Advanced topics

Nils Myszkowski, PhD

Outline

Measurement invariance
Latent class models
Monotonicity
Unfolding models
Overfitting
Network psychometrics
Collateral information

Aims of this session

- At the end of this session, you should:
  - Get a sense of current issues in psychological testing
  - Gain a basic understanding of advanced concepts like measurement invariance

Outline

Measurement invariance

Latent class models

Monotonicity

Unfolding models

Overfitting

Network psychometrics

Collateral information

Measurement invariance

- Measurement invariance is an important (yet under-investigated) assumption in psychological testing.
- It states that the measurement device being used (i.e., the test) functions equivalently (i.e. is invariant) across different groups of individuals, in different contexts, at different times, in different languages, etc.

Measurement invariance

- For example, say we are studying a vocabulary test, the test should function in the same way for all examinees that are going to take it. For example, all items should have the same degree of difficulty, so that differences observed in scores cannot be imputable to the test changing properties.
- If you take the analogy of a measuring tape, you can only compare the height of two chairs if the tape has fixed properties (for example, it's not elastic): If so (or at least to a sufficient extent), we can say that the tape is invariant across the measurement of the chairs.

Measurement invariance

- Mathematically, measurement invariance implies that the function that links item responses (i.e., the observations) and the construct (i.e., the person's latent attribute) is the same across groups. In other words, the item-response function is fixed/invariant.
- If we assume a congeneric model to be true:

$$
\begin{aligned}
x_{i,j} &= \tau_{i,j} + e_i \\
\tau_{i,j} &= a_j \times \tau_i + b_j \\
e_i &\sim \mathcal{N}(0, \sigma_j^2)
\end{aligned}
$$

- Here, (strict) measurement invariance exists across groups of individuals, if the items have fixed values for a_j (loadings), b_j (intercepts) and σ_j^2 (residual variances).

Measurement invariance

Degrees of invariance

- Measurement invariance is typically investigated using Multigroup Confirmatory Factor Analysis (CFA), and typically in several steps:
  - Structural invariance
  - Weak/Metric invariance
  - Strong/Scalar invariance
  - Strict invariance

Degrees of invariance

- The process is generally as follows:
  - There is evidence of *structural invariance* if the factor structure being assumed fits well separately in the different groups.
  - If there is structural invariance, we then compare loadings across groups. If the loadings are similar (a_j), we then say that we have evidence of *metric/weak invariance*.
  - If we have weak invariance, we then compare intercepts (b_j) across groups. If they are similar, we then say that we have *scalar/strong* invariance. It is generally thought to be the requirement when comparisons across groups.
  - If we have strong invariance, we then compare residual variances (σ_j^2) across groups. If they are similar, we then say that we have *strict* invariance. Strict invariance implies that the instrument is equally reliable across groups.

Measurement invariance

Differential Item Functioning (DIF)

- In the context of Item-Response Theory, the question of measurement invariance is in general referred to as the study of *Differential Item Functioning*.
- It refers to the same idea, and uses a similar process of model comparisons where we gradually constrain models to have equal item parameters across the groups.
- An instrument is invariant if it does *not* present Differential Item Functioning (i.e., the items do not function differently across groups).
- Measurement invariance can also be studied in network models.

Measurement invariance

Violations

- The violation of measurement invariance is a threat to the validity of comparisons that may be made across test takers. This is because any difference observed could be imputable to the test functioning differently.
- The lack of measurement invariance implies that comparisons may be biased and/or unfair. Because measurement occurs in many "contest" contexts (personnel selection, college admissions, etc.) measurement invariance (or the lack thereof) has implications on *fairness* in selection.

Measurement invariance

Measurement invariance vs. selection invariance

- While the lack of measurement invariance suggests the presence of a bias, evidence for measurement invariance does not imply that there necessarily is no bias.
- The concept of the absence of bias in selection contexts is often referred to as *selection invariance*.

Measurement invariance

Longitudinal measurement invariance

- Measurement invariance is also an important assumption when studying variations across time. This is often referred to as *longitudinal measurement invariance*.
- For example, if you are comparing means of stress before vs. after an intervention, you are making the assumption that your measurement of stress is invariant across time, so that any differences found are not attributable to the functioning of the instrument.
- Longitudinal measurement invariance is notably threatened by effects such as training effects, acquiescence, fatigue, or observer-expectancy bias.

Outline

Measurement invariance
Latent class models
Monotonicity
Unfolding models
Overfitting
Network psychometrics
Collateral information

Latent class models

Kind vs. continuum

- A common assumption underlying most psychometric practices (e.g., Cronbach's α, Exploratory Factor Analysis, Confirmatory Factor Analysis, Item-Response Theory analysis) is that a latent attribute (e.g., conscientiousness) is assumed to cause the observed item scores (e.g. "Agree" to "I am never late at work.").
- However, the variable type of the underlying attribute is rarely discussed.
- In fact, it is in general assumed that this latent variable is a continuous variable with a Normal distribution. Is it always the case?

Latent class models

Kind vs. continuum

- When a latent attribute is assumed to be continuous, we often refer to it as a *trait*.
- When a latent attribute is assumed to be categorical, we often refer to it as a *class*. There can be 2 or more classes, and they can be ordered (or not).
- It is hard (although some have advanced possible methods) to empirically compare the two, in particular because a variable could, for example, also be comprised of classes within which there is a continuum (e.g., two groups of non-clinical vs. clinical depression, within which there are continua of depression levels), which is referred to as a *mixture*.

Latent class models

Latent class analysis

- *Latent Class Analysis (LCA)* is similar to "traditional" exploratory/confirmatory factor analysis, except that the latent attribute is a categorical variable.
- Latent class models also exist in the Item-Response Theory (IRT) tradition (i.e., based on logistic relations between latent variables and item scores), they are generally referred to as *Latent Class IRT* models.

