The UK's European university

University of Kent

Response distortions in self-reported and other-reported measures:
Is there light at the end of the tunnel?

Anna Brown

Response distortions: Summary of research

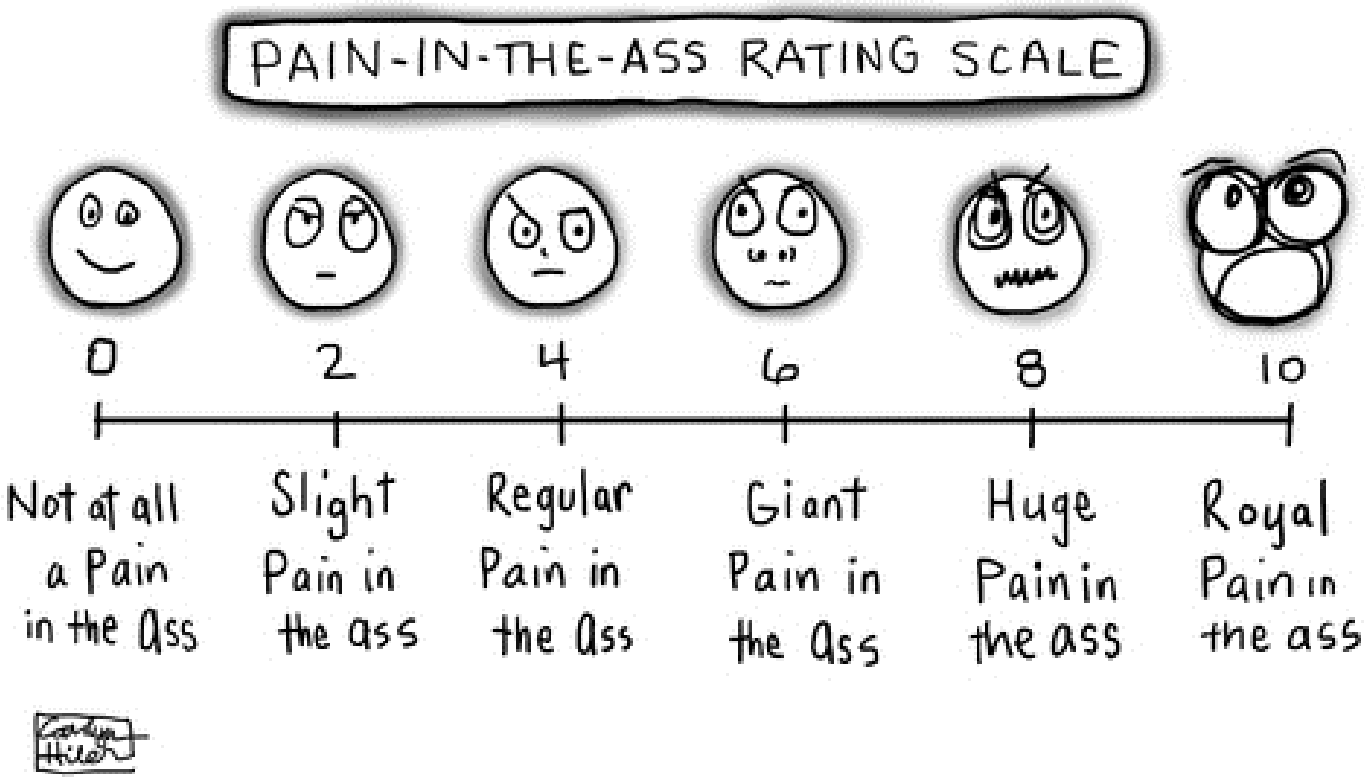
Acknowledgements

Eunike Wetzel (University of Konstanz)
Jan Böhnke (University of York)

Respondent-reported measures

- We ask people to describe themselves or others on a set of psychological characteristics
  - It may be the easiest and cheapest option out of imperfect alternatives
    - What perfect options are there to measure personality?
  - It may be the only available option
    - What other options are there to measure social attitudes?

We base our scaling on….



University of Kent

What is response bias?

- The "**systematic** tendency to respond …. on some basis other than the specific item content" (Paulhus, 1991)
  - Nuisance to measurement of intended constructs

- For example,
  - tendency to use extreme response categories,
  - tendency to agree with statements as presented,
  - tendency to give positive appraisal to someone who you quite like as a person

Types of response biases in self-reports

Independent of item content

- Careless responding
  - Not paying attention to item content
- Response styles
  - Systematic tendencies to prefer certain response categories over others

Depends on item content

- Socially desirable responding
  - Tendency to provide responses in line with social norms
  - Unintentional: self deception
  - Intentional: faking; simulation / dissimulation

Types of response biases in reports by others

- The same biases occur as in self-ratings
  - Inattentiveness, response styles
  - Socially (politically) desirable representation of ratee
- In addition, rater biases
  - Leniency / severity
  - Halo effect
    - over-generalisation of all behaviours, cognitive bias of exaggerated coherence (Thorndike, 1920; Kahneman, 2011)

How prevalent are response biases?

