Test theories

Nils Myszkowski, PhD

Outline

What are "testing theories"?

Classical Test Theory

Modern Test Theory

Other approaches

Aims of this session

- At the end of this session, you should be able to:
  - Understand the main differences between the major testing theories
  - Understand the practical implications of these differences
  - Identify the approach used in a given psychometric analysis

Outline

What are "testing theories"?

Classical Test Theory

Modern Test Theory

Other approaches

Testing theories

- There are different approaches to how psychological testing works.
- These different approaches are not only different conceptually, but they have practical implications on how to use item/test scores.
- In this session, we will see the major approaches, which are Classical Test Theory and Modern (Item-Response) Test Theory.
- We will also cover new approaches, such as Network Psychometrics and Formative Measurement.

Before we start

- Before we start, let us consider a typical testing situation, where an individual/case i has responded an item j.
- The combination of the two has produced an observed score $x_{i,j}$.
- This process is repeated over a sample of several individuals/cases (i) and a sample of items (j).

What are testing theories?

- Testing theories are frameworks to make intepretations (inferences) based on a set of observed scores $x_{i,j}$.
- Testing theories are (more or less directly) concerned with how $x_{i,j}$ can be used to infer on a construct or more generally the functionning of an individual.

An example

- The question is essentially "Now that we have the item scores for a sample of individuals, what do we do with them?"

Item.1	Item.2	Item.3	Item.4	Item.5	Item.6	Item.7	Item.8	Item.9	Item.10	Item.11	Item.12
1	4	5	2	3	1	2	1	3	1	2	4
3	4	2	8	3	3	2	8	3	1	2	8
1	4	5	4	3	2	2	3	3	2	2	1
2	4	4	2	3	3	2	4	3	2	2	4
2	4	5	2	3	2	2	1	1	2	2	4

Outline

What are "testing theories"?

Classical Test Theory

Modern Test Theory

Other approaches

Classical Test Theory

- Classical Test Theory (or "CTT") goes by other names, including notably True Score Theory.
- It is certainly by far the most popular approach to testing in most textbooks and instruction.
- It is implicitly applied often (for example, when computing Cronbach's α for reliability or sum scores to estimate a construct).
- It has however shortcomings, which we will discuss.

Classical Test Theory

- Definition
- Variants
- Implications and practical advantages
- Limitations

Classical Test Theory

- CTT is defined as a framework in which "the raw score obtained by any one individual is made up of a true component and a random error component" (Kline, 2005).
- Its central idea is essentially that, for an individual i and item j, there is a true (pure, or error-free) score (τ_i), and that an observed score is that true score, with a random error part (e).
- It is generally summarized mathematically by:

$$x_{i,j} = \tau_i + e$$

With the error component being random (normally distributed with a zero mean and a σ^2 variance):

$$e_i \sim N(0, \sigma^2)$$

A note on local independence

- Note that CTT implicitly formulates the assumption that the item scores $x_{i,j}$ are not related to one another, beyond them being explained by the same true score τ_i.
- This assumptions is usually referred to as the assumption of **local independence**.

Let's go further

- There are different (more or less constrained) versions of CTT (see Lord & Novick, 1968).

$$x_{i,j} = \tau_i + e$$
$$e \sim N(0, \sigma^2)$$

- What varies between them is:
  - Is the "true" score identical or different across the j items? (τ_i or $\tau_{i,j}$) ?
  - Do items produce the same amount of error? (is there an overall σ^2 or different σ_j^2)

Classical Test Theory

Definition

Variants

Implications and practical advantages

Limitations

CTT for parallel tests

- The CTT variant for **parallel** tests is the most constrained version.

$$x_{i,j} = \tau_i + e$$
$$e \sim N(0, \sigma^2)$$

- The true score τ_i does not depend on the item. It is a fixed value for an individual i.
- All items produce errors of the same magnitude σ^2.
- In other words, here items are perfectly interchangeable: They produce the same amount of errors and have the same true score.

CTT for τ-equivalent tests

- The CTT variant for τ-**equivalent** (read "tau-equivalent") tests is less constrained.

$$x_{i,j} = \tau_i + e$$
$$e \sim N(0, \sigma_j^2)$$

- The true score τ_i does not depend on the item. It is a fixed value for an individual i.
- Items produce errors of different magnitude σ_j^2.
- In other words, here items only differ in the magnitude of their errors. You can see it as items being more or less "noisy" measures of the construct.

