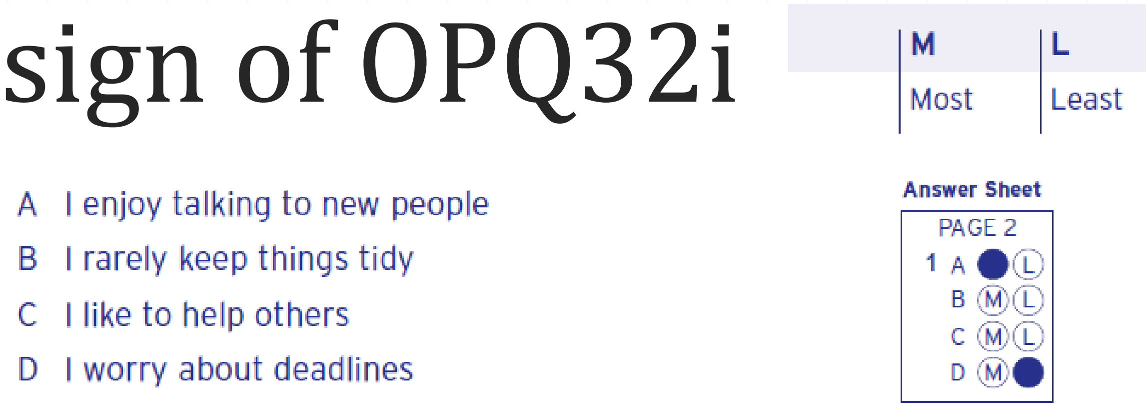
	Level of measurement		
Block size	Binary	Ordinal	Ratio
n = 2	Choice between 2 alternatives	Graded comparison between 2 alternatives	Composition with 2 alternatives
n ≥ 3	Ranking (full or partial); Ranking with ties (Q-sort)	Graded block (paired graded comparisons)	Composition with 3 or more alternatives

Summary: Analysis of comparative questionnaire data

- We adopt the outcome of pairwise comparison as the universal data unit
- We assume that utility maximisation is the basis for outcome of **any** comparison (binary, ordinal, ratio)
- We adopt utility difference as the universal latent variable underpinning the pairwise outcome
  - The normal utility difference $y^*_{\{i,k\}}$ is the response tendency for $y_{\{i,k\}}$
    - In preferential choice, the latent tendency is dichotomised
    - In graded preference, it is categorised
    - In composition, it is directly observed

Applications

Redesign of OPQ32i



- OPQ32i was used in assessment for managerial and professional roles worldwide
- Measured 32 work-related traits with 416 items arranged in 104 partial ranking quads
  - Yielded ipsative scores
- Thurstonian IRT model was applied to create OPQ32r (Brown & Bartram, 2009)
  - Changed format from quads to triplets – cognitive simplicity
  - Took out least informative items based on the item parameters

Development of a Big 5 measure (FCFFM)

- Using the TIRT model as the basis, Brown and Maydeu-Olivares (2011) developed a FC questionnaire from scratch
  - Used 60 IPIP items, the five factor markers subset by Goldberg (1992)
  - 12 items per factor; 8 positively and 4 negatively keyed
  - Triplets (n = 3); equal number of pairs with items keyed in the same direction and items keyed in opposite directions
- TIRT was fitted to a sample of N=438 (RMSEA=.025)
  - Very similar inter-scale correlations to the Likert model (but slightly less inflated intercorrelations)
  - Mono-trait hetero-method correlations were very similar to reliabilities
- Later modified to be used as Compositional (Brown, 2016), and Graded Blocks (Brown & Maydeu-Olivares, 2017)

Re-analysis of old data that were not thought comparative

- Picture story exercise (PSE) consists of drawn pictures showing people
  - Respondents write stories describing what is happening in each of the pictures
  - PSE is supposed to measure implicit motives
  - Each story is scored based on how much each motive was mentioned
- PSE has been shown to have **good** external validity but very **poor** reliability (“**reliability paradox**”)
- Lang (2014) considered stories as expressions of competing motives
  - Implicitly comparative data, only outcomes of comparisons (between competing motives) gets observed
  - “Dynamic Thurstonian model” – utility maximisation, plus the principle of diminishing strength of motive after it gets expressed
  - Showed that the PSE was reliable all along, but used the wrong model (Cronbach’s alpha inappropriate)