- Inattentive responding is common in basic research and social surveys
  - 10-12% in Meade & Craig (2012)
- Response styles are common in all applications
  - Up to 20% misreport on reversed items (Swain et al., 2008)
  - There are cultural differences (e.g. van Herk et al., 2004)
- Socially desirable responding is common. For the *intentional* component,
  - 47% of US applicants admit to exaggerating positive attributes and 62% to deemphasising negative (König et al., 2011)
  - "Ideal-employee" factor has been consistently found in high stakes assessment (Schmit & Ryan, 1993; Klehe et al., 2012)
  - Having political goals is common for raters (Murphy et al, 2004)
- Rater biases are common
  - Leniency and halo effects are commonly found (Ng et al., 2011; Barr & Raju, 2003; Murphy et al., 1993)

Sources of variance in responses

- The basic measurement model assumes only **two** types of sources influence the response
  - True scores – psychological constructs we intend to measure
  - Random error

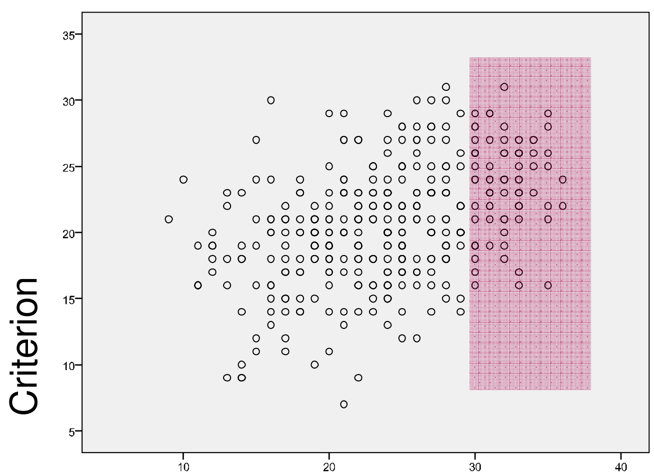
- A third source often exists – conscious and unconscious ***response distortions*** *(or* ***biases)***
  - Systematic error
  - If not included in the model, it will mask itself as true score

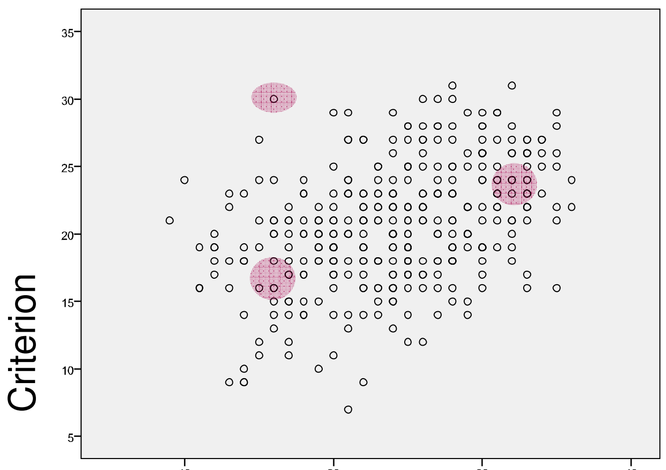
Why do I worry about response biases?

- Response biases are irrelevant sources of variance, and if left uncontrolled, they lead to **biased** test scores
  - Test no longer measures what we intended to measure (validity is affected)
- Decisions based on test scores that are **biased** in any way can lead to
  - breach of equal opportunities legislation
  - a sense of grievance
  - wrong selection decisions
  - invalid conclusions in basic research
- Fairness is the ultimate concern

“Valid” distortions?

- Some argue that biases do not matter if criterion-related validity is maintained
  - For example, high stakes assessments still predict performance (Ones et al., 2007)
    - employees continue “managing impression” after hire





“Valid” distortions?

- I argue that the key issue is construct validity
  - What does our test measure that predicts a criterion?
    - Faking is *“saying what you think you ought to say rather than what you really want to say. We have a word for that – “civilization.”* (Kevin Murphy, in Morgeson et al., 2007)
    - We may as well admit that when used in high stakes, the test measures what people think they are ought to say rather than their “personality”
    - We may compare who we select on the basis of this construct versus the “personality” basis

What should we do?

- I think that anyone who relies on respondent-reported measures used *in contexts where certain biases are prevalent*, should be concerned

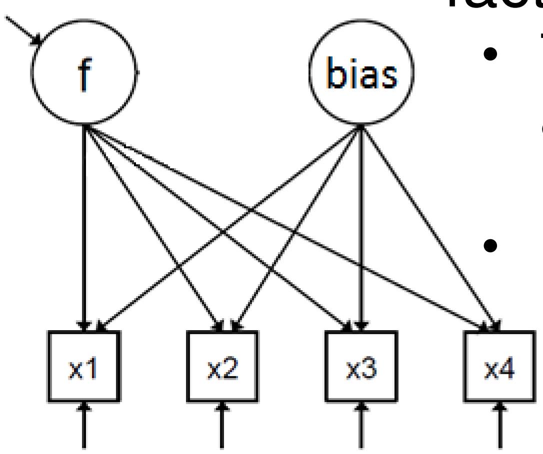
- To remedy the situation, one could
  - **Detect** biases after they have occurred, and adjust (**correct**) the test scores statistically
  - **Prevent** biases before they occur
  - Abandon respondent-reported measures and come up with something better