CTT for essentially τ-equivalent tests

- The CTT variant for **essentially τ-equivalent** tests is even less constrained.

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

- The true score part of the equation does not only depend on the individual τ_i, there is also an added constant which differs per item, b_j (often referred to as an intercept). This conveys the idea that items may have different locations/difficulties. For example, for an easy item with $b_j = 2$, we expect to see higher scores $x_{i,j}$ for a given individual than for a more difficult item $b_{j'} = 0$.

CTT for congeneric tests

- The CTT variant for **congeneric** tests (Joreskog, 1971) is even less constrained.

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

- The true score part includes an additive constant b_j, like in the essentially τ-equivalent model, but now, it also includes a multiplicative component a_j (often referred to as a item loading), which represents the idea that items may be more or less, positively or negatively related to the true score $\tau_{i,j}$.

Classical Test Theory

Assumptions to implications

- It is important to note that these variants formulate different assumptions about the similarities between the items.
- Different interpretations can or cannot be made depending on the model that is chosen.

A major advantage

- A major advantage of the most constrained (parallel and τ-equivalent) CTT is that the expectation of the observed score is the true score.

$$x_{i,j} = \tau_i + e$$
$$e \sim N(0, \sigma_j^2)$$

- Implies that $E(x_{i,j}) = \tau_i$.
- This means that, as the number of items increase, the mean of the observed scores tends to the true score.

A major advantage

- Therefore, in this view, if we build a sufficient number of items, the mean of all observed scores will be close to their expectation, and thus to the true score. In other words we will have estimated the true score.
- Another way to see this is that, by repeating measurement, the random error components will end up (more or less) cancelling themselves, leading to us obtaining an estimate of the true score.
- Parallel and τ-equivalent CTT thus enable sum/average score as a way to estimate the true score.
- It is thus very practical.

Classical Test Theory

- Definition
- Variants
- Implications and practical advantages
- Limitations

Disadvantages of τ-equivalence and parallel CTT

- This has however important drawbacks, because the assumptions of τ-equivalence and parallel CTT are quite strict:
  - Items are not necessarily of the same difficulty.
  - Items may not all relate with the same strength with the true score.
- In other words, these forms of CTT make assumptions that are actually hardly realistic.

...and more general problems with CTT

- Normal errors do not work well with binary (or non continuous or bounded) items.
- The true score is not a construct: It remains dependent upon the test. A true (or estimated true) score of 5 is only meaningful when considering the test.
- Most of the time the assumptions of CTT are not tested.
- CTT is a tautology more than a testable model, in that the relation between the observations and the theoretical attribute is "fixed axiomatically" (Borsboom, 2002, p.429). In other words CTT is necessarily true and not open to discussion or research.
- CTT does not reflect a psychological explanation for the scores.

Outline

What are "testing theories"?

Classical Test Theory

Modern Test Theory

Other approaches

Modern Test Theory

Modern Test Theory

Definition

Examples

Implications

Definition

- **Modern Test Theory**, also called **Item-Response Theory** (IRT), adopts a different perspective than CTT.
- It is "a system of models that defines one way of establishing the correspondence between latent variables and their manifestations." and "uses latent characterizations of individuals and items as predictors of observed responses" (Ayala, 2013).
- In other words it is not exactly a theory, but a class of models that relate psychological attributes and item attributes to observations.

Definition

- Because it is a class of models more than a theory, we could formulate IRT broadly as an framework that consists of modeling observed scores $x_{i,j}$ as a function of the person's psychological latent attribute θ_i and the item's attribute(s) a_j, b_j, c_j, etc.

$$x_{i,j} \sim f(\theta_i, a_j, b_j, c_j, \ldots)$$

- A central novelty here, illustrated by $\sim$ algebraically, is that IRT forms testable models.
- The term "Item-Response Theory" comes from the function f, which is called in this framework the **Item-Response Function**.

Origins

- The item-response function takes many forms, but it is certain that originally, IRT was developed to accommodate for ability test items (pass-fail) varying in difficulty (which is problematic in CTT because of the Normal distributional assumption).
- These challenges were recognized by CTT theoreticians, which led to different conceptualizations of difficulty, that were originally based on ordering items (Guttman and Mokken scaling), leading to formalizing the response function as logistic with Rasch modeling.
- Later, Rasch/IRT models were (and continue to be) extended in several ways to accommodate various types of measures (ordinal, continuous, categorical, response times, counts, etc.).