Latent class models

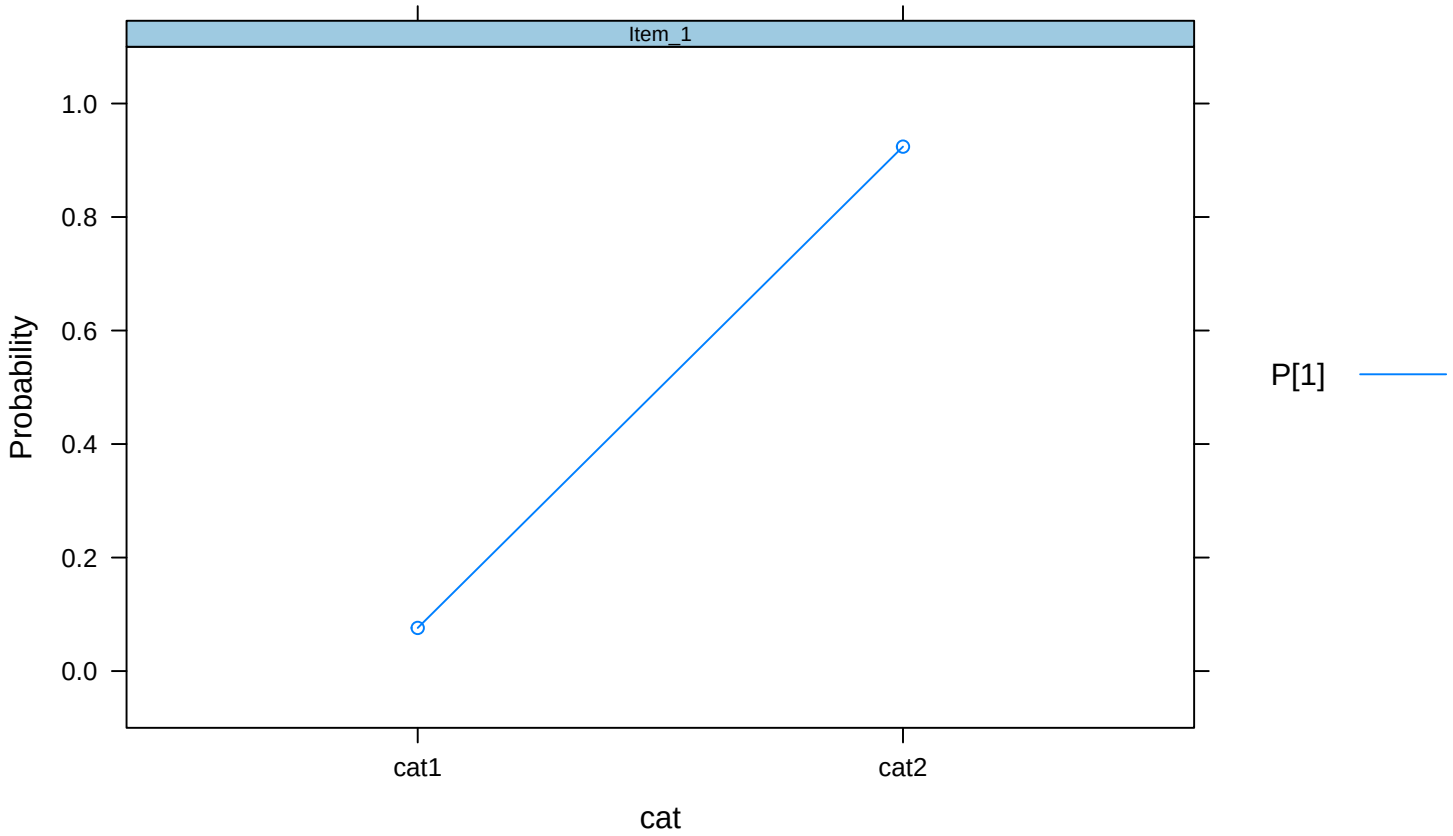

Figure: Based on the `mirt` package (Chalmers, 2012), predicted probabilities of 5 items according to a unidimensional Latent Class IRT model

Latent class models

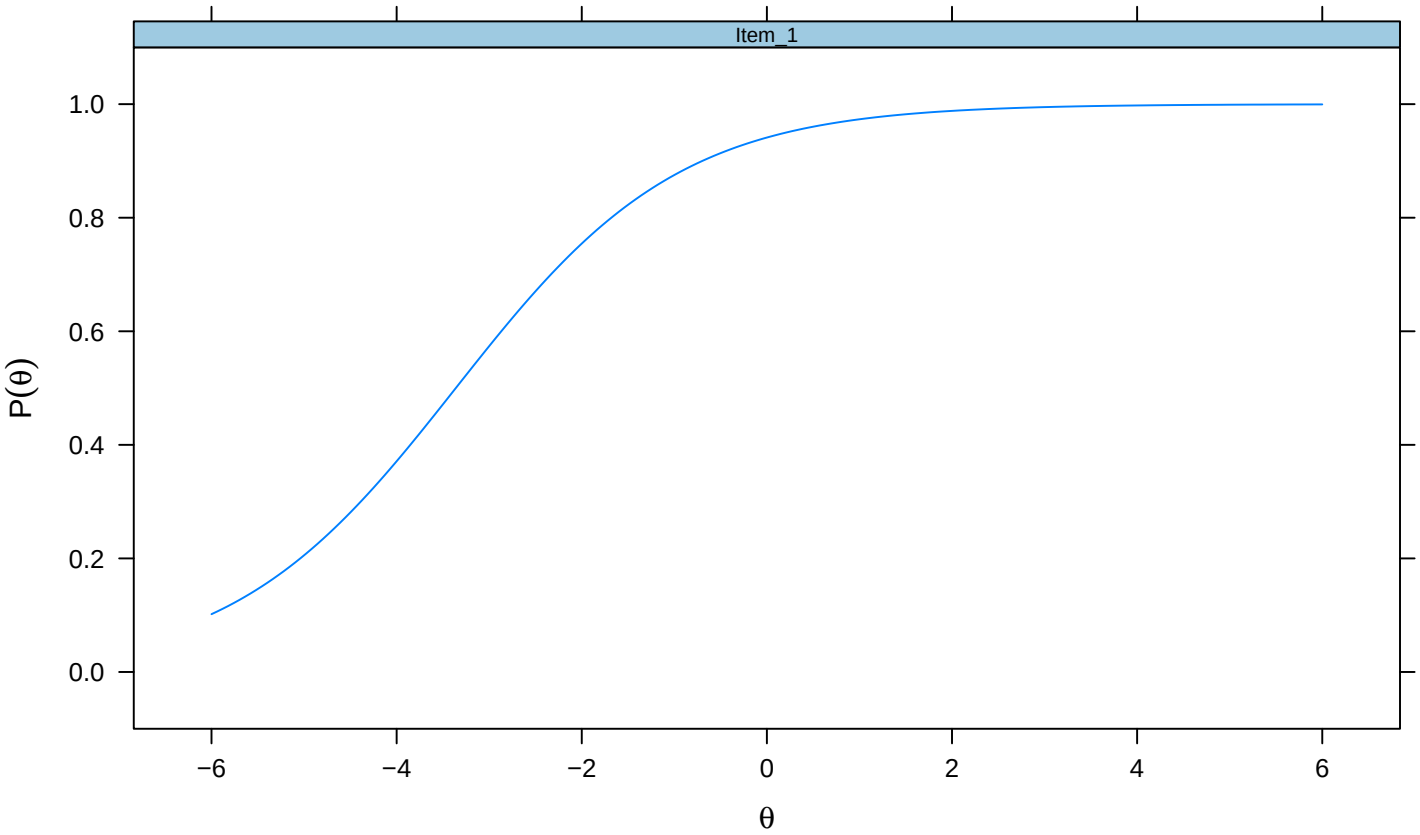

Figure: Compare with a unidimensional latent *trait* IRT model

Outline

19

Measurement invariance
Latent class models
Monotonicity
Unfolding models
Overfitting
Network psychometrics
Collateral information

Monotonicity

- Most measurement models make the assumption of *monotonicity*.
- This assumption states that, as the latent trait increases, there is a monotonous increase (or decrease for reversed items) in the expectation of the item.
  - As extroversion increases, so does the score expectation for "I like going to parties.", and this is true at every level of extroversion and for all items.
  - As math ability increases, so does your probability of answering correctly to a math problem, and this is true at every level of math ability and for all items.