Detection and correction methods

- Manifest / Observed indices
  - Index quantifying the extent of certain bias is created
    - Frequency indices for response styles
    - Lie / Social Desirability scales
  - Observed test score is corrected using the index
    - E.g. the regression residual of trait score on the index is assumed free of bias (Webster, 1958)
- Latent variables
  - Response biases are part of the measurement model (via latent traits, or latent classes)
    - The extent to which bias affects the measurement model fit can be appraised
  - Latent (and estimated) trait scores are controlled for biases

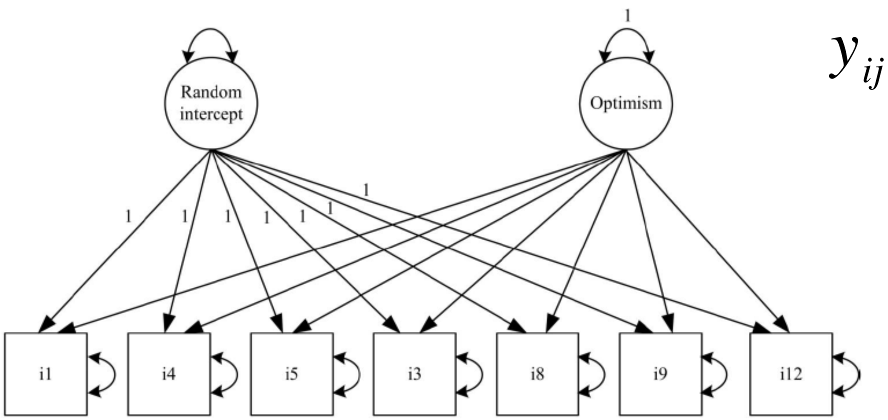
Bias as latent trait

- We may assume that individuals vary in the extent they engage in some biasing behaviour, and represent the individual differences as a latent factor
- Every response indicates not only its dedicated trait(s), but also some biasing factor
  - The approach has many uses and modifications and can be used for modelling many biases (e.g. Podsakoff et al., 2003)
  - Model identification can be a problem and often requires special designs
    - For example, having content-independent items (or “anchoring vignettes”) just to identify biases



Example 1: Acquiescence bias

- Acquiescence (or ‘yea’–saying) is the individual tendency to agree with items as presented
- Acquiescence bias becomes obvious when some people agree to both, positively and negatively worded items.
  - What should be opposite ends of the same factor, come out as two separate factors in EFA
- Personal tendency to acquiesce can be modelled as random intercept
  - Response for item i and person j



From: Maydeu-Olivares & Coffman (2006)

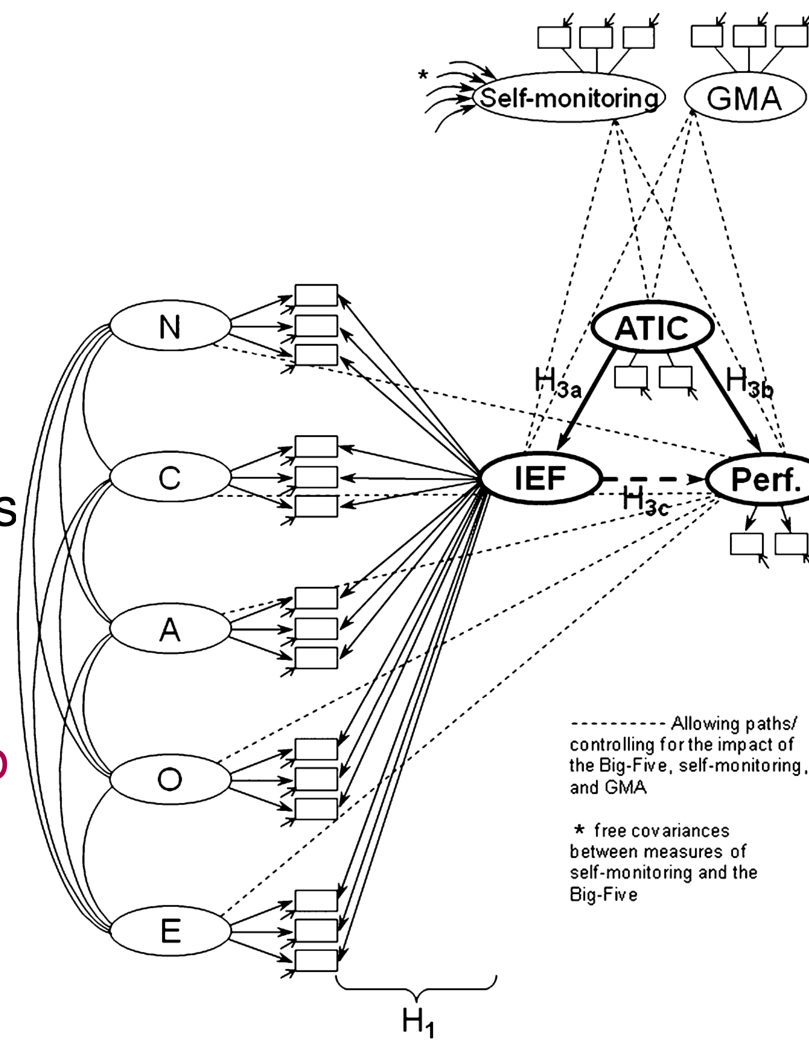
$$y_{ij} = \mu_i + \delta_j + \lambda_i f_j + \varepsilon_{ij}$$

