Modern test theory in action

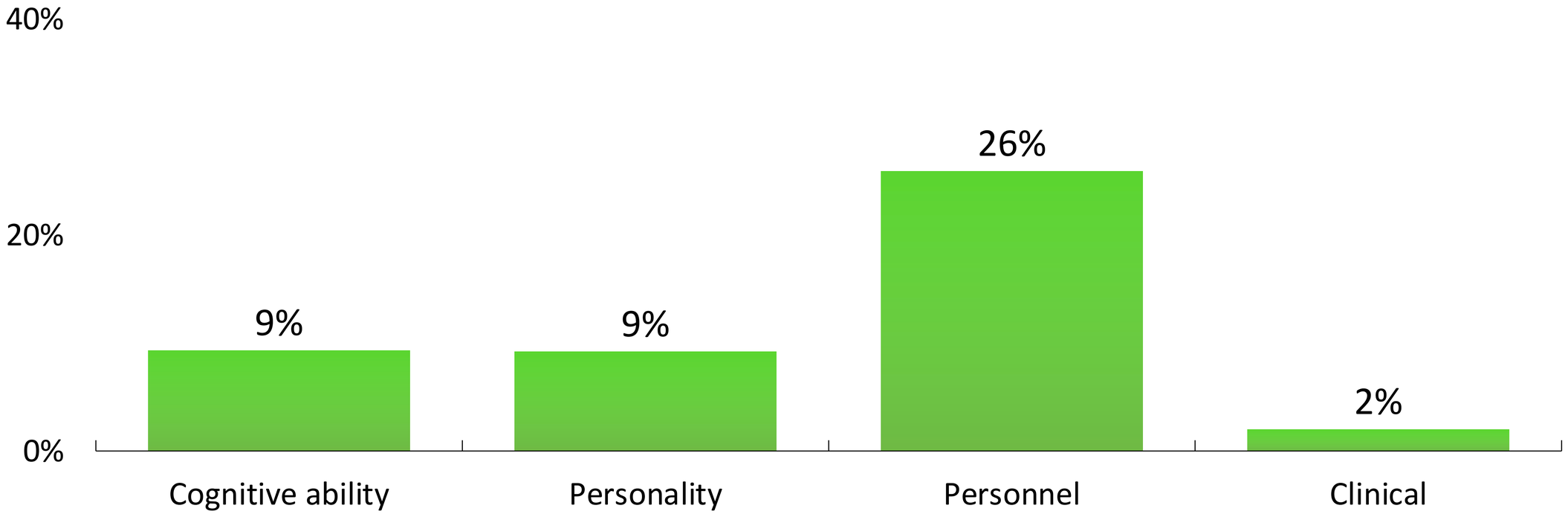
Solving real-world challenges in psychological measurement

Nils Myszkowski – July 10th, 2025

About me

- Associate Professor of Psychology at Pace University, NYC
- Teaching : Psychometrics, statistics and research design, in doctoral clinical/school programs, using SPSS (PsyD) and R (PhD)
- Main research aim : To improve the measurement of creative potential (and human potential in general), for research, policy and assessment
- Research partners : PISA Creative Thinking methods group, Claire Wladis (CUNY, NSF Grant), Martin Storme & Pinar Çelik (IESEG/CNRS, France), Baptiste Barbot (UC Louvain, Belgium)

Occurrence of “Item Response Theory” in Publications on Psychological Assessment with Tests (De Boeck, 2018)



Barriers

(Borsboom, 2006; De Boeck, 2018)

We are rarely interested in individual items…

Classical test theory methods are more accessible…

In simple situations, sum scores are strongly correlated with IRT scores…

Unawareness of real-life assessment problems that modern test theory can solve

→

Today's presentation !

Barriers
(Borsboom, 2006; De Boeck, 2018)

We are rarely interested in individual items…

Classical test theory methods are more accessible…

In simple situations, sum scores are strongly correlated with IRT scores…

Unawareness of real-life assessment problems that modern test theory can solve

Today's presentation !