Monotonicity

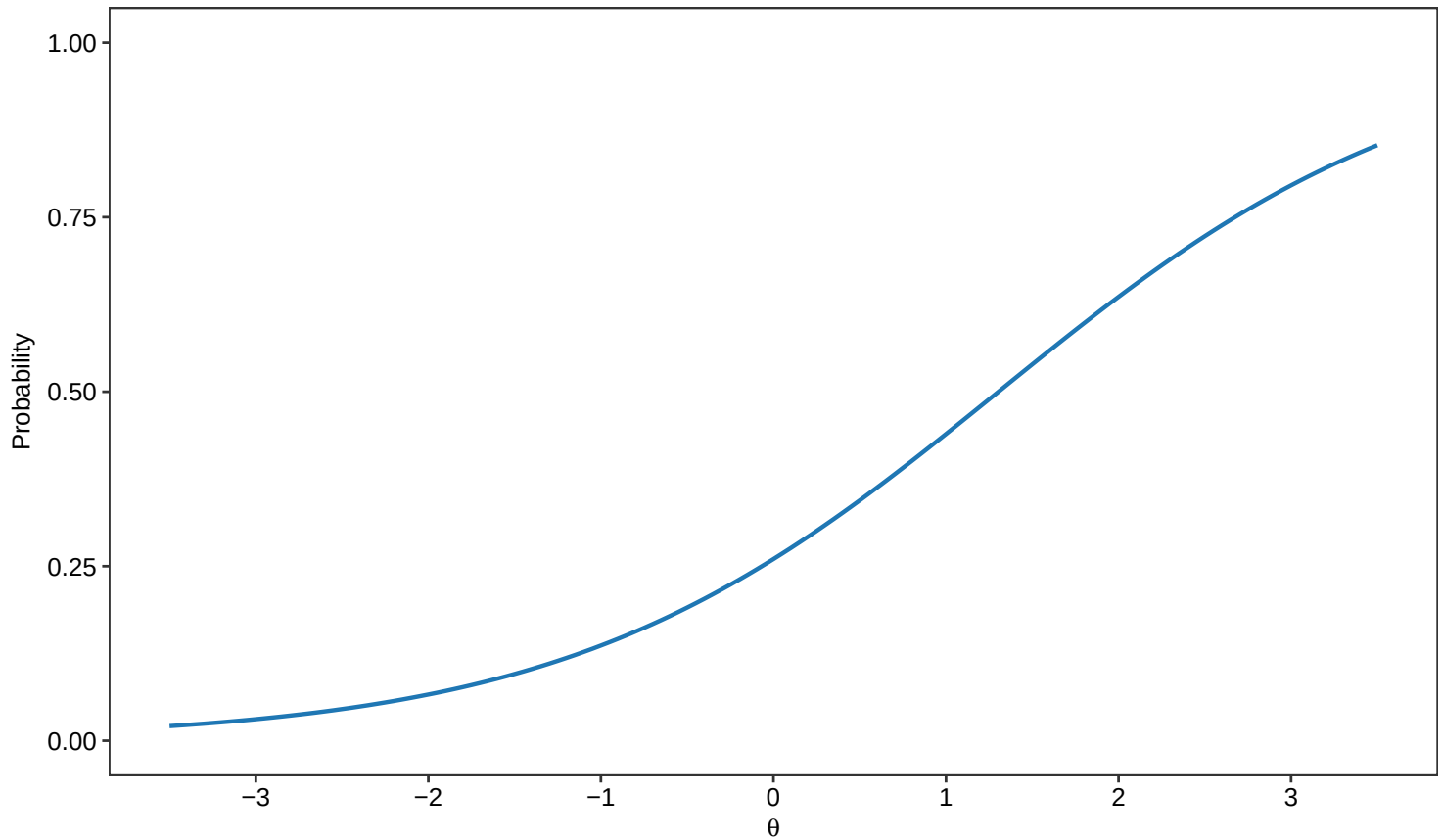

Figure: As math ability increases (x axis), so does the probability of answering correctly to a math problem (y axis).

Assuming monotonicity

- Psychometric models often take the form of monotonous functions, especially:
  - Linear functions (CTT, Congeneric CFA/EFA models)
  - Logistic functions (Binary IRT)
- Because these functions are assumed to represent the relation between the construct and the item responses, these models *assume* monotonicity, usually without *testing* it.

Testing monotonicity

- To investigate monotonicity, we typically use non-parametric IRT methods (i.e., IRT methods where the response function is more data-driven), such as:
  - Mokken Scale Analysis (MSA), where we typically look at an item's (average) observed score at different levels (usually 3-4 levels) of "rest-scores" (the sum score of the rest of the items).
  - Non-parametric IRT (i.e. IRT models that follow more closely the data without taking a pre-specified shape like the logistic function). Splines models are often used.

Monotonicity

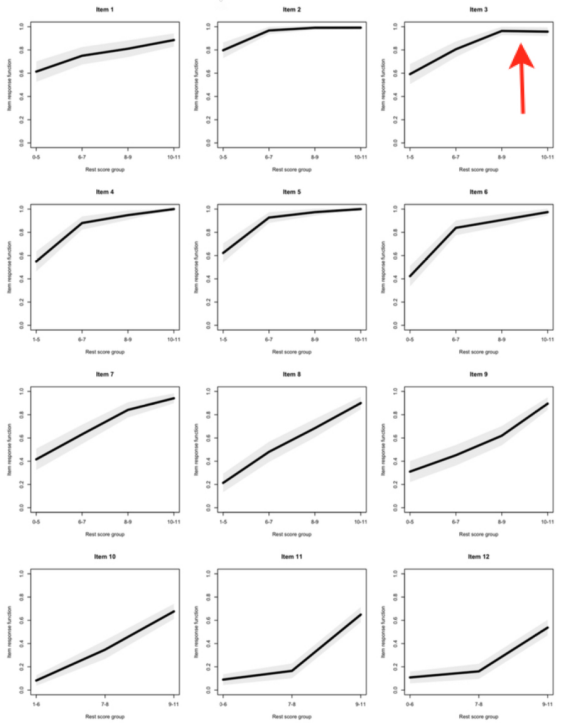

Figure 2. Item response functions of the last series of the Standard Progressive Matrices (SPM-LS) items (with 95% confidence intervals).