- In school children data, RI accounts for about 10% of variance in item responses

Example 2: "Ideal-employee" factor

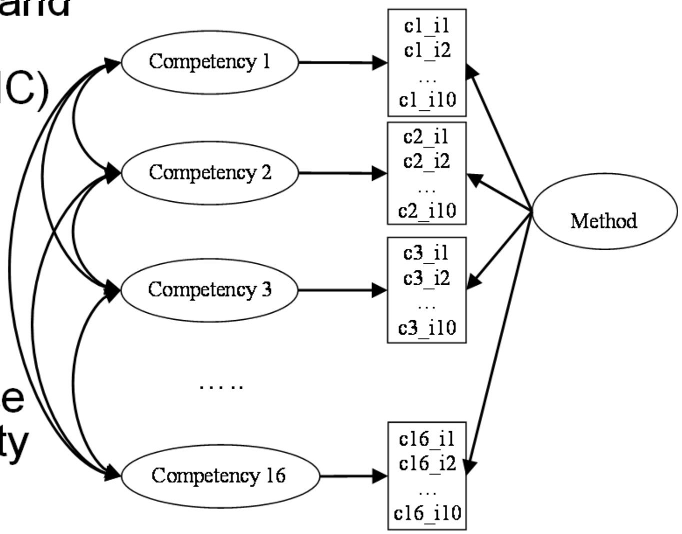
- A common factor explaining inflated correlations between all desirable characteristics has been found in applicant data (e.g. Schmit & Ryan, 1993)
- The "ideal-employee" factor has varying factor loadings – the most desirable behaviours affected most
- Klehe et al. (2012) showed that the relationship between ideal-employee factor and job performance is explained by ability to identify criteria (ATIC)

Illustration: Klehe et al. (2012)



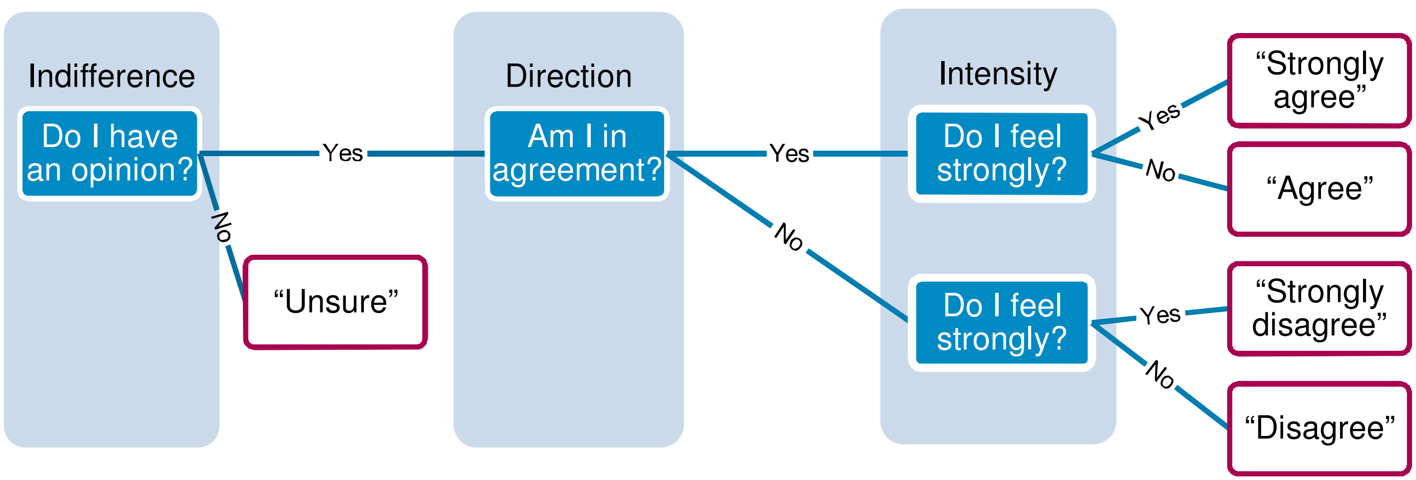
Example 3: Correcting biases in 360 appraisals

- Organizational appraisal data is notorious for response biases
- Study by Brown, Inceoglu and Yin (partly reported at SIOP 2014)
  - Large sample (N=4,675) of self-, peer, boss and subordinate assessments
  - Inventory of Management Competencies (IMC)
    - 16 competencies; 160 items
- Method factor represented non-uniform distortions similar to those of "ideal-employee" in both self- and other assessments
  - Explained around 50% of systematic variance
  - Controlling for method factor improved validity of competency scores
    - meaningful second-order factor structures
    - better inter-rater agreement (ave. ICC = 0.39)
    - better convergent correlations with an external measure (ave. self = .42; others = .25).