Modern Test Theory

- Definition
- Examples
  - IRT for binary items
  - IRT for ordinal items
  - IRT for continuous items
- Implications

Logistic models

- There are several prominent IRT models, but we will examine briefly the most commonly encountered.
- We will briefly discuss models for binary, polytomous and continuous data.
- We will only focus on IRT models.
- All the models presented here make the assumption of local independence.

Modern Test Theory

- Definition
- Examples
  - IRT for binary items
  - IRT for ordinal items
  - IRT for continuous items
- Implications

Binary data

- Binary IRT is generally based on logistic models. A general form for these models is Birnbaum's (1968) "3PL" (3-Parameter Logistic) model:

$$p_{i,j}(\theta_i, a_j, b_j, c_j) = c_i + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}}$$

- Where items vary in discrimination (a_j), difficulty (b_j) and probability of guessing (c_j).
- Common particular constrained cases are the "2PL" model (where $c_j = 0$) and Rasch model (where $c_j = 0$ and $a_j = 1$).

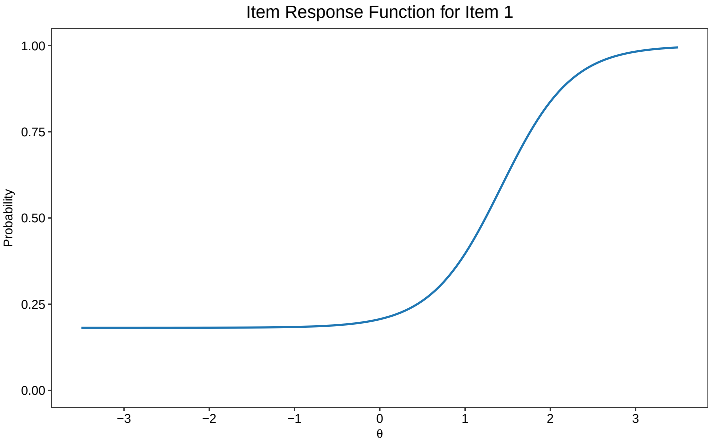


Figure: The predicted probability of success for item 1 with a 3PL model

Modern Test Theory

Examples

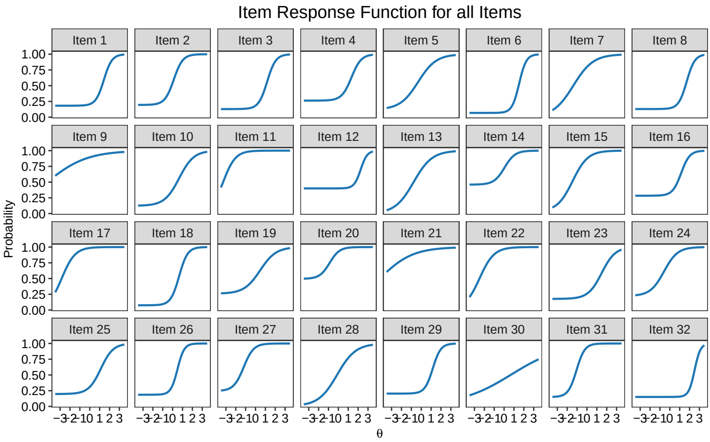