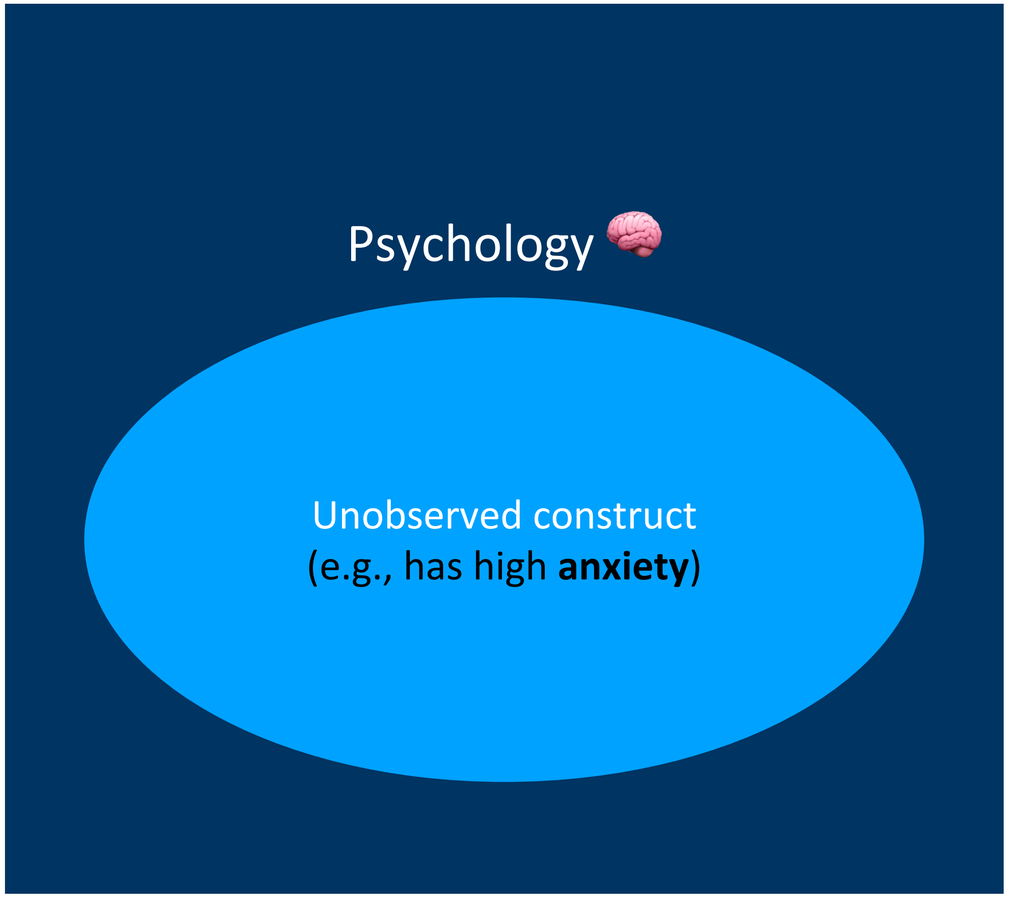
Today's presentation

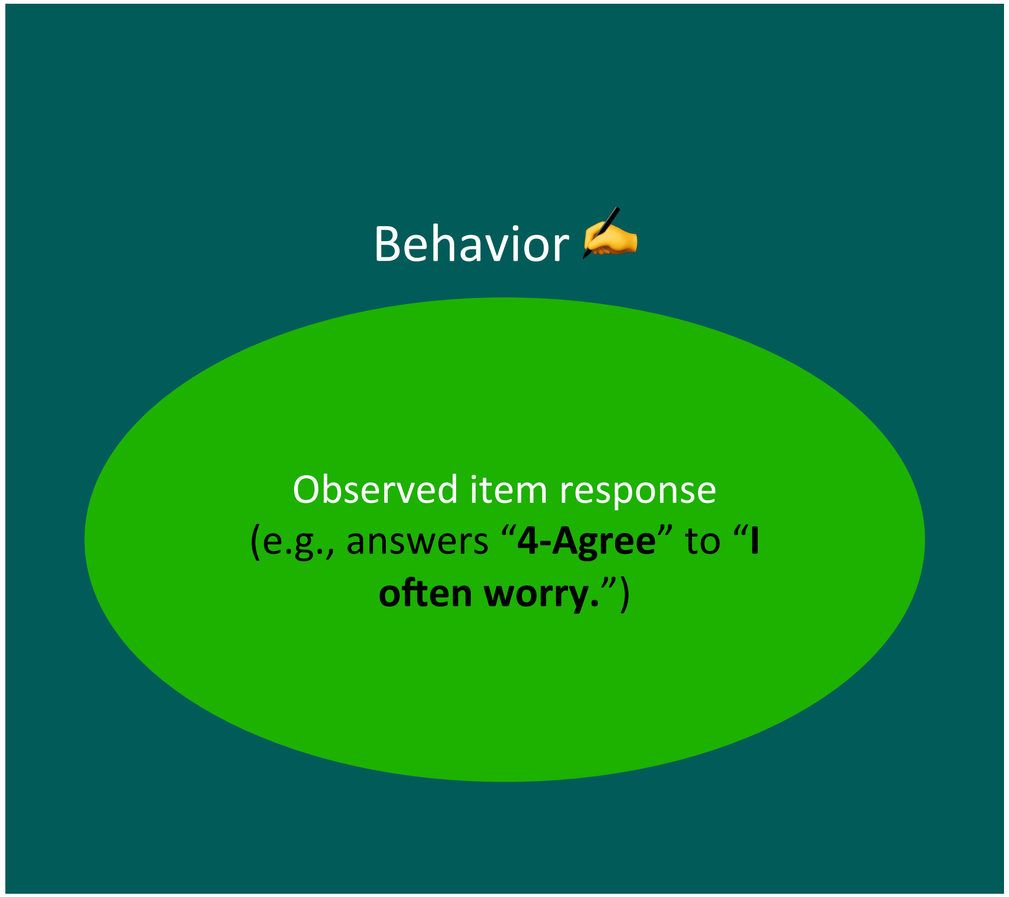
Solving real-world challenges in psychological measurement

1. A primer on modern test theory: More than a set of rules
2. Accommodating non-standard formats: Generalized models
3. Boosting reliability with distractors: Nested logit models
4. Disentangling response processes: IRTree models
5. Going beyond answers: Joint hierarchical models
6. Conclusions and upcoming plans

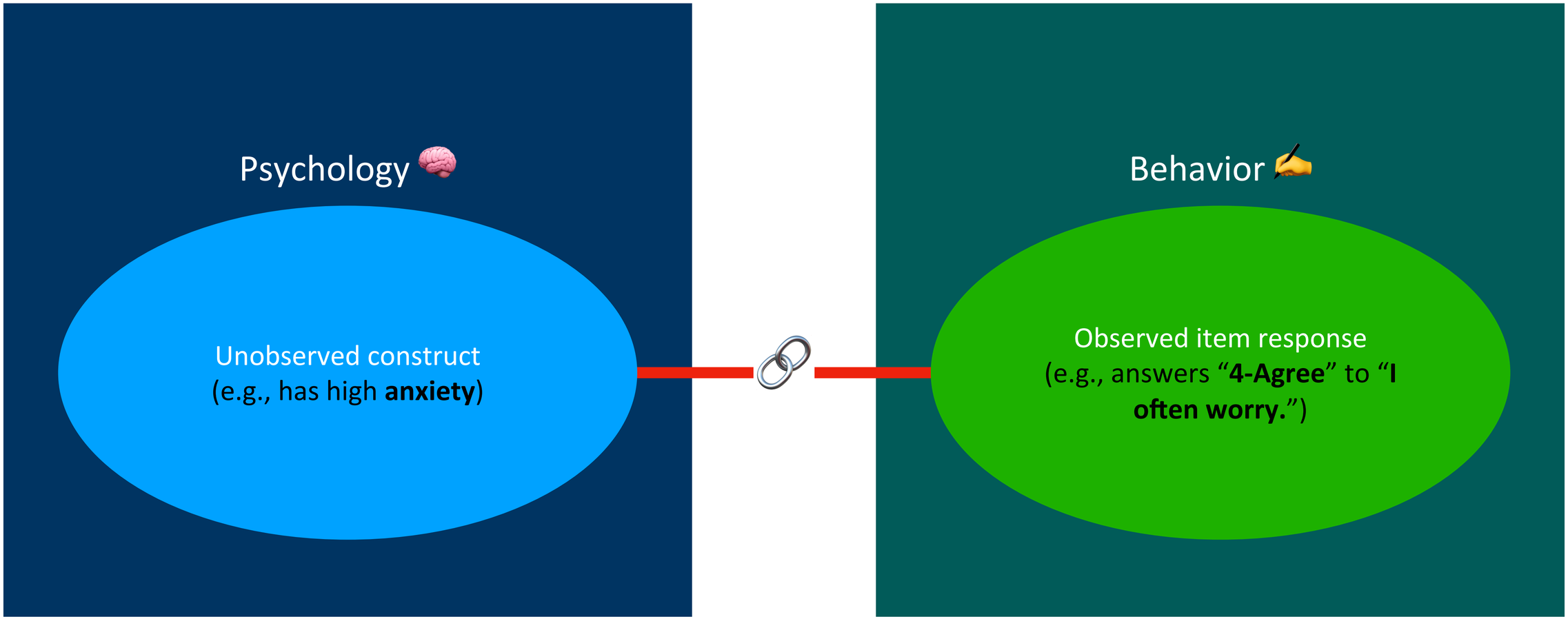
A primer on modern test theory: More than a set of rules

What do test theories do ?