Bias as response process model

- Response process as a decision tree (Böckenholt, 2012)

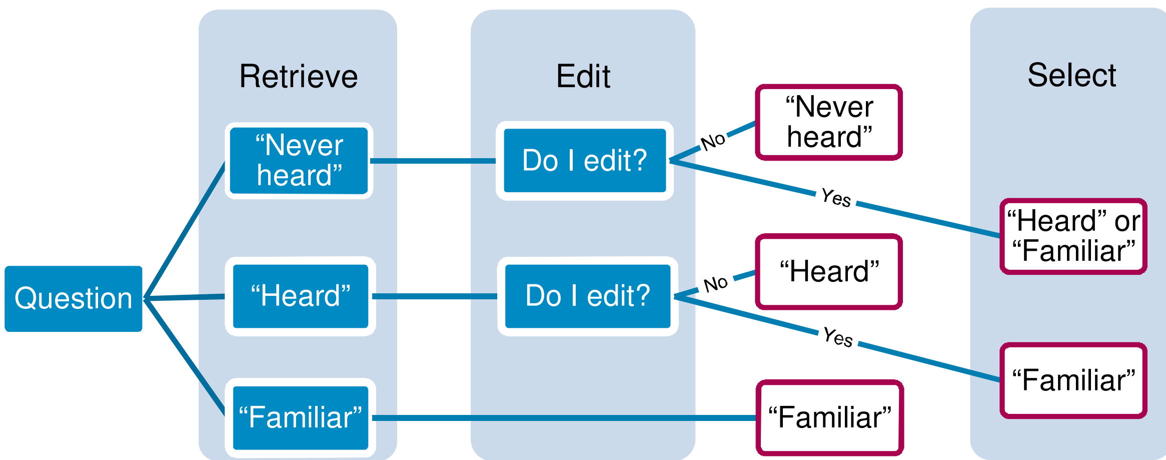


- 3 pseudo items are created to indicate a 3-step process

	Indifference	Direction	Intensity
Strongly disagree	0	0	1
Disagree	0	0	0
Unsure	1	-	-
Agree	0	1	0
Strongly agree	0	1	1

Example 4: Motivated misreports

- Bockenholt (2014) proposed the "Retrieve-Edit-Select" decision model to account for self-enhancement
  - Assumes that editing can happen only in one direction
    - For example, people over-report knowledge but do not under-report it