Figure: The predicted probability of success for all items with a 3PL model

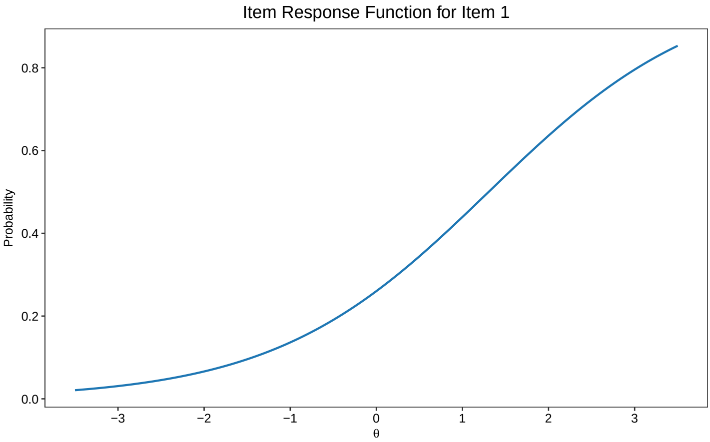


Figure: The predicted probability of success for item 1 with a 2PL model

Modern Test Theory

Examples

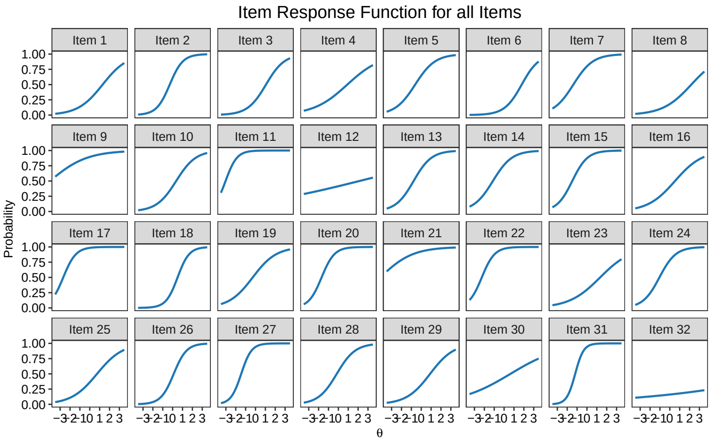


Figure: The predicted probability of success for all items with a 2PL model (note that the lower asymptote is now 0)

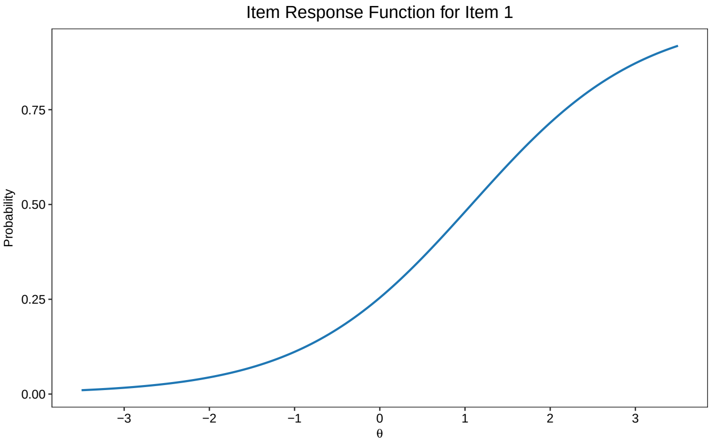


Figure: The predicted probability of success for item 1 with a Rasch model

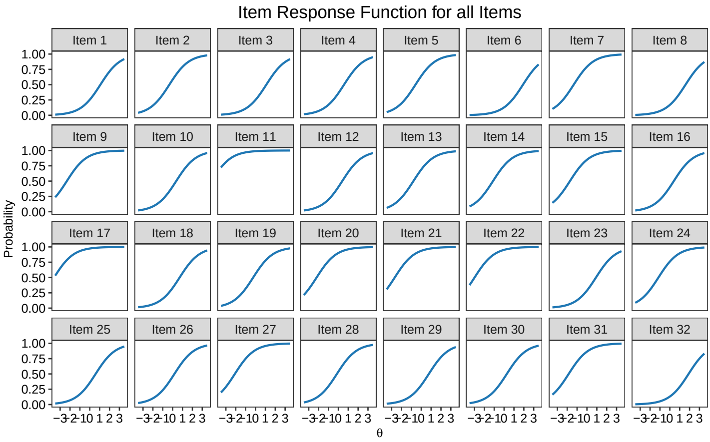


Figure: The predicted probability of success for all items with a Rasch model (note that the all slopes are the same)

Modern Test Theory

Polytomous data

- Models for ordinal responses are combinations of logistic models for each threshold.
- For example, for a 4-point Likert scale, we model simultaneously the probability to respond 2 vs 1, 3 vs. 2 and 4 vs 3. (another approach is to combine models for probabilities of 1 vs. more, 2 vs. more and 3 vs. more, 4 vs more, which usually gives similar results).
- The most popular models (the Generalized Partial Credit Model and Graded Response Model) accommodate for variations between items in discrimination, difficulty and category structure.