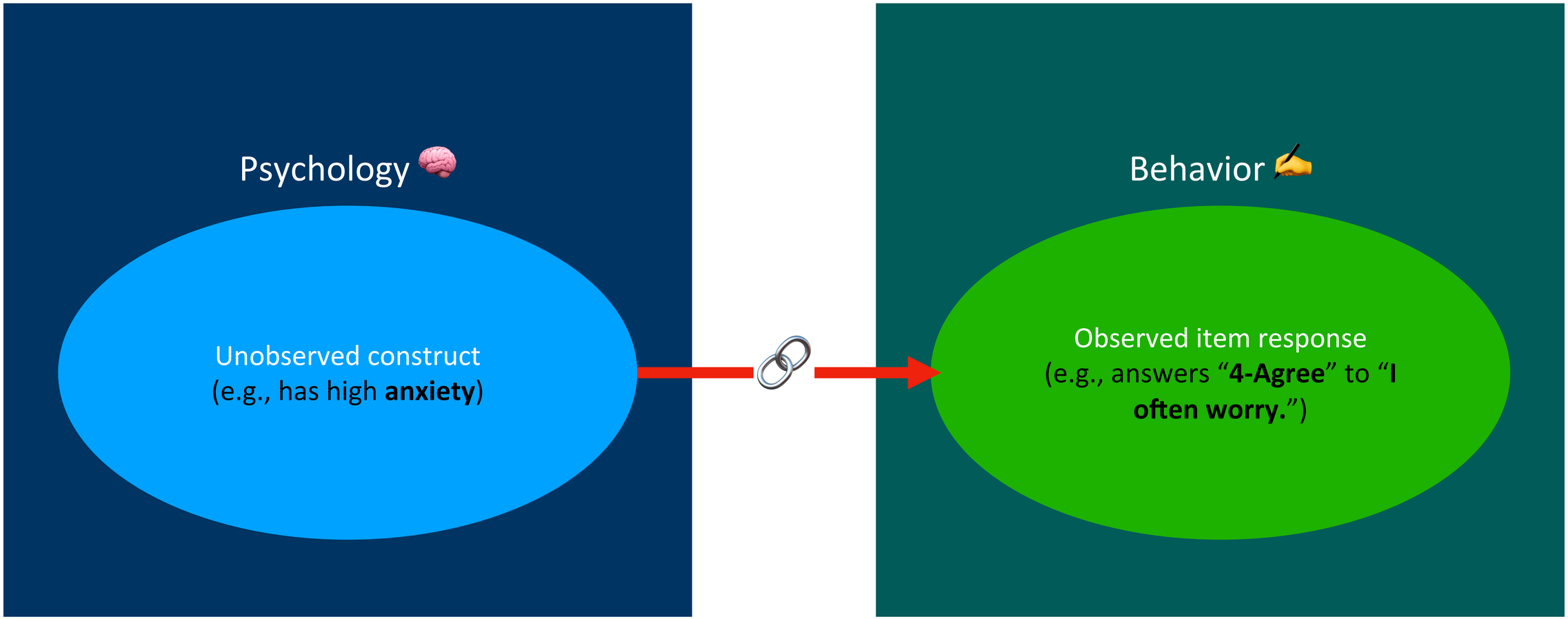




What do test theories do ?



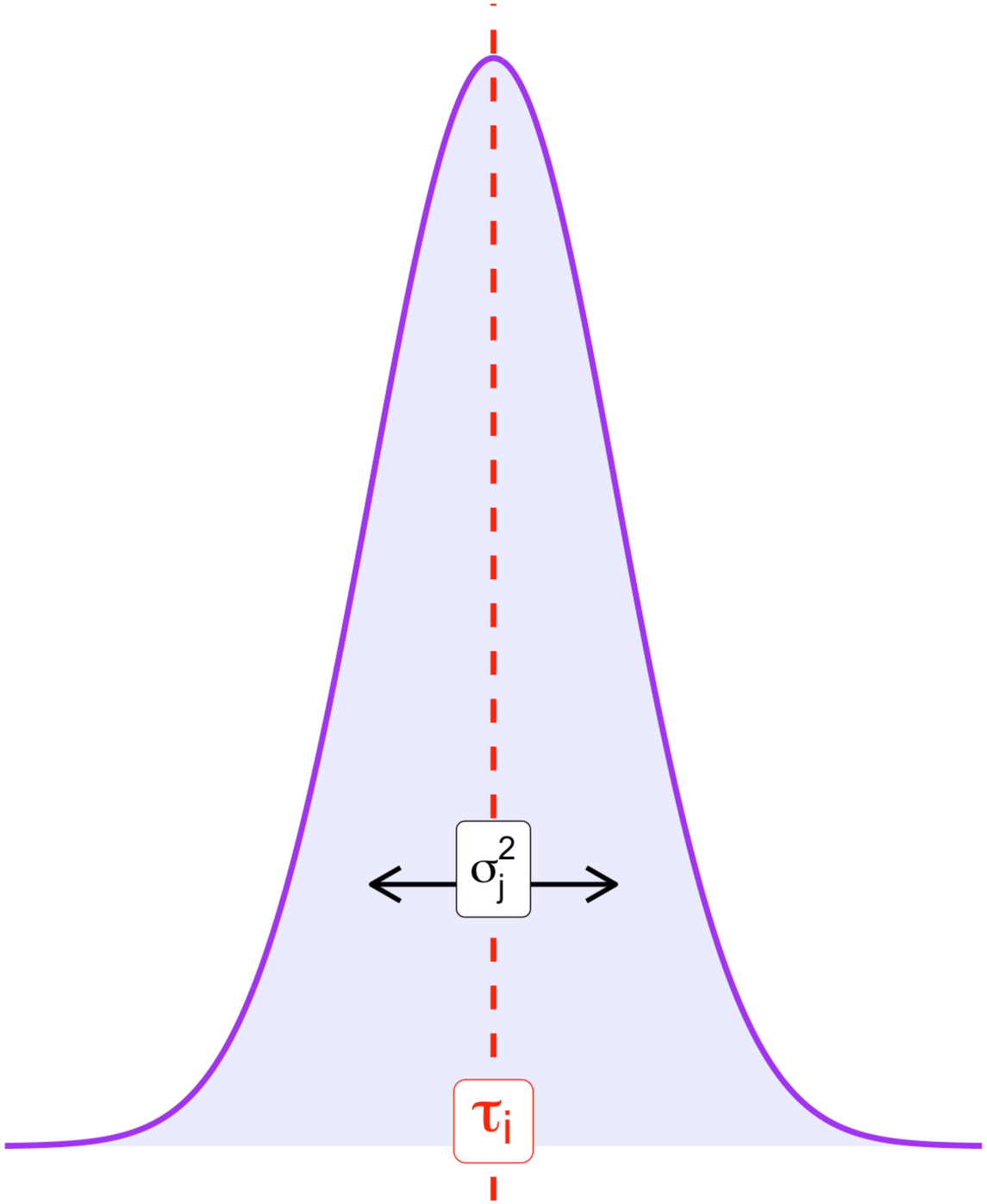
What do reflective test theories do ?



Classical Test Theory (CTT)