Figure: Item response curves in a Mokken Scale Analysis (Myszkowski, 2020), where one item shows a slight non-monotonicity.

Monotonicity

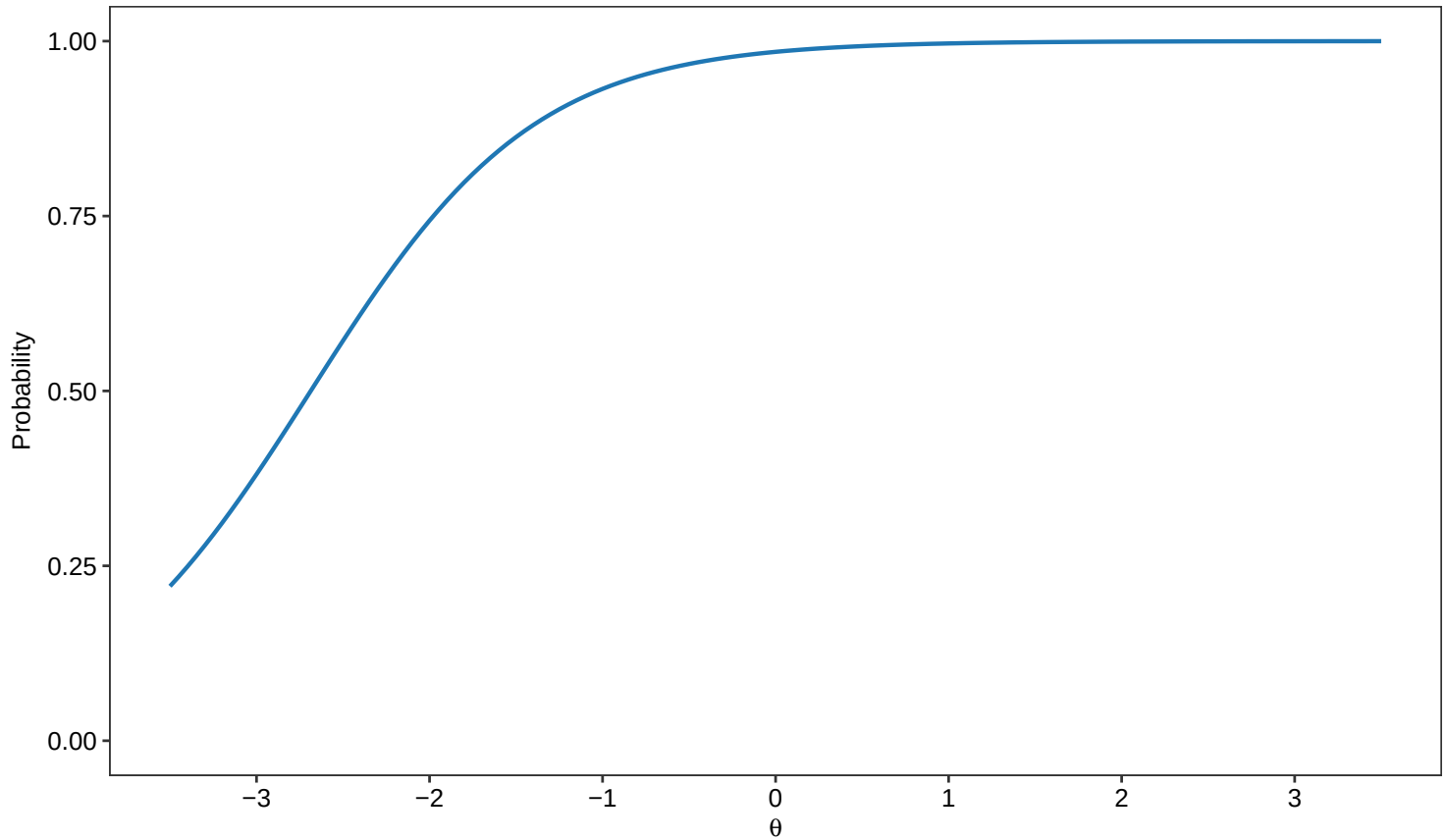

Figure: A logistic IRT response function with a violation of monotonicity.