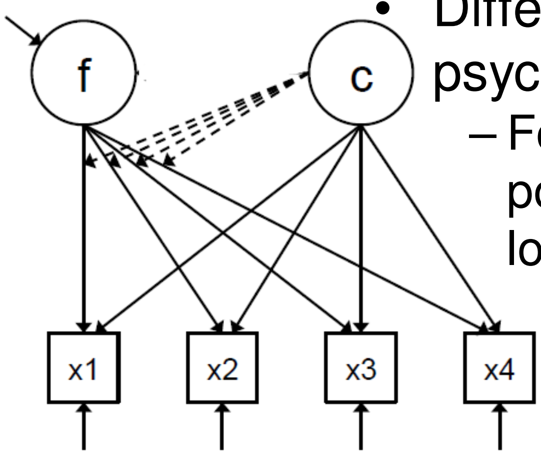


- Modelled latent traits θ^R, θ^E, θ^S

University of Kent

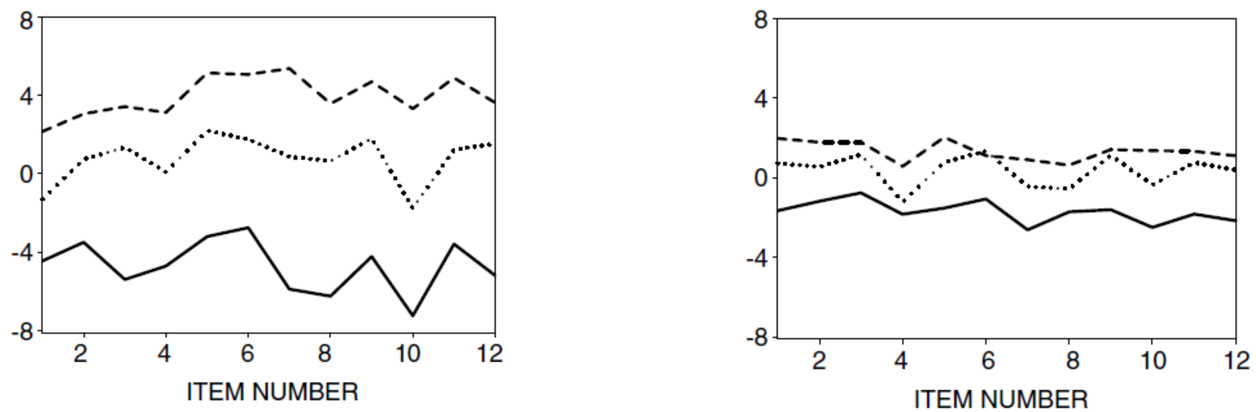
Bias as latent class

- We may assume that respondents come from several unobserved (latent) classes
  - Observed distributions are in fact mixtures of unobserved subpopulation distributions
- Model parameters may differ between classes
  - Differing thresholds (or intercepts) may indicate extreme responding
  - Differing factor loadings may indicate different psychological constructs underlying responses
    - For example, class of individuals endorsing both positive and negative items may show all positive factor loadings



Example 5: Extreme responding

- Rasch mixture modelling has been used to identify classes of extreme and mid-scale respondents
  - For instance, Austin et al. (2006) identified 2 classes with systematically different item thresholds controlling for the latent trait



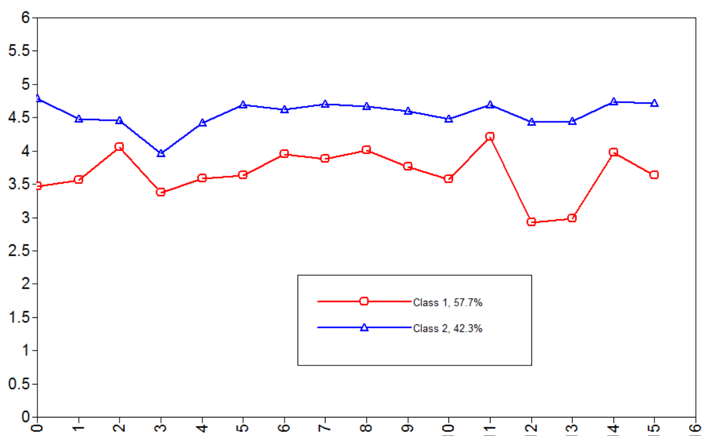
– Extreme responders (29%) have narrow thresholds (endorsing extreme categories is easier)

Example 6: Faking behaviour

- Re-analysis of Brown (2008) study: Instructed faking / Honest conditions
  - One job description was used as target; should yield the same **ideal profile**
  - Scale scores (item means) on 16 personality traits were analysed

Latent class analysis (LCA)

- 2 classes give excellent separation (entropy = .984)
  - "Ideal" and "honest" profiles



LCA with known class

- Do the latent classes coincide with the 2 conditions?
- Latent transition probabilities

	Class 1	Class 2
Honest	.082	.918
Instructed faking	.971	.029

- Unfortunately, LCA does not achieve such results in real operational data

University of Kent

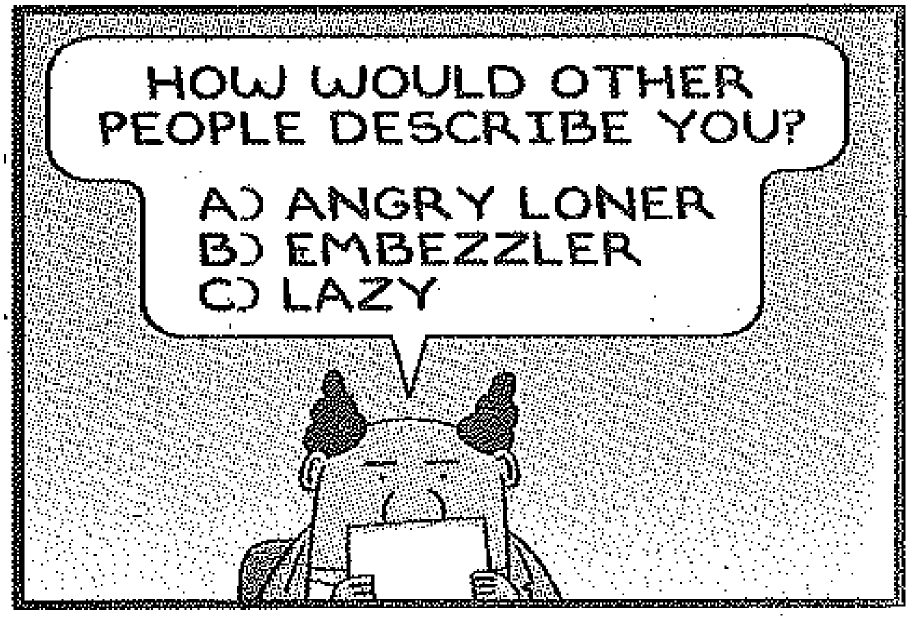
Limitations of correction methods

- No real-world data have only one type of bias
  - Modelling several biases is problematic
  - Special study designs are often required to separately identify biases
- Biases are complex to model properly
  - Every model is a great simplification of reality
    - For example, latent class models assume that there is no individual difference in the extent of bias within classes
  - Some biases are much more difficult to deal with than others
    - For example, faking is a challenge to model because the cognitions behind this process vary dramatically between people (Kuncel & Tellegen, 2009; Robie et al., 2007; Brown, 2014)