An **axiomatic** approach to psychometrics

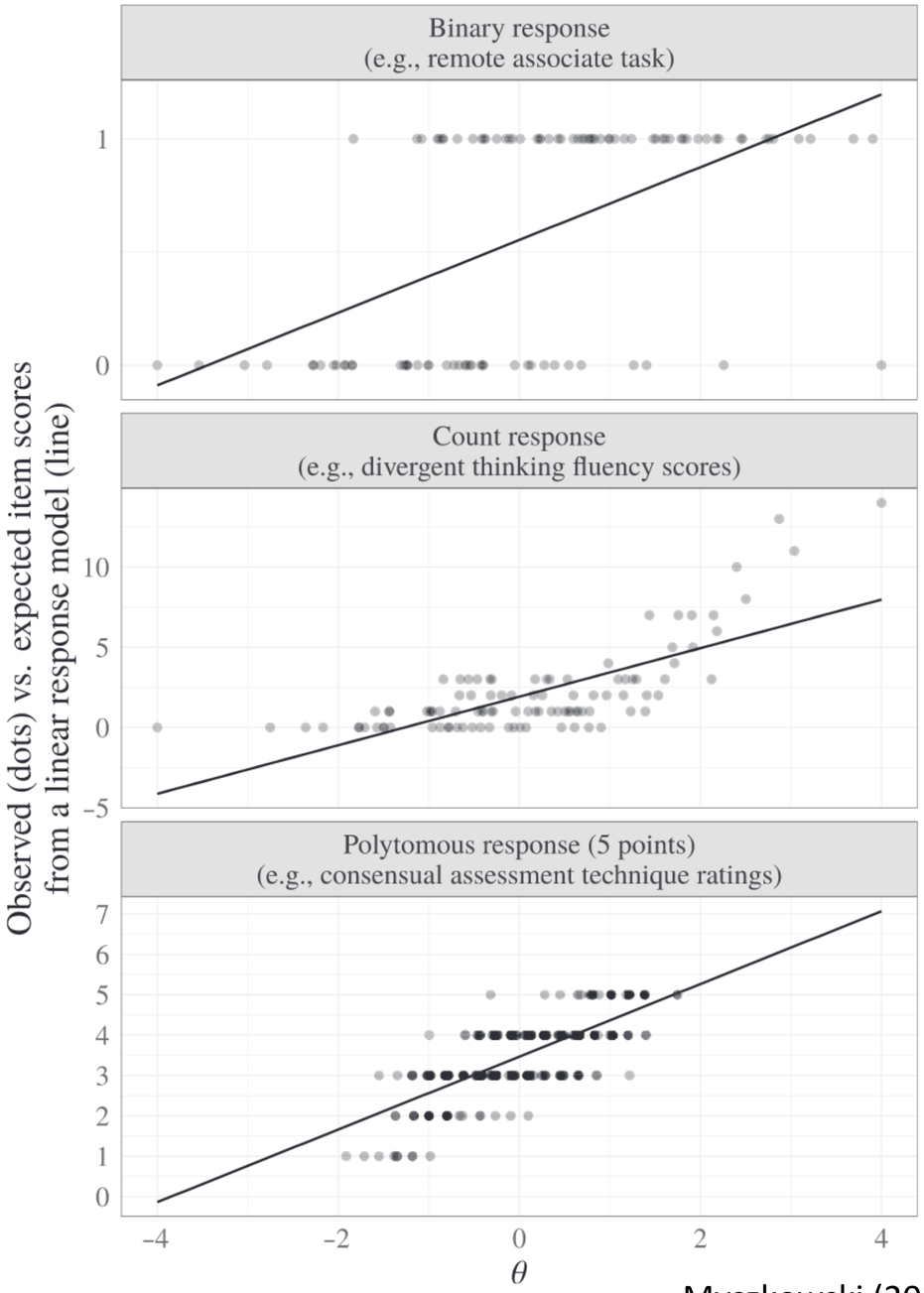
- $Item\ score = True\ score + Error$
- For person i and item j (tau-equivalent CTT):
  - $X_{ij} = \tau_i + \varepsilon_{ij}$, with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$
- The expectation of observed scores is the true score.
- We can obtain person locations (“true scores”) by **calculation** (averaging/summing over items).



Some problems

Linear/Gaussian issues

- Assumes:
  - A **linear** relation between the attribute and expected item scores
  - **Normally** distributed item scores with fixed variance
- Not always reasonable !



Myszkowski (2024)

What about modern response theory?

Mplus

HOME ORDER CONTACT US LOGIN MPLUS DISCUSSION

Item Response Theory (IRT)

Mplus offers IRT analyses using 1PL, 2PL, 3PL, 4PL, partial credit, and generalized partial credit models. Due to the general modeling framework of Mplus, the IRT modeling includes unique features that combine multidimensional analysis; two-level, three-level, and cross-classified

---

SAS THE POWER TO KNOW.

Home Support Learn Connect

Go to Documentation Home

SAS/STAT(R) 13.1 User's Guide

PDF

Contents Topics About

- The CLUSTER Procedure
- The CORRESP Procedure
- The DISCRIM Procedure
- The DISTANCE Procedure
- The FACTOR Procedure
- The FASTCLUS Procedure
- The FMM Procedure
- The FREQ Procedure
- The GAM Procedure
- The GENMOD Procedure
- The GLIMMIX Procedure

Search this docu

Previous Page | Next Page

The IRT Procedure (Experimental)

Overview Getting Started Syntax Details Examples References

Basic Features

The IRT procedure enables you to estimate various item response theory models. The following list summarizes some of the bas

uses the Rasch model; one-, two-, three-, and four-parameter models; and graded response model with logistic or probit link

---

STATA

Home / Products / Features / IRT

IRT (item response theory)

Order

Explore the relationship between unobserved latent characteristics such as mathematical aptitude and the probability of correctly answering test questions (items). Or explore the relationship between unobserved health and self-reported responses to questions about mobility, independence, and other health-affected activities. IRT can be used to create measures of such unobserved traits or place individuals on a scale measuring the trait. It can also be used to select the best items for measuring a latent trait. IRT models are available for binary, graded, rated, partial-credit, and nominal response items. Visualize the relationships using item characteristic curves, and measure overall test performance using test information functions. And much more.

Learn about IRT (item response theory).

Binary response models

- One-parameter logistic (1PL)
- Two-parameter logistic (2PL)
- Three-parameter logistic (3PL)

Watch One-parameter logistic (1PL) models.
Watch Two-parameter logistic (2PL) models.
Watch Three-parameter logistic (3PL) models

Ordinal response models

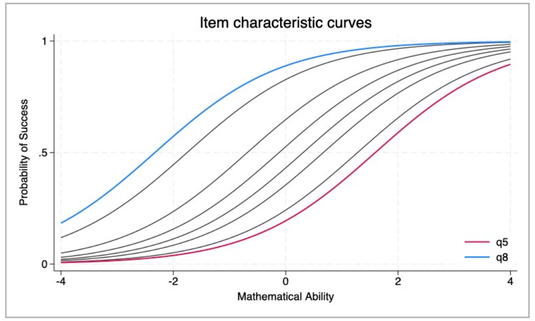
- Graded response
- Partial credit
- Generalized partial credit
- Rating scale

Watch Graded response (GRM) models.
Watch Rating scale (RSM) models.

Categorical response model

- Nominal response

Watch Nominal response (NRM) models.