Monotonicity

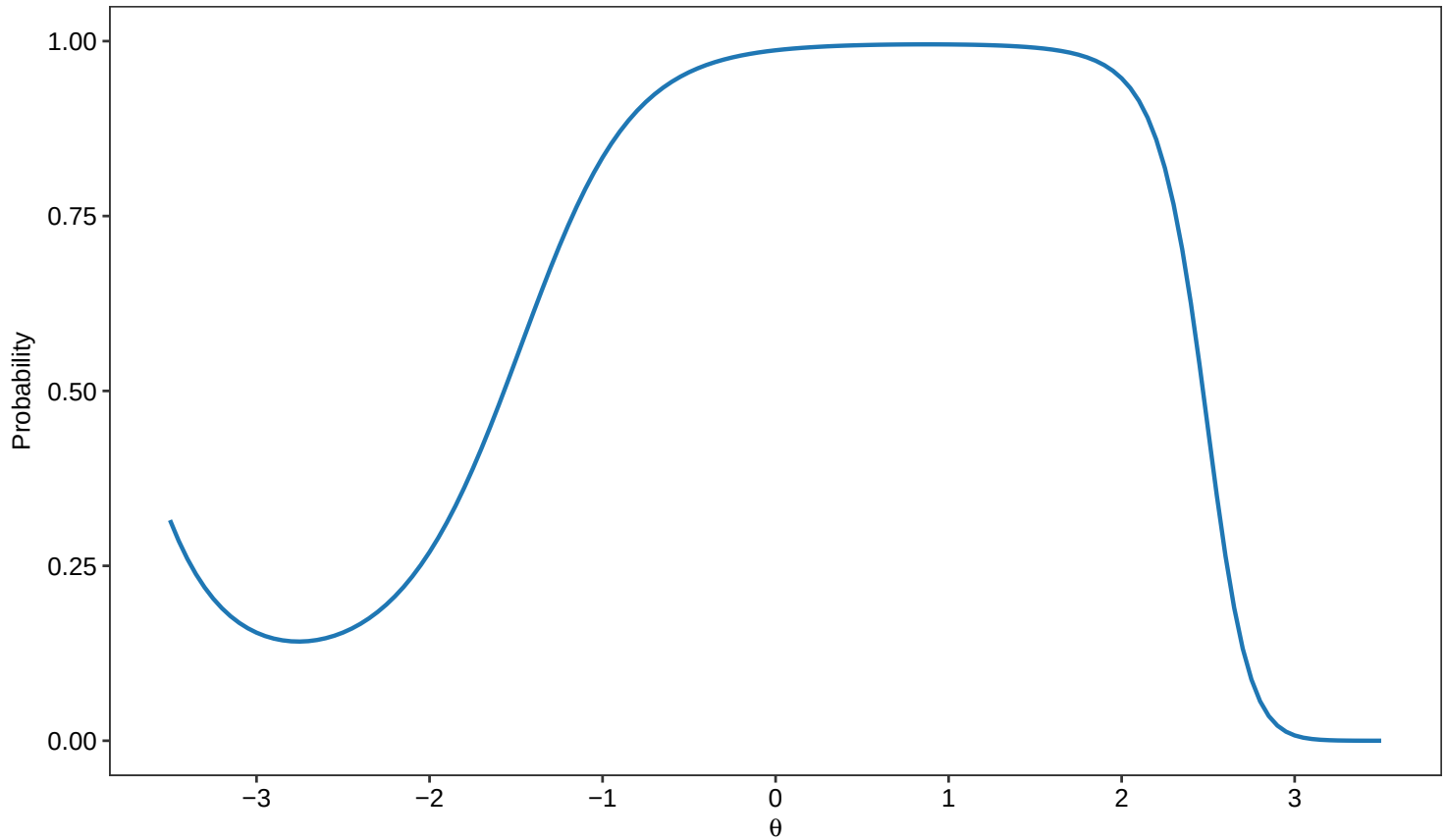

Figure: The same item with a spline IRT response function, showing a violation of monotonicity.

Outline

Measurement invariance
Latent class models
Monotonicity
Unfolding models
Overfitting
Network psychometrics
Collateral information

Unfolding models

Unfolding models

- Sometimes, rather than expect the item response to increase or decrease monotonously with the trait, we actually expect that it first goes in one direction (monotonously), then reaches an optimal/ideal point, and then reverses in the other direction (monotonously).
- In other words, it is the *distance* between the person and the item that is predictive of the item score.
- *Unfolding models* (also called *ideal point models*) allow to deal with these situations. They essentially work by changing the response function (the mathematical function that links the latent variable and the item response).

Unfolding models

Unfolding models

- For example, consider the item "I am concerned about artificial intelligence." as an indicator of AI enthusiasm. The extreme AI enthusiast may say "no". But so would a person who may not be interested at all! There is thus an ideal point of enthusiasm, that is not too low and not too high, that would lead to endorsing the item ("Yes" for a binary item, "Totally agree" for a Likert scale item).

Unfolding models

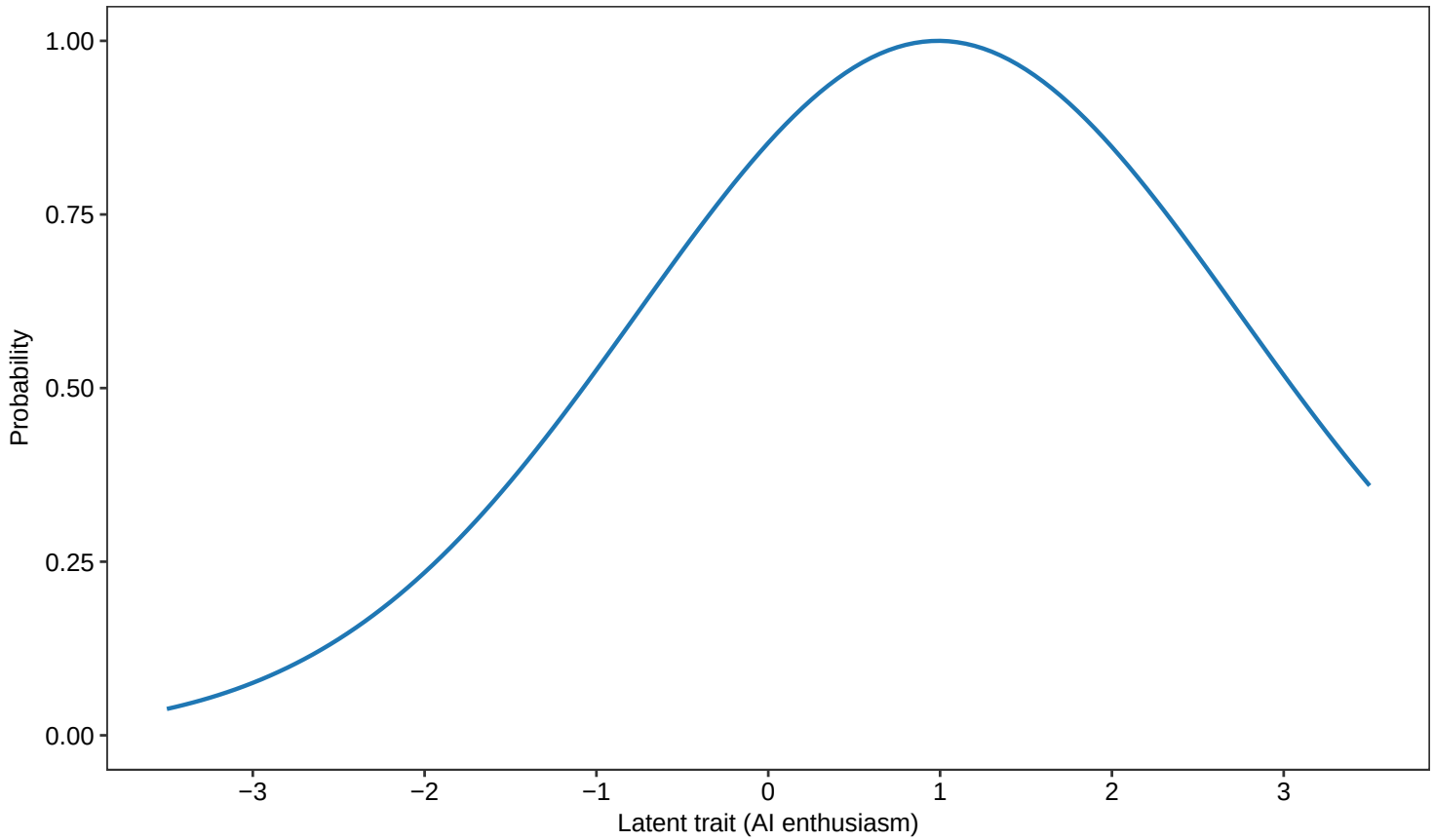

Figure: Based on the `mirt` package (Chalmers, 2012), response function of an ideal point model for a binary item