Modern Test Theory

Examples

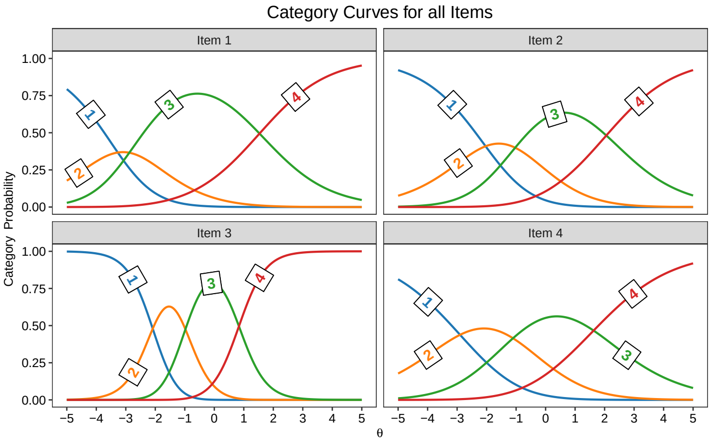


Figure: Predicted probabilities for a 4 point Likert scale with a Generalized Partial Credit Model)

Modern Test Theory

Examples

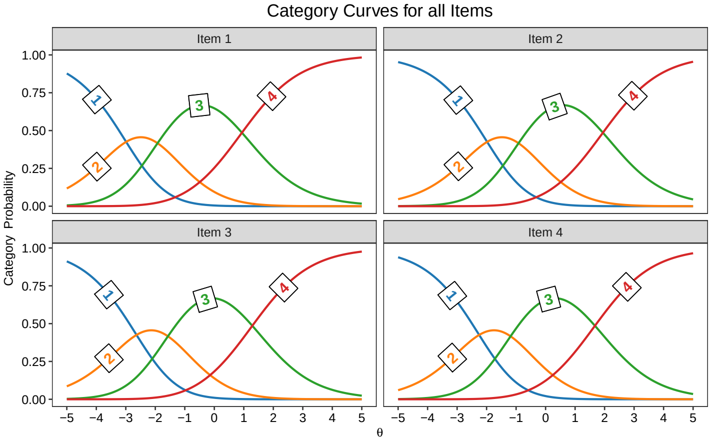


Figure: The more constrained Rasch Rating Scale does not allow items to differ in discrimination or category structure (but only in location/difficulty))

Modern Test Theory

Definition

Examples

IRT for binary items

IRT for ordinal items

IRT for continuous items

Implications

Continuous data

- It has been pointed (Mellenberg, 1994) that the congeneric model (Joreskog, 1971) is actually a linear form of IRT for unbounded continuous item responses.

$$x_{i,j} = (a_j \tau_i + b_j) + e$$
$$e \sim N(0, \sigma_j^2)$$

- Another way to represent this is to consider it as a general linear model of θ_i with an item slope a_j and intercept b_j:

$$x_{i,j} \sim f(a_j \theta_i + b_j)$$

Continuous data

- The congeneric model can therefore be though as a form of IRT model.
- Since other CTT formulations are constrained versions of congeneric CTT, this in fact implies that CTT models in general can be seen as nested within IRT models.
- Note: In practice, CTT techniques and IRT models remain very separate (e.g., usually separate software, separate journal articles, etc.).

Modern Test Theory

- Definition
- Examples
- Implications

Model fit

- An important feature of IRT models is that they can be tested.
- Indeed, since these models form predictions of item scores, these predictions can be compared with the observations.
- We will see how when we discuss structural validity in the validity session.

Factor scores

- In the estimation of these models, the attribute θ_i is estimated, and thus can be used as a measure of the attribute.
- In addition, in IRT, the error of measurement is estimated as conditional upon the latent attribute. In other words, IRT encompasses the idea that a test or an item may be more or less appropriate for different levels (for example, a test could be too easy or too hard to be properly measuring someone's ability).
- Thus individual scores θ_i also come with estimates of measurement accuracy (called Standard Error of Measurement, see session on reliability).