Control panel interface

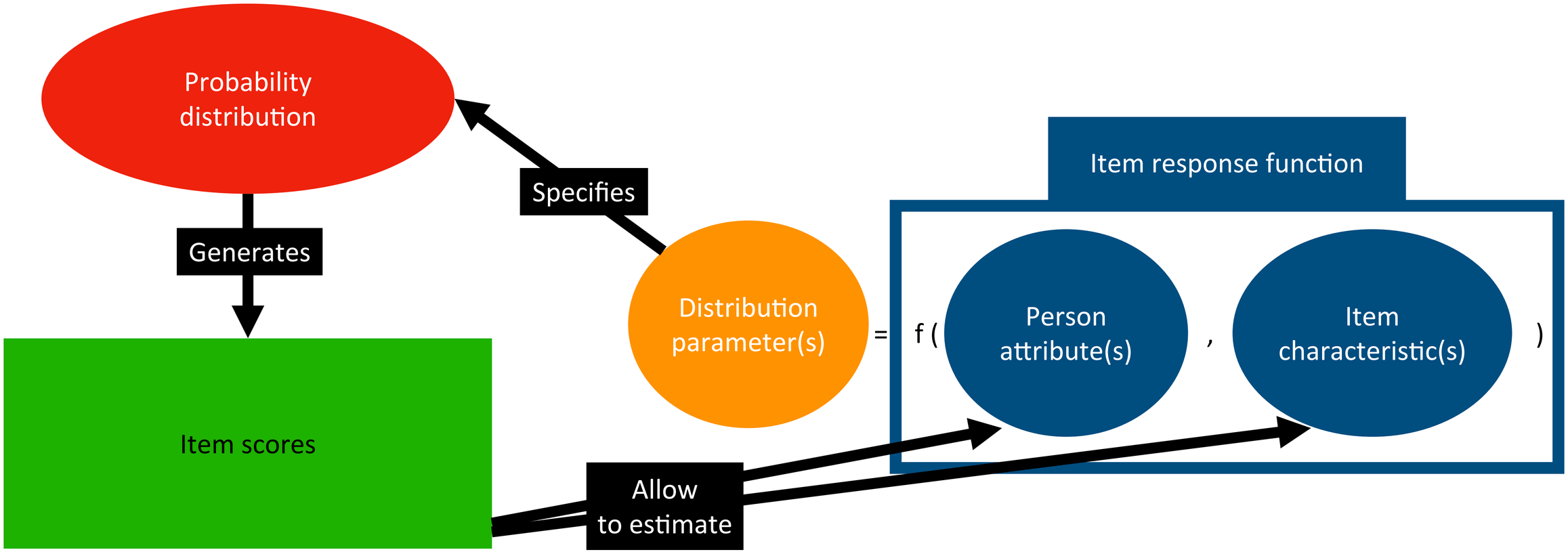
- Access all IRT features
- Easily select response type and item variables
- Even create hybrid models
- Estimate models
- Select and customize graphs
- Manage reporting of results

Modern test theory / item response theory (IRT)

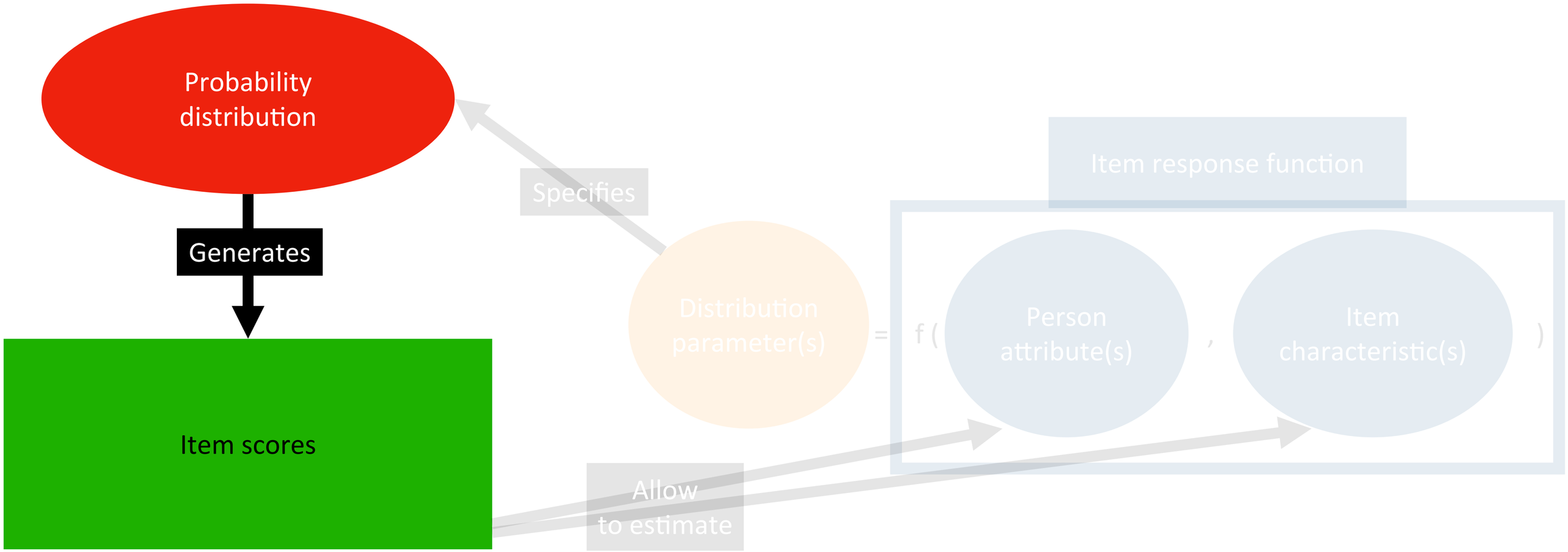
A **probabilistic** approach to psychometrics

- IRT “is like the theory of statistical estimation. IRT uses latent characterizations of individuals and items as predictors of observed responses.” (De Ayala, 2022, p.5).
- **Conceptual shift**: from rationalizing scoring procedures to proposing a probabilistic explanation for item scores.
- Requires **statistical estimation** to obtain person locations.

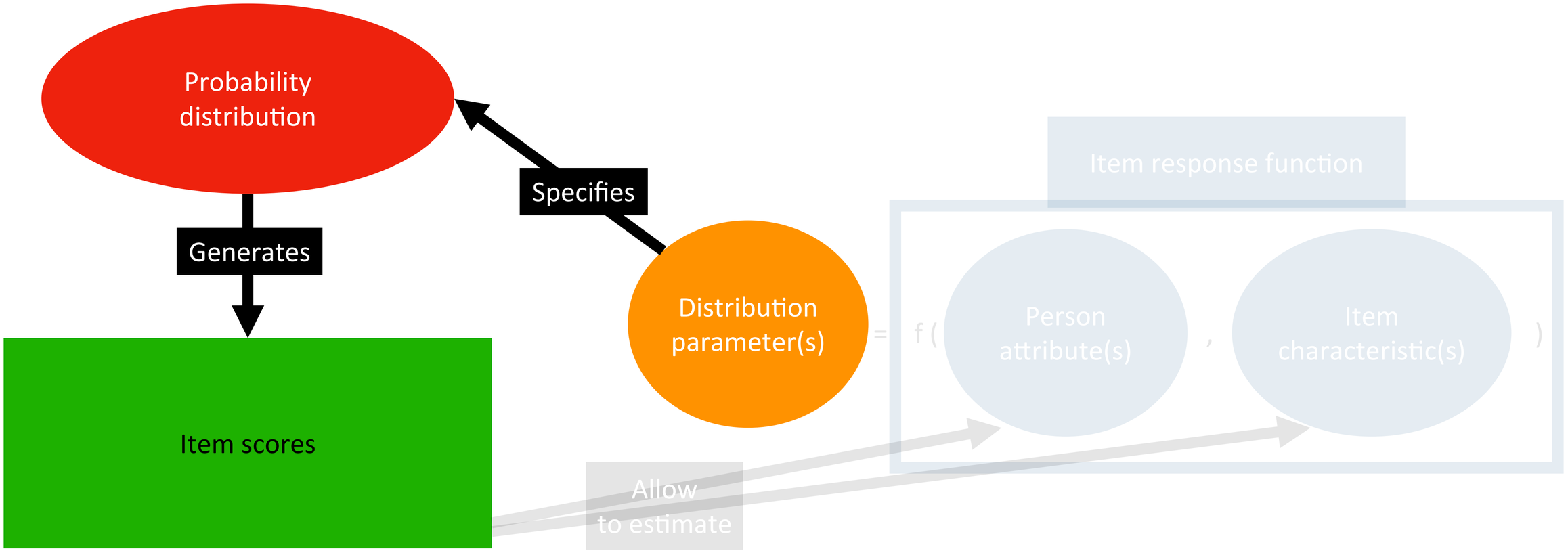
Anatomy of an IRT model
(in general)