Unfolding models

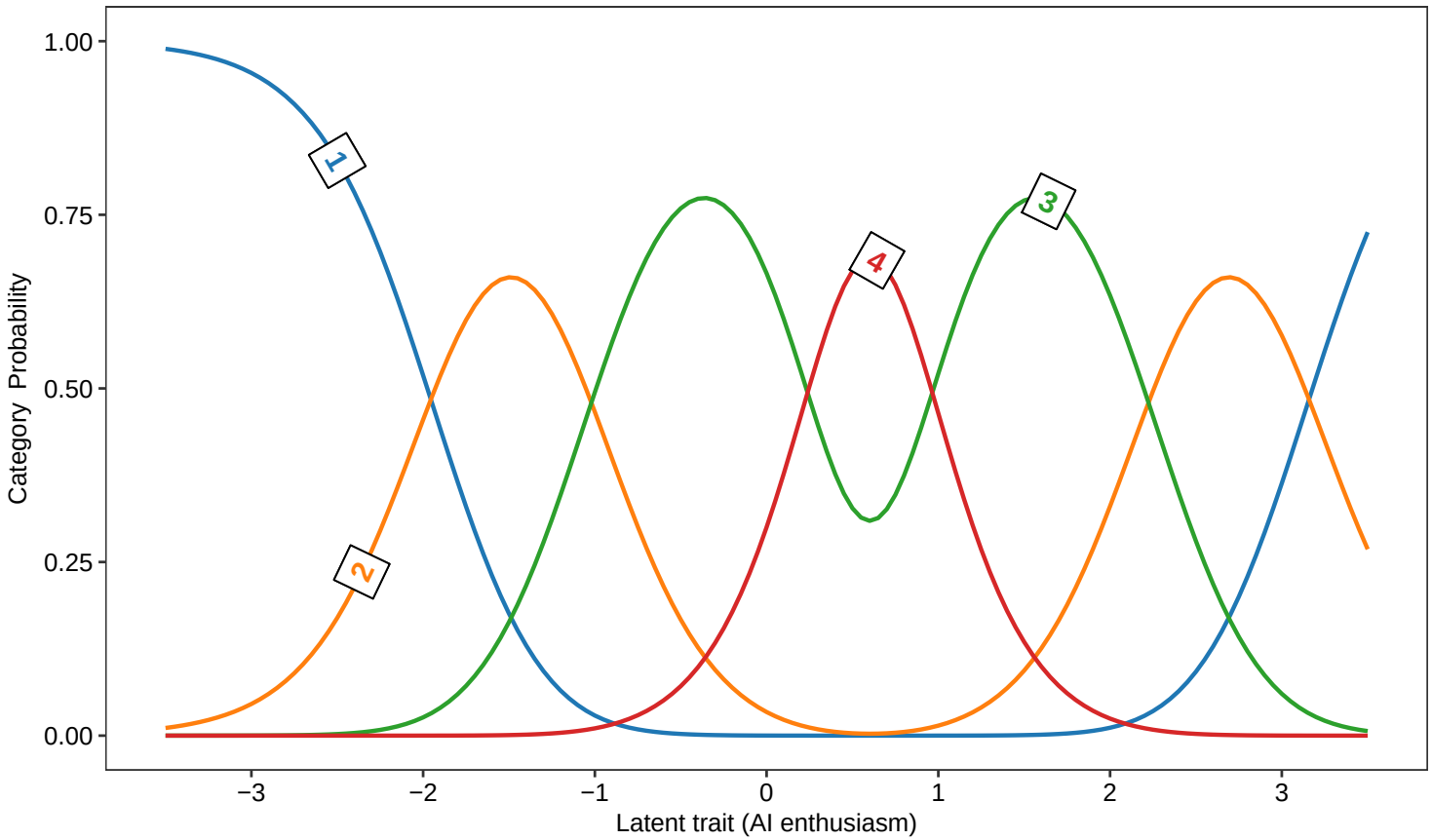

Figure: Based on the `mirt` package (Chalmers, 2012), response function of an ideal point model (the Generalized Graded Unfolding Model) for a polytomous item.

Outline

Measurement invariance
Latent class models
Monotonicity
Unfolding models
Overfitting
Network psychometrics
Collateral information

Overfitting

- Psychometric models are statistical models.
- A frequent danger in statistical modeling is to explore and confirm in the same dataset, which usually results in false positives.
- In psychometrics, this issue is found when *the same dataset* is used to explore for a measurement model (for example, with Exploratory Factor Analysis to identify a structure) and then to confirm the structure (for example, with Confirmatory Factor Analysis).
- In other words, you should not explore and confirm in the same dataset, or you will risk overfitting (having an artificially inflated fit for the confirmatory model).

Overfitting

Solutions

- What are the solutions?
  - Sample again after exploring, and cross-validate (confirmatory analyses) in the second sample.
  - Split the sample in two prior to analysis, and use the first sub-sample for exploration, and the second one for confirmatory analysis.
  - Go only for confirmatory analyses if you have a clear hypothesis

Overfitting

2.3. Overall data analysis strategy

2.3.1. Data splitting

As explained earlier, our approach consisted of two main phases: 1) Investigating the structural validity of the VAST, and consequently creating the VAST-R, and 2) confirming the good structural validity of the VAST-R. Because here the confirmatory analyses performed in the second step are directly derived from the item selection subsequent to the first exploratory analysis, the same data could not be used. Using the same sample for both steps in our case would present a danger of *overfitting*, meaning that we would end up with a new instrument that would by design have good properties in the sample, but whose properties could not necessarily be generalizable to another sample (Hastie et al., 2011).

For this reason, we first split the original sample, using a *2-fold* data split (also called *holdout*) method for cross-validation (Hastie et al., 2011; Zhang, 1993). To do so, we used the R package 'caret' (Kuhn, 2008; Kuhn & Johnson, 2013). Using this package (heavily used in

Figure: Data splitting strategy for exploration (Exploratory IRT) and confirmation (Confirmatory IRT) (Myszkowski & Storme, 2017)

Overfitting

predictive modeling in R), we partitioned the data into two "equivalent" parts. By equivalent, we mean that observations of the original dataset were classified in one group or the other randomly, but ensuring that the two groups have approximately the same distribution of VAST total scores – thus ensuring groups of similar "T". Concretely, this is achieved by using quotas per quintile. It consists in randomly assigning observations to the groups within each quintile of the distribution of the VAST scores (Kuhn, 2008). As a result, we assumed that the different levels of "T" were equivalently represented in each group. "Sample 1" was used as the subsample used in the first study (exploration of the VAST and construction of the VAST-R), and "Sample 2" was the subsample used in the second study (cross-validation of the structural validity of the VAST-R).