Modern Test Theory

Implications

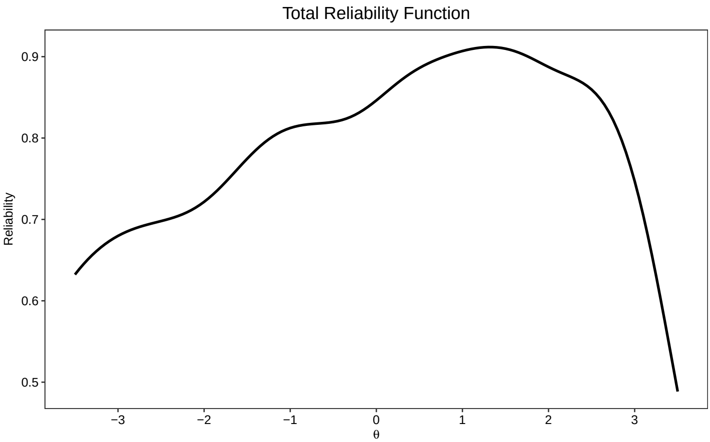


Figure: We see here that reliability is a function of the latent attribute. In other words, some levels of the attribute are better measured than others.

Challenges

- There are still challenges to the use of IRT.
  - It is a statistical model: It performs better with a lot of data.
  - It is a statistical model: It may not fit well the data (but at least you know it!)
  - It is complex and not completely implemented in the most popular statistical packages (SPSS especially).

Outline

What are "testing theories"?

Classical Test Theory

Modern Test Theory

Other approaches

54

Other approaches

- Behavior Domain Theory
- Network Psychometrics

Behavior Domain Theory

- IRT and CTT are based on the idea that a latent construct is the cause for the observations. This conceptualization of measurement is generally referred to as **reflective** measurement or reflective models.
- **Behavior Domain Theory** (also called **Formative Measurement**) makes the exact opposite formulation, where the construct is a composite formed by the items (Bollen & Diamantopoulos, 2015).
- This is based on the idea that, in some cases, the scores are not caused by a common construct, but are samples of a common domain of behavior.

Behavior Domain Theory

- In this theory, the scores do not have a common cause, so they may or may not be appear correlated.
- For example if measuring one's artistic ability, different domains could be the musical domain, the performing arts domain, the visual domain, etc.
- Here more ability in a domain implies more artistic ability. But more artistic ability does not imply more musical ability, for example.
- In other words, the observations for the latent construct, rather than being cause by it.
- Note that this approach is however largely criticized in psychometrics (it is much more largely used in business fields), because it is unclear how it can be tested.

Other approaches

Behavior Domain Theory

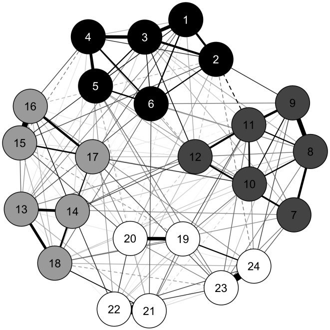
Network Psychometrics

Network Psychometrics

- Another very recent approach is to consider the existence of a latent construct as, itself, an assumption.
- And thus, one may investigate the relations between observations directly, without interpreting them under the lens of reflecting or forming a latent construct.
- This approach thus makes the assumption that there are local dependencies but no common unobserved cause/true score. It then tries to make sense of the local dependencies directly.

Other approaches

Network Psychometrics



Concern

- 1: Thinking about what my future will be like
- 2: Realizing that today's choices shape my future
- 3: Preparing for the future
- 4: Becoming aware of the educational and vocational choices that I must make
- 5: Planning how to achieve my goals
- 6: Concerned about my career

Control

- 7: Keeping upbeat
- 8: Making decisions by myself
- 9: Taking responsibility for my actions
- 10: Sticking up for my beliefs
- 11: Counting on myself
- 12: Doing what's right for me

Curiosity

- 13: Exploring my surroundings
- 14: Looking for opportunities to grow as a person
- 15: Investigating options before making a choice
- 16: Observing different ways of doing things
- 17: Probing deeply into questions I have
- 18: Becoming curious about new opportunities

Confidence

- 19: Performing tasks efficiently
- 20: Taking care to do things well
- 21: Learning new skills
- 22: Working up to my ability
- 23: Overcoming obstacles
- 24: Solving problems

Figure: A network model of a career adaptability scale (Myszkowski, Celik & Storme, in preparation)

Network Psychometrics

- This approach is catching a lot of attention recently, and is especially relevant to uncover structures of behaviors (for example, relations between occurrences of symptoms of a disorder). They are especially useful with longitudinal data.
- Making networks sparse (i.e., selecting which local dependencies to retain and which to remove) can be challenging and/or it can be challenging to make these decisions reproducible. Using these models usually require knowledge beyond traditional psychometrics (e.g., in regularization).
- Network models can be combined with latent variable models in various ways.