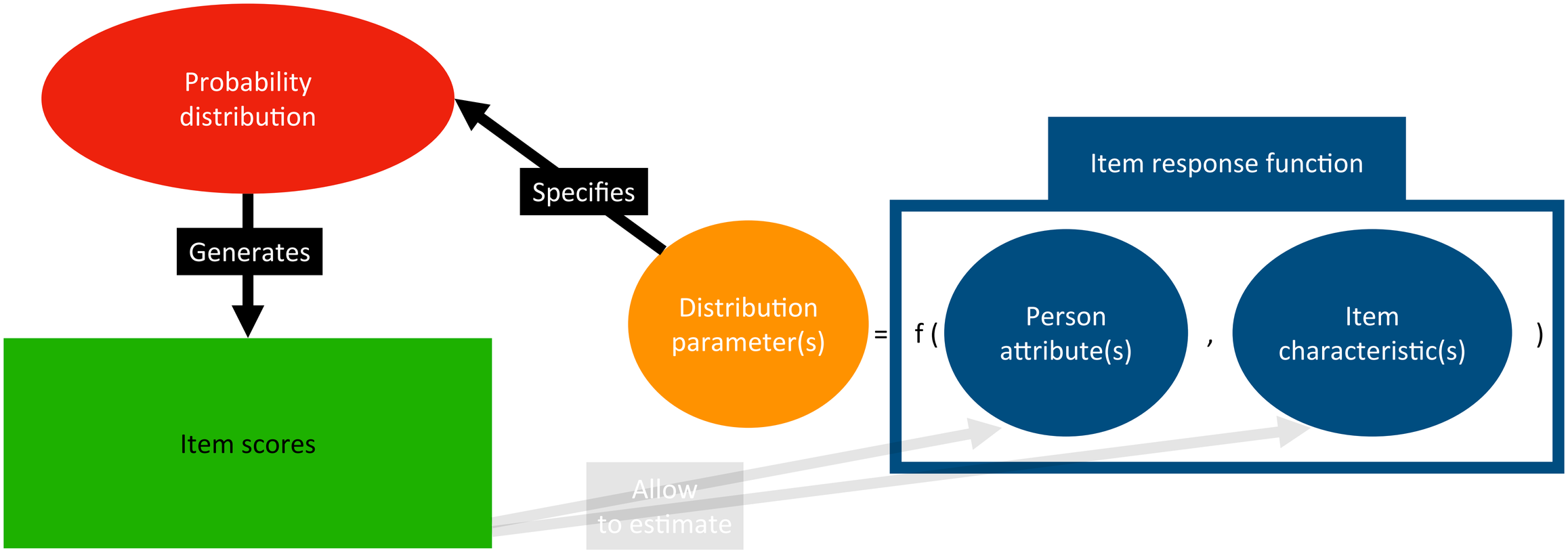
Anatomy of an IRT model
(in general)



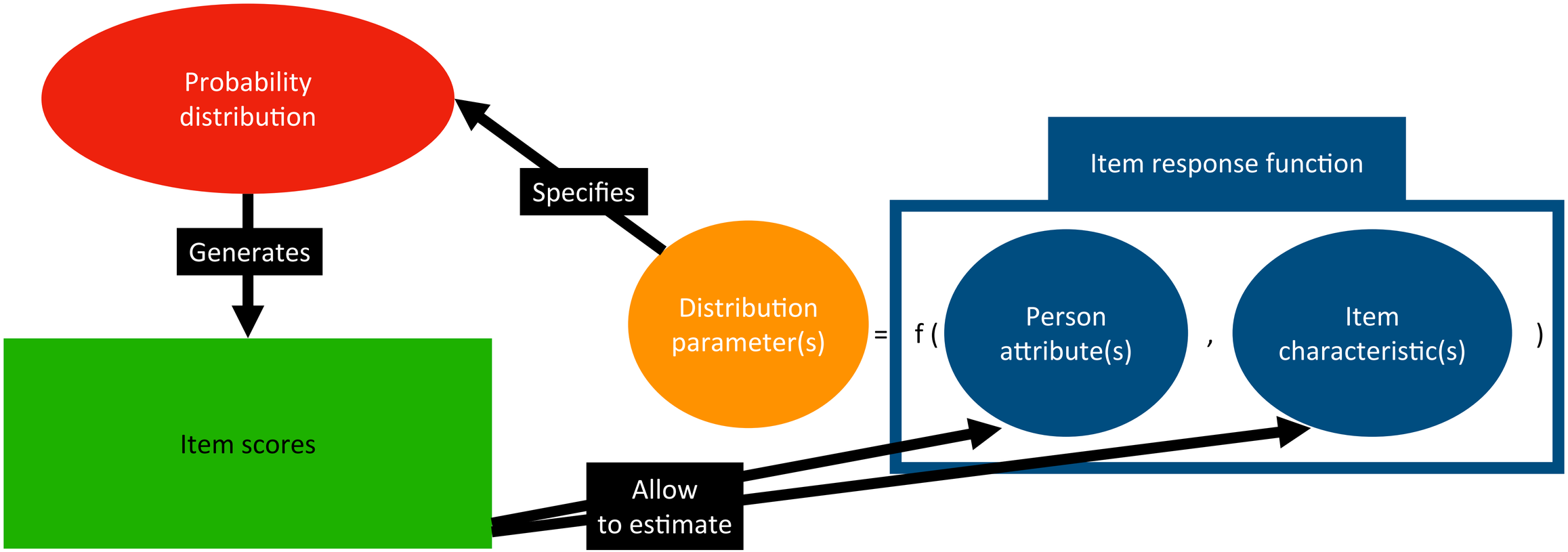
Anatomy of an IRT model
(in general)



Anatomy of an IRT model
(in general)