(some) Prevention methods

- Test-taking motivation
  - Lack of motivation increases careless responding
  - Motivation to meet the selection criteria increases socially desirable responding (Schmit & Ryan, 1993)
- Rater calibration
  - Calibrating own ratings against others reduce leniency
  - Rating the same competency for different people, rather than different competencies for the same person reduce halo (Kahneman, 2011)
- Item wording
  - Negatively worded items are difficult to process (“item verification difficulty”; e.g. Swain et al., 2008)
- Response format
  - Response options must be labelled thoughtfully to avoid idiosyncratic interpretation (e.g. Hernandez et al., 2006)
  - Forcing choice between items controls for all uniform biases (e.g. Cheung & Chan, 2002)

Forced choice



- Comparisons “calibrate” options against each other, reducing cognitive biases (Kahneman, 2012)
- Finer differentiation between similar stimuli
- Direct comparison - no rating scale and hence no idiosyncratic use of the rating options

Forced-choice: Mechanism for bias prevention

- According to Thurstone's (1927) **law of comparative judgement**, respondent chooses stimulus with the highest utility (t)
  - If $t_A - t_B > 0$, then item A is chosen
  - If $t_A - t_B < 0$, then item B is chosen
- If item utilities are biased with fixed linear effects *c* (arbitrary, *c* > 0) and *d*,

$$t'_{\mathrm{A}} = ct_{\mathrm{A}} + d, \qquad t'_{\mathrm{B}} = ct_{B} + d,$$

  - The difference of utilities has the same sign (Brown, 2010)

$$t'_{\mathrm{A}} - t'_{B} = \left(ct_{\mathrm{A}} + d\right) - \left(ct_{\mathrm{B}} + d\right) = c\left(t_{\mathrm{A}} - t_{\mathrm{B}}\right)$$

- FC format eliminates all multiplicative and additive effects acting uniformly within blocks

Example 7: Preventing biases in 360 appraisals

- Study by Brown, Inceoglu and Yin (continued from Example 3)
  - Large sample (N=4,675) of self-, peer, boss and subordinate assessments
  - Inventory of Management Competencies (IMC)
    - 16 competencies; 160 items
- Forced-choice rankings modelled with Thurstonian IRT (Brown & Maydeu-Olivares, 2011)
- Estimated trait scores yielded as good construct and external validities as the bias-corrected Likert ratings, and slightly better rater agreement (ave. ICC = 0.41).
  - This is impressive considering the lower reliability of FC scores
- The multidimensional forced-choice response format is an effective bias prevention method in self- and others- ratings

Limitations of prevention methods

- Some prevention methods have very small effects
- Prevention methods seem to be most effective against unmotivated biases
  - (which probably emerge due to us creating bad questionnaires in the first place)
- But when test developers go against human willpower, things get tough
  - Working with forced choice taught me that it is effective for prevention of response styles, leniency and halo
    - Recommended in cross-cultural research and assessments by others
  - But if someone wants to misrepresent their personality, they can do it, whether you are forcing choice or not
    - I can always swap my true choices to misrepresent myself

Is there light at the end of the tunnel?

Some thoughts on the effectiveness of the proposed methods and challenges ahead

So is there light at the end of the tunnel?

- Fighting biases can be very frustrating
- We can continue with developing detection and correction methods
  - Fast estimation methods and advancing psychometrics will help
- But in my opinion, we should focus on prevention
  - What is the point in investing all efforts in fancy models, and continue using poorly designed tests?
  - It is not enough to manipulate factors with small effects on biases
- It is time to think outside the box, and be critical of established practices

A question to you

- A question to those who use abstract rating options such as
  - Strongly disagree / disagree / neither agree nor disagree / agree / strongly agree
- If you do not want the responses to be affected by the tendency to agree, or the strength of agreement, why ask about agreement at all?
  - Additional factor is introduced
- Why not use response categories that represent intervals on the trait of interest?

In social conversation, how do you usually behave?

talkative – an easy talker – talk when necessary –

prefer listening – refrain from talking

(McDonald, 1999)

And another question to you

- A question to those who use personality measures for selection, and feel faking is normal because it reflects the adherence to social norms
- Why don’t you just ask the respondents:
  - What kind of person do you think we would like to recruit? (the “ideal-employee” image as they see it)

  AND
  - How motivated are you to get this job?
- Taken together, the ability to identify criteria (ATIC) and motivation presumably explain a lot of variance in job performance
  - And there is little reason to fake the above measures

A Plea for Process in Personality Prevarication

- "a focus on the response process that test takers go through will accelerate our understanding of faking behavior" (Kuncel, Goldberg & Kiger, 2011)
- This is true for all biases
- If we understand the process, we can
  - (At least) detect and correct it better
  - Prevent the negative impact of faking by creating better assessments

It is time for qualitative research

- I have been carrying out research of test taker cognitions in high stakes assessments
  - Qualitative interviews
  - Free descriptions of motivations and cognitions after taking a personality tests for selection
- It made me realise that
  - the prevalence of faking is high (and higher than estimated in the literature),
  - the motivation and cognitions are complex and different from person to person,
  - the problem is more serious than most admit,
  - the problem will only get worse with more exposure to psychological testing.