Lang, J.B. (2014). **A Dynamic Thurstonian Item Response Theory of Motive Expression in the Picture Story Exercise: Solving the Internal Consistency Paradox of the PSE**. *Psychological Review*, 121(3), 481–500

Growing area of research

- For binary choice data, other models exist:
  - E.g. Zinnes-Griggs (1974); Andrich (1989, 1995); MUPP (Stark, Chernyshenko & Drasgow, 2005)
  - These models assume different measurement models for utilities and can be classified using a common framework (Brown, 2016a)
- For graded preferences and compositional data, I am not aware of any alternatives to the Thurstonian models

Future directions

- Computerized Adaptive Testing (CAT)
  - Already works with MUPP (“TAPAS”, Stephen Stark and colleagues)
  - We are working on CAT with TIRT (my PhD student Yin Lin)
    - Lin & Brown (2017) looked into the influence of context (which block the item is in) on item parameters
- Latent classes rather than latent factors underlying preferences
  - For example, the use of FC formats for assessments of personality types
- To what extent can the comparative formats prevent faking? A much more in-depth research is needed

Thank you

Anna Brown, PhD

Senior Lecturer in Psychological Methods and Statistics
School of Psychology, University of Kent, Canterbury, CT2 7NR, UK

E-mail: a.a.brown@kent.ac.uk
Staff page: https://www.kent.ac.uk/psychology/people/browna/
Personal website: http://annabrown.name

References

- Brown, A. & Maydeu-Olivares, A. (November 22, 2017). **Ordinal Factor Analysis of Graded-Preference Questionnaire Data**. *Structural Equation Modeling: A Multidisciplinary Journal, 25(4), 516-529.*
- Brown, A., Inceoglu, I., & Lin, Y. (2017). **Preventing Rater Biases in 360-Degree Feedback by Forcing Choice**. *Organizational Research Methods, 20(1),* 121-148.
- Lin, Y., & Brown, A. (April 28, 2016). **Influence of Context on Item Parameters in Forced-Choice Personality Assessments**. *Educational and Psychological Measurement*, 77(3), 389-414.
- Brown, A. (2016b). **Thurstonian Scaling of Compositional Questionnaire Data**. *Multivariate Behavioral Research,* 51:2-3, 345-356.
- Brown, A. (2016a). **Item Response Models for Forced-Choice Questionnaires: A Common Framework**. *Psychometrika, 81(1),* 135–160.
- Brown, A. & Maydeu-Olivares, A. (2013). **How IRT can solve problems of ipsative data in forced-choice questionnaires**. *Psychological Methods, 18(1), 36-52*.
- Brown, A. & Maydeu-Olivares, A. (2012). **Fitting a Thurstonian IRT model to forced-choice data using Mplus**. *Behavior Research Methods, 44,* 1135–1147.
- Brown, A. & Maydeu-Olivares, A. (2011). **Item response modeling of forced-choice questionnaires**. *Educational and Psychological Measurement, 71(3), 460-502.*
- Maydeu-Olivares, A. & Brown, A. (2010). **Item response modeling of paired comparison and ranking data**. *Multivariate Behavioural Research, 45, 935-974*.

References - continued

Book chapters

- Brown, A. & Maydeu-Olivares, A. (2018). **Modeling forced-choice response formats**. In Irwing, P., Booth, T. & Hughes, D. (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development,* pp. 523-570. London: John Wiley & Sons.
- Brown, A. (2015). **Personality Assessment, Forced-Choice**. In Wright, J. D. (Ed.), *International Encyclopedia of the Social and Behavioural Sciences, 2nd Edition.* Elsevier.

Psychometric tests and software

- Brown, A. (2012) ***Mplus syntax builder for testing forced-choice data with the Thurstonian IRT model. Software and User's Guide***. http://annabrown.name/software
- Brown, A. & Maydeu-Olivares, A. (2011). ***Forced-choice Five Factor markers***. Retrieved from PsycTESTS.
- Brown, A. & Bartram, D. (2009-2011). ***OPQ32r Technical Manual***. Surrey, UK. SHL Group.

Conference proceedings

- Brown, A. & Bartram, D. (2009). **Doing less but getting more: Improving forced-choice measures with IRT**. In: *Society for Industrial and Organizational Psychology conference*, 2-4 April 2009, New Orleans.