2.3.2. Verifying equivalence between the samples

The assignment process being the result of a randomized assignment from each quintile, the two resulting group sizes were not assumed to be perfectly equal – but at least very similar (Kuhn, 2008). In our case, group sizes were very close ($n_1 = 275$, $n_2 = 272$). The two groups also logically had very close VAST means ($M_1 = 35.8$ years, $M_2 = 35.8$ years, Cohen's $d = 0.01$, $t(545) = 0.11$, $p = 0.91$). Additionally, there were no significant group differences in terms of mean ages ($M_1 = 20.7$ years, $M_2 = 20.8$ years, Cohen's $d = 0.08$, $t(545) = -0.88$, $p = 0.38$) or gender distributions ($n_{1,\text{ Male}} = 120$, $n_{1,\text{ Female}} = 155$, $n_{2,\text{ Male}} = 107$, $n_{2,\text{ Female}} = 165$, Cramer's $V = 0.04$, $\chi^2(1) = 0.87$, $p = 0.35$). This indicates that we had two samples that were similar by different aspects, though independent.

Figure: Data splitting strategy for exploration (Exploratory IRT) and confirmation (Confirmatory IRT) (Myszkowski & Storme, 2017)

Outline

Measurement invariance
Latent class models
Monotonicity
Unfolding models
Overfitting
Network psychometrics
Collateral information

Network psychometrics

Network psychometrics

- Network psychometrics represent the relations between the item scores without using latent variables (i.e., they assume that there is no variable that is a common cause to the item scores).
- The starting point of network psychometrics is usually a matrix of relations (e.g., correlations, partial correlations, etc.) between all items.

Network psychometrics

Dense to sparse networks

- In a *dense network*, all variables are connected (all nodes are connected by edges).
- The first step of network analysis is thus usually to make to form a *sparse network* instead, meaning that we remove the small relations (remove some edges).
  - Using thresholds (e.g., no r below $.1$ or above $-.1$)
  - Using statistical significance as a criterion to keep relations
  - Using regularization methods (e.g., LASSO regularization)
- Some methods have a better reproducibility than others.

Network psychometrics

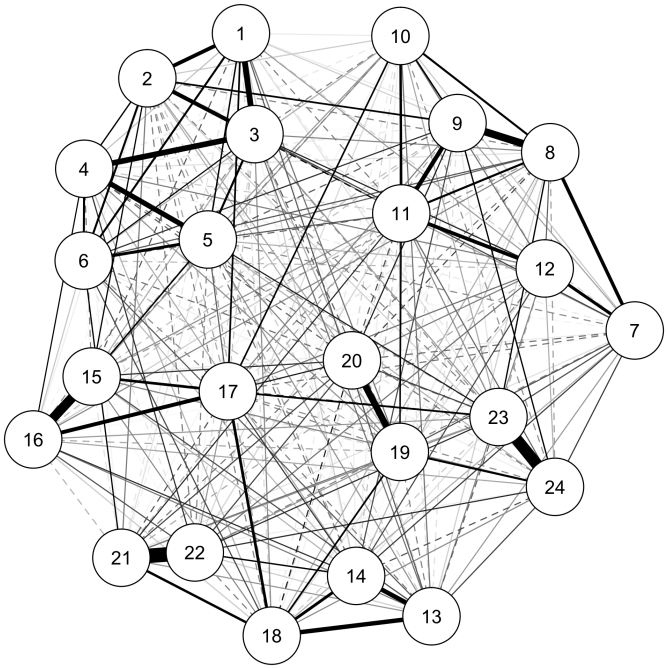

Figure: Graph representation of a dense network

Network psychometrics

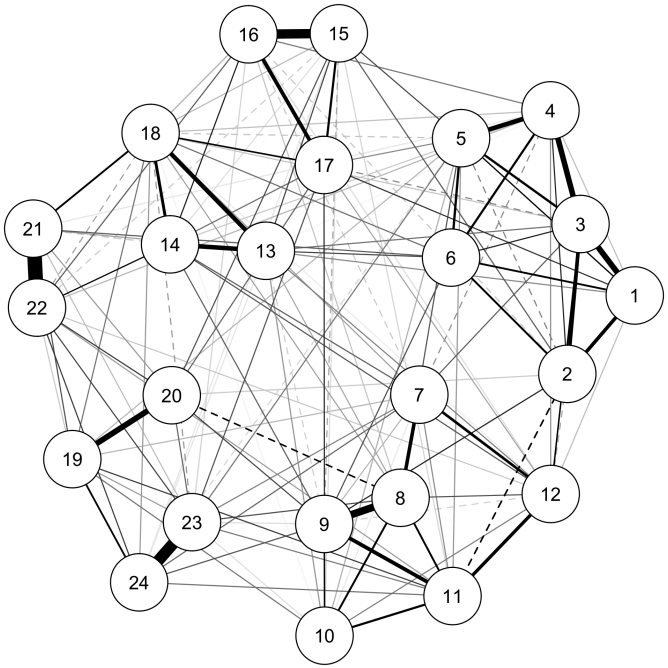

Figure: Graph representation of a sparse network (same data)

Network psychometrics

To factor models

- Once we have these models, we can notably identify clusters of items, which we can then use to build a factor model that can be tested with, for example, Confirmatory Factor Analysis.
- This exploratory technique finds clusters of items, and thus is an alternative to Exploratory Factor Analysis (EFA). We often refer to it as Exploratory Graph Analysis (EGA, Golino & Epskamp, 2017).
- There are algorithms to identify clusters automatically from a network which can be applied in this context, such as the Louvain method (Blondel, Guillaume, Lambiotte & Lefebvre, 2008).