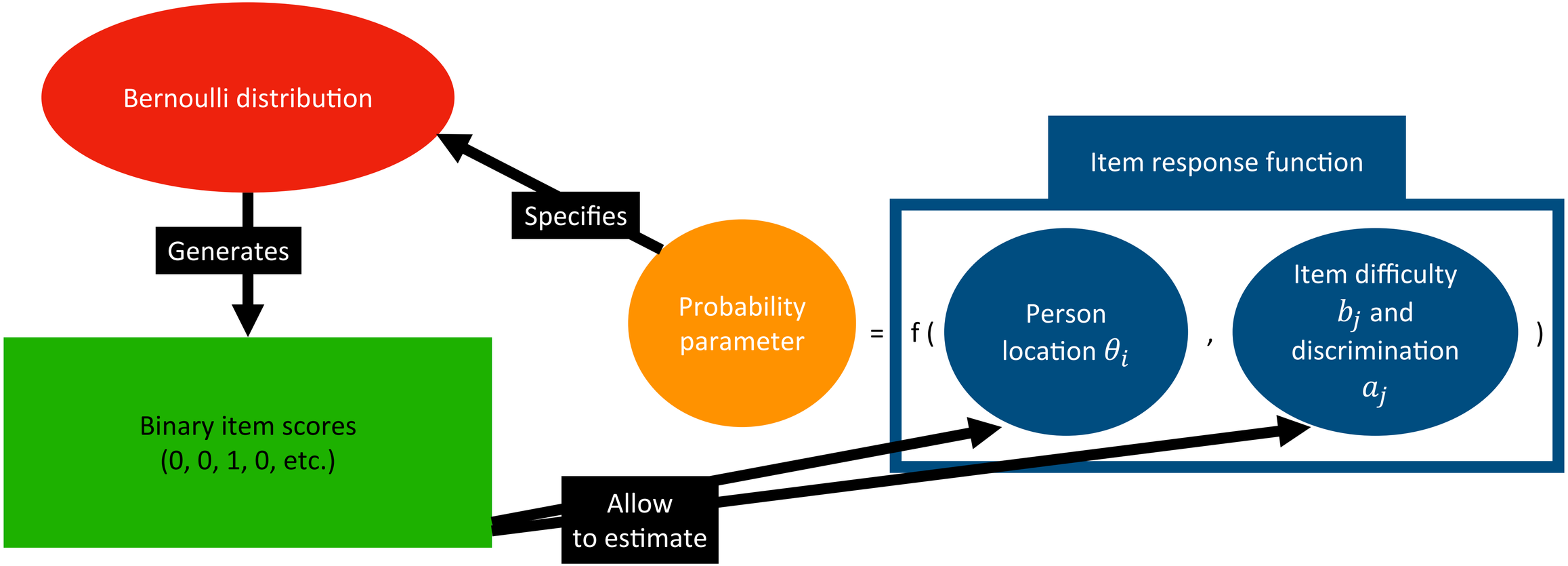


Anatomy of an IRT model
(in general)



Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$

Bernoulli distribution

Generates

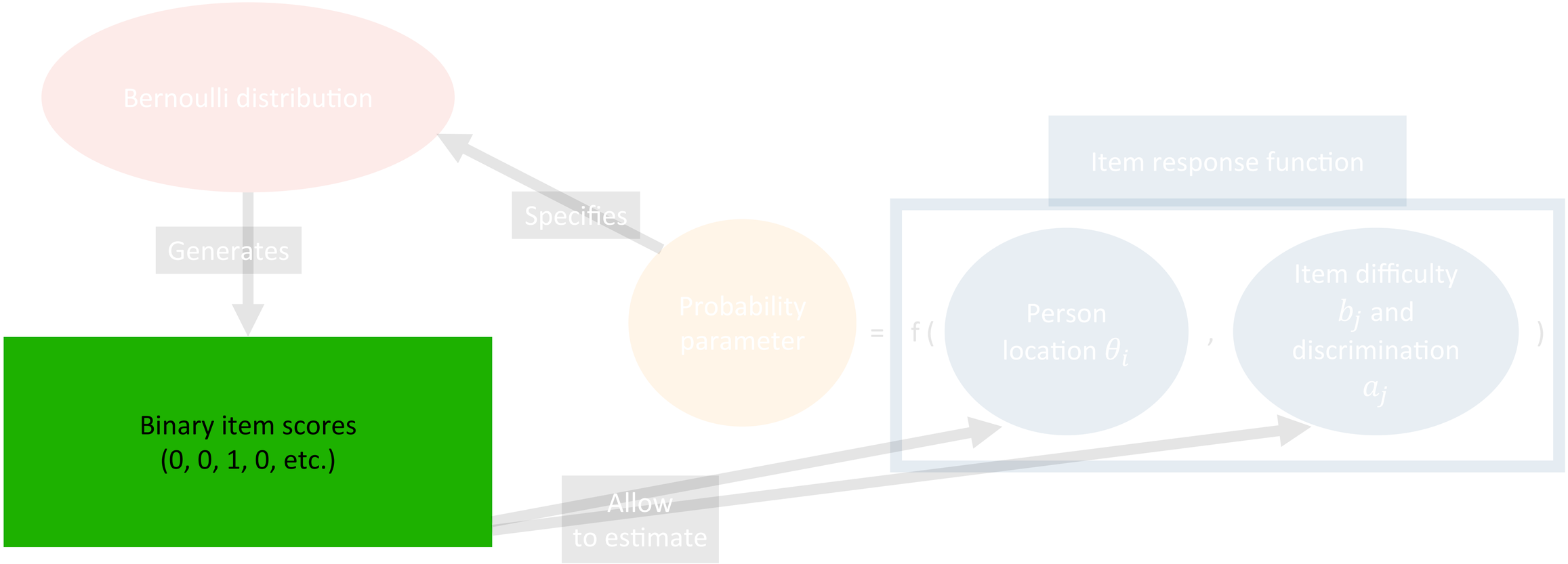
Binary item scores
(0, 0, 1, 0, etc.)

	Item.1	Item.2	Item.3	Item.4	Item.5
1	1	1	1	0	1
2	1	1	1	0	0
3	1	1	1	1	1
4	0	1	0	1	1
5	1	1	1	0	1
6	1	1	1	1	1
7	1	1	1	1	1
8	1	0	1	0	1
9	1	0	1	0	1
10	1	1	1	1	1

m difficulty
b_j and
crimination
a_j

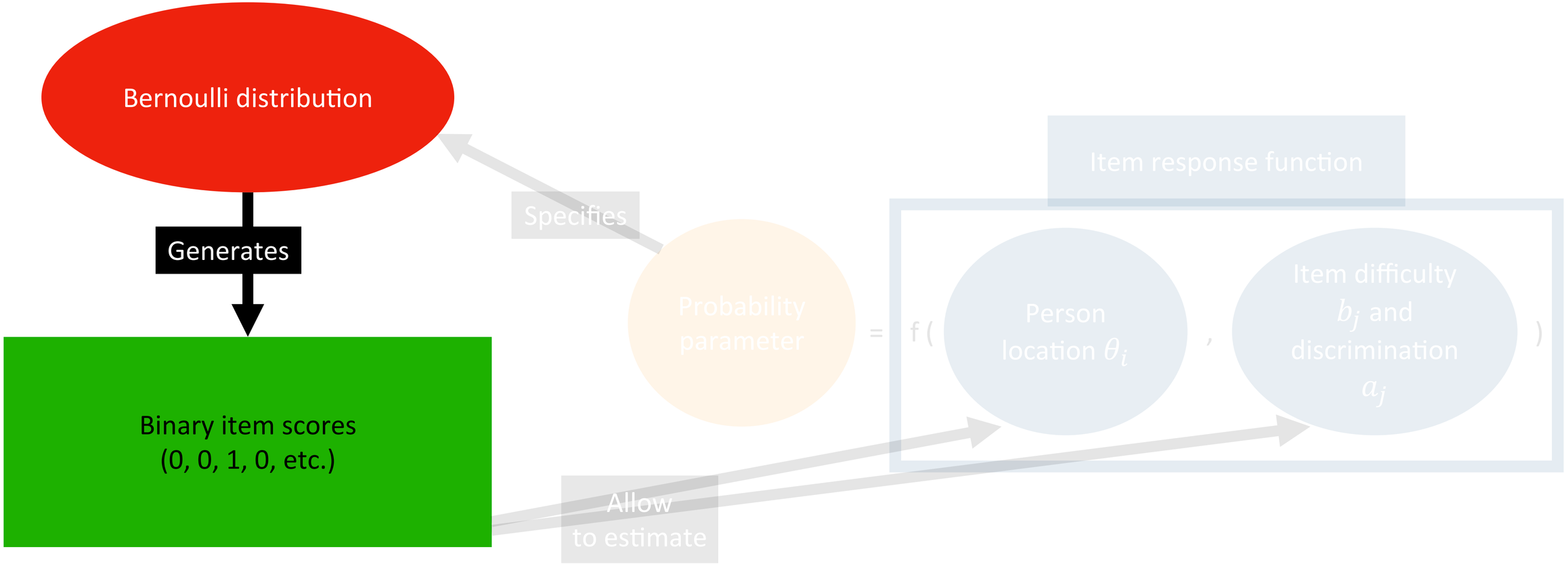
Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$

Bernoulli distribution

Generates

Binary item scores
(0, 0, 1, 0, etc.)