No simple answers

FC reduces faking

? When facing two equally desirable items, the respondent will fall back on true response (Gordon, 1951).

- *"I found this [FC] questionnaire more friendly because all statements were about good things, so I could relax and think about my personality"*

FC facilitates faking

? Direct comparison of items facilitate acute differentiation of their desirability levels (Feldman & Corah, 1960).

- *"…it was hard to chose which option was really me and tended to go with the one that my employer would be more likely to want."*

Conclusions

- Response biases matter because they can distort the true scores on attributes of interest
    - Construct validity is affected
- Detection / correction and prevention methods exist that can help, but there are many problems
- A more critical and fresh approach is needed
    - Investing time in creating a new type of assessment rather than in fixing problems in the old one
- Understanding the response process is crucial
    - in detecting and correcting the response biases,
    - in preventing response biases from occurring by creating assessments more resistant to them (or even free from them)

a.a.brown@kent.ac.uk

THANK YOU FOR LISTENING!

References

Wetzel, E., Boehnke, J., & Brown, A. (2016). **Response biases**. In Leong, F. et al. (Eds.), The ITC International Handbook of Testing and Assessment. Oxford University Press.

- Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. Personality and Individual Differences, 40, 1235-1245.
- Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. Organizational Research Methods, 6(1), 15-43.
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. Psychological Methods, 17(4), 665-678.
- Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. Psychometrika, 79(3), 515–537.
- Brown, A. (2008) The impact of questionnaire item format on ability to ‘fake good’. Paper presented at the International Test Commission conference, 13-16 July 2008, Liverpool.
- Brown, A. (2010). How Item Response Theory can solve problems of ipsative data. Doctoral dissertation, Universitat de Barcelona.
- Brown, A. (2014) “Faking good” on personality tests: Test takers’ cognitions and the forced-choice format. Paper presented at the International Test Commission conference, 2-5 July 2014, San Sebastian.
- Brown, A., Inceoglu, I., Lin, Y. & Bartram, D. (2014). Examining bias and validity of measurement in 360-degree feedback. Paper presented at the Society for Industrial and Organizational Psychology conference, 18-20 May 2014, Honolulu.
- Brown, A. & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. Psychological Methods, 18(1), 36-52.
- Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. Structural Equation Modeling, 9, 55-77.
- Feldman, M. J., & Corah, N. L. (1960). Social Desirability and the Forced Choice Method. Journal of Consulting Psychology, 24(6), 480-482.

References (cont.)

- Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. Journal of Applied Psychology, 35(6), 407.
- Kahneman, D. (2011). Thinking, fast and slow. London, UK: Allen Lane.
- Klehe, U. C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. Human Performance, 25(4), 273-302.
- Konig, C. J., Hafsteinsson, L. G., Jansen, A., & Stadelmann, H. (2011). Applicants self-presentational behaviour across cultures: Less self presentation in Switzerland and Iceland than in the US. *International Journal of Selection and Assessment*, *32*(3), 223-246.
- Kuncel, N. R., Goldberg, L. R., & Kiger, T. (2011). A plea for process in personality prevarication. Human Performance, 24(4), 373-378.
- Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. Personnel Psychology, 62(2), 201-228
- Maydeu-Olivares & Coffman (2006). Random intercept factor item analysis. Psychological Methods, 11, 344-362
- McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812–821.
- Meade, A. W., & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. Psychological Methods, 17(3), 437-455.
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. Personnel Psychology, 60(4), 1029-1049.
- Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error - A critical analysis. Journal of Applied Psychology, 78(2), 218-225.

References (cont.)

- Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. Journal of Applied Psychology, 89(1), 158-164.
- Ng, K. Y., Koh, C., Ang, S., Kennedy, J. C., & Chan, K. Y. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. Journal of Applied Psychology, 96(5), 1033-1044.
- Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. Personnel psychology, 60(4), 995-1027.
- Paulhus, D. L. (1991). Measurement and control of response bias. Measures of Personality and Social Psychological Attitudes, 1, 17-59.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. Journal of Applied Psychology, 88(5), 879.
- Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. Journal of Business and Psychology, 21(4), 489-509.
- Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. Journal of Applied Psychology, 78(6), 966.
- Scott D. Swain, Danny Weathers, and Ronald W. Niedrich (2008) Assessing Three Sources of Misresponse to Reversed Likert Items. Journal of Marketing Research: February 2008, Vol. 45, No. 1, 116-131.
- Thorndike, E. L. (1920). A constant error in psychological ratings. Journal of Applied Psychology, 4, 25-29.
- van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. Journal of Cross-Cultural Psychology, 35(3), 346-360.
- Webster, H. (1958). Correcting personality scales for response sets or suppression effects. Psychological Bulletin, 55(1), 62-64.