Network psychometrics

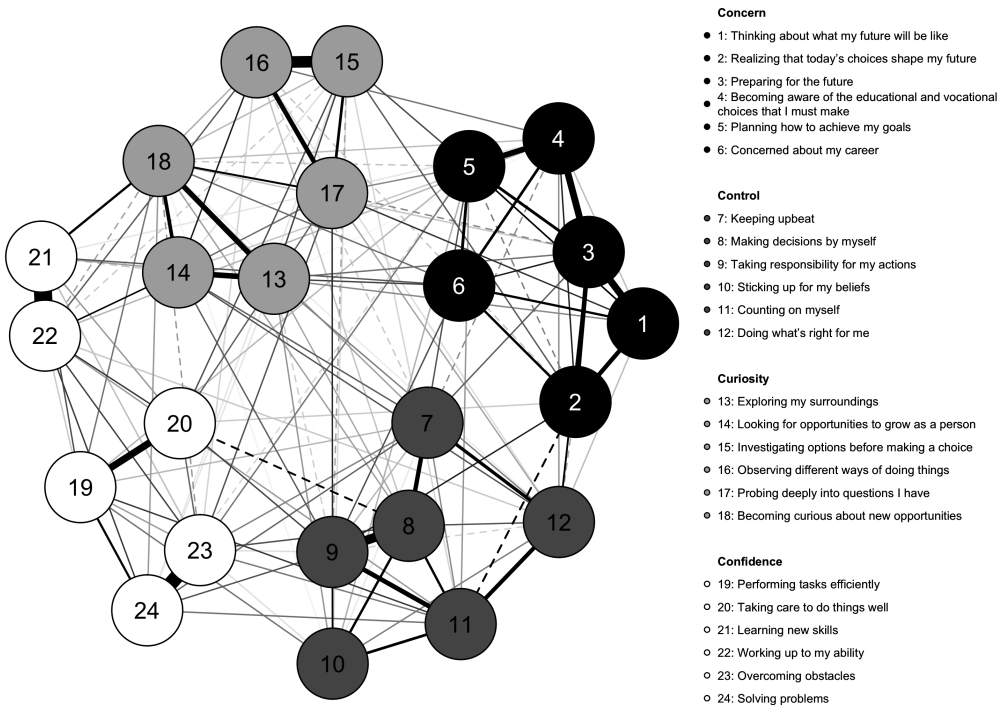

Figure: 4 clusters identified from the network using the Louvain method

Testing sparse networks

- Dense networks are saturated models and therefore have a trivial (uninformative) perfect fit.
- However, sparse networks can be tested using the same framework as Confirmatory Factor Analysis (Kan et al., 2020).
- New software packages like `psychonetrics` (Epskamp, 2021) allow CFA vs. network models comparisons based on the traditional fit indices of Structural Equation Modeling (CFI, TLI, RMSEA, etc.), as well as blends of the two (e.g. the items are explained by latent variables which are organized as a network).

Outline

Measurement invariance
Latent class models
Monotonicity
Unfolding models
Overfitting
Network psychometrics
Collateral information

Collateral information

- Computerized testing now allows the collection, storing and processing of additional information that can be used in test scoring.
  - Response times
  - Number of response changes
  - Which wrong answer ("distractor") was selected
  - Eye movements

Collateral information

- These additional sources of information can be used to:
  - Detect "aberrant" behaviors (e.g., cheating, guessing, faking, overthinking, etc.)
  - Increase the accuracy of the estimation of the constructs
  - Better understand how a test works
  - Better understand relations between constructs

Collateral information

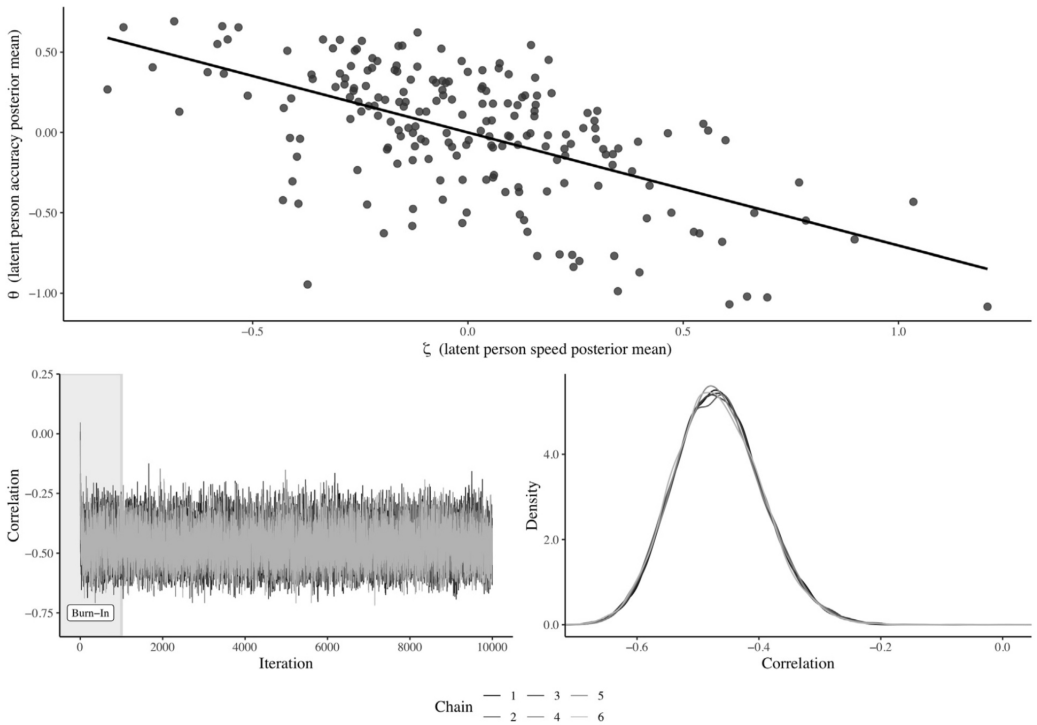

Fig. 2. Correlation between person speed and accuracy for the 3PNO-2PLN model.

Figure: Examining the speed and accuracy trade-off in visual aesthetic sensitivity tests (Myszkowski, 2019)

Collateral information

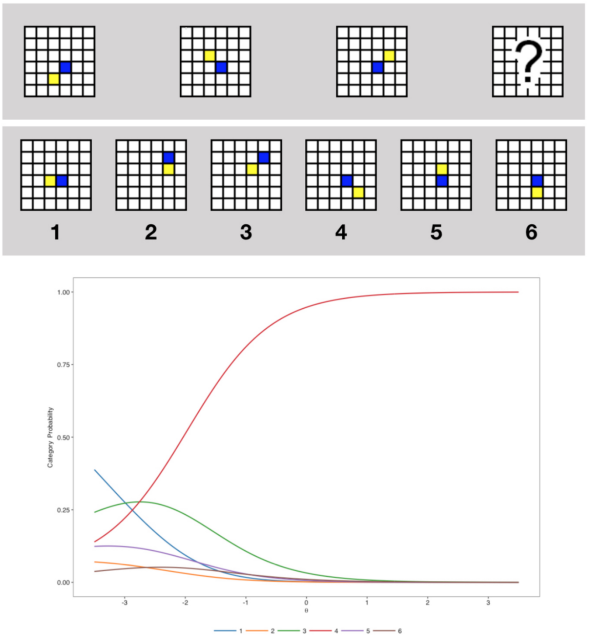

Figure 1. An example item of the GF20 (**top**) and the associated category characteristic curves as estimated by the 3-Parameter Nested Logit model (**bottom**). The correct response (4) is increasingly probable as θ_j increases. However, the response category 3—which is the only distractor response where the blue and the yellow squares are (correctly) not adjacent—would be more probably selected by individuals with low abilities ($\theta_j \approx -2.7$), while the category 1 would be more probably selected by individuals with even lower abilities ($\theta_j < -3$)—thus showing that the choice of distractor may be informative of θ_j.

Figure: Using distractor responses to boost test reliability (Storme et al., 2019)