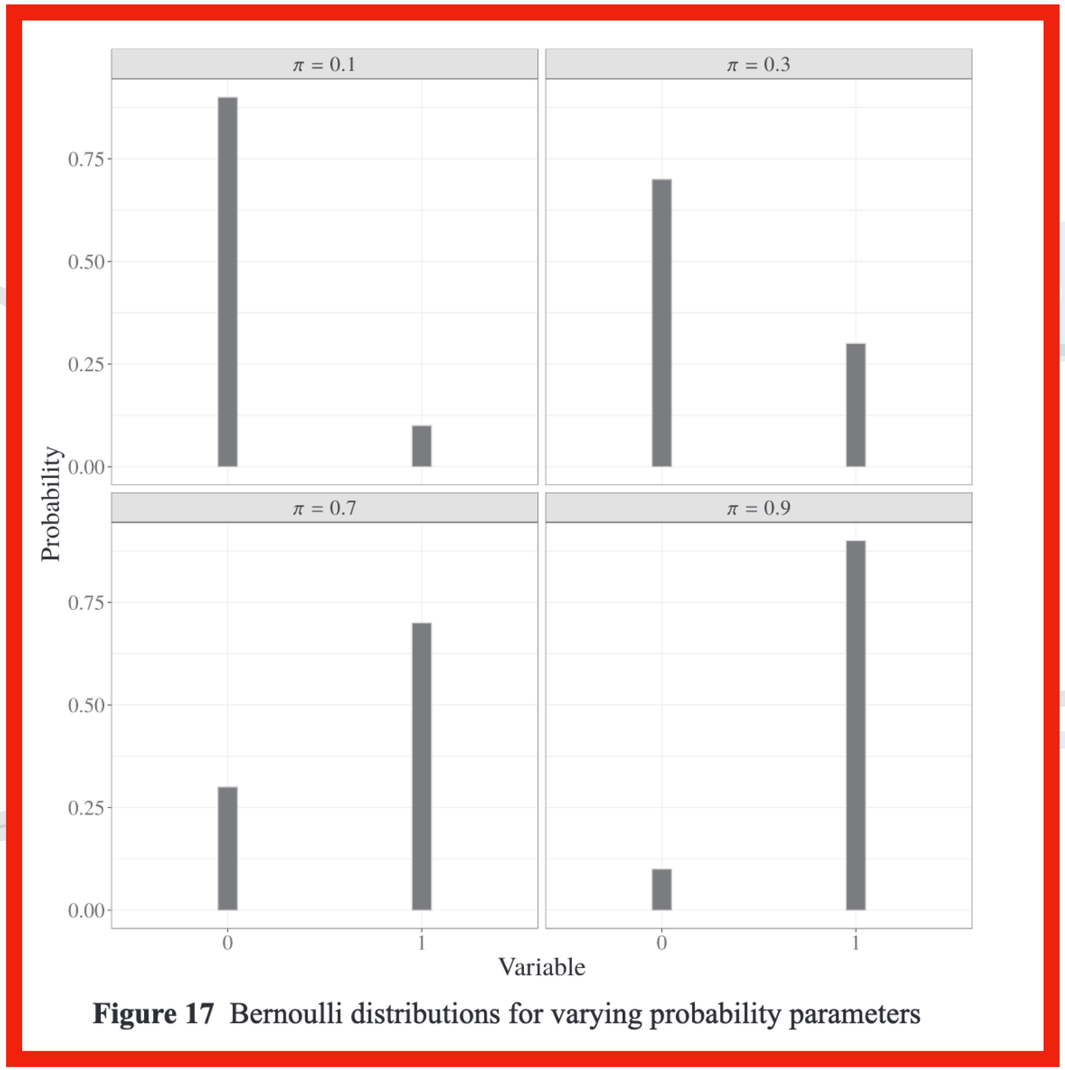


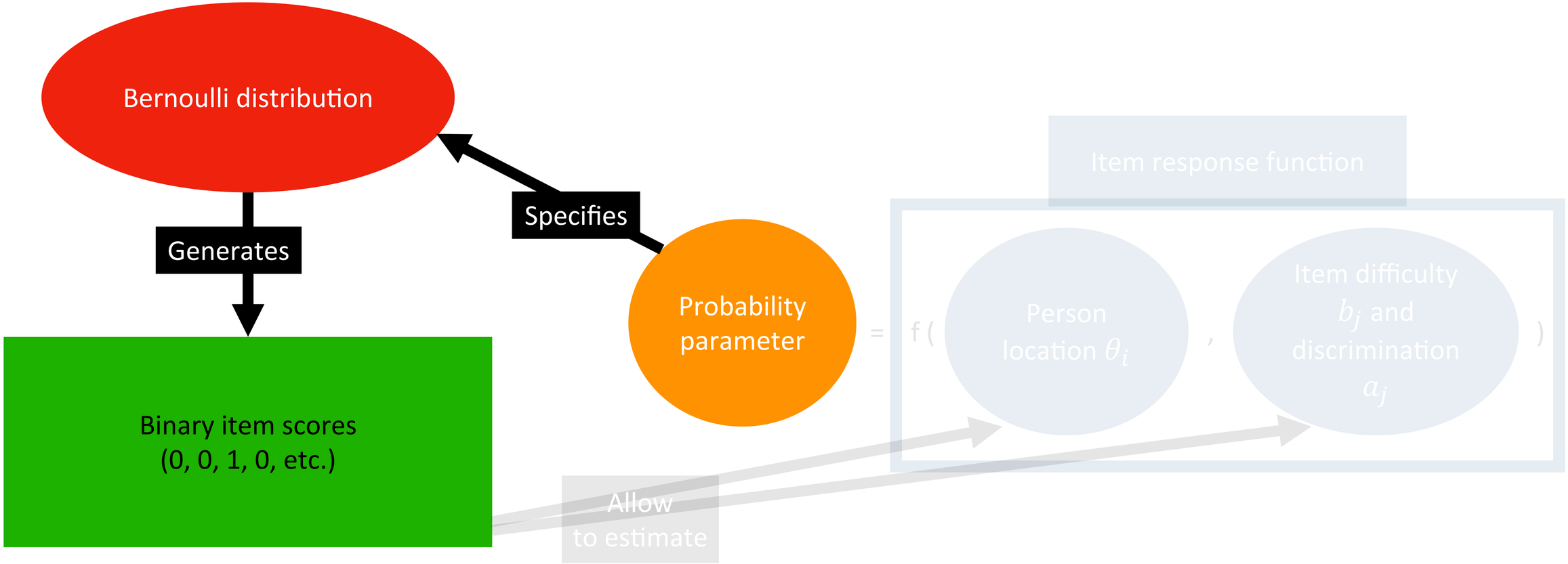
Figure 17 Bernoulli distributions for varying probability parameters

nse function

Item difficulty b_j and discrimination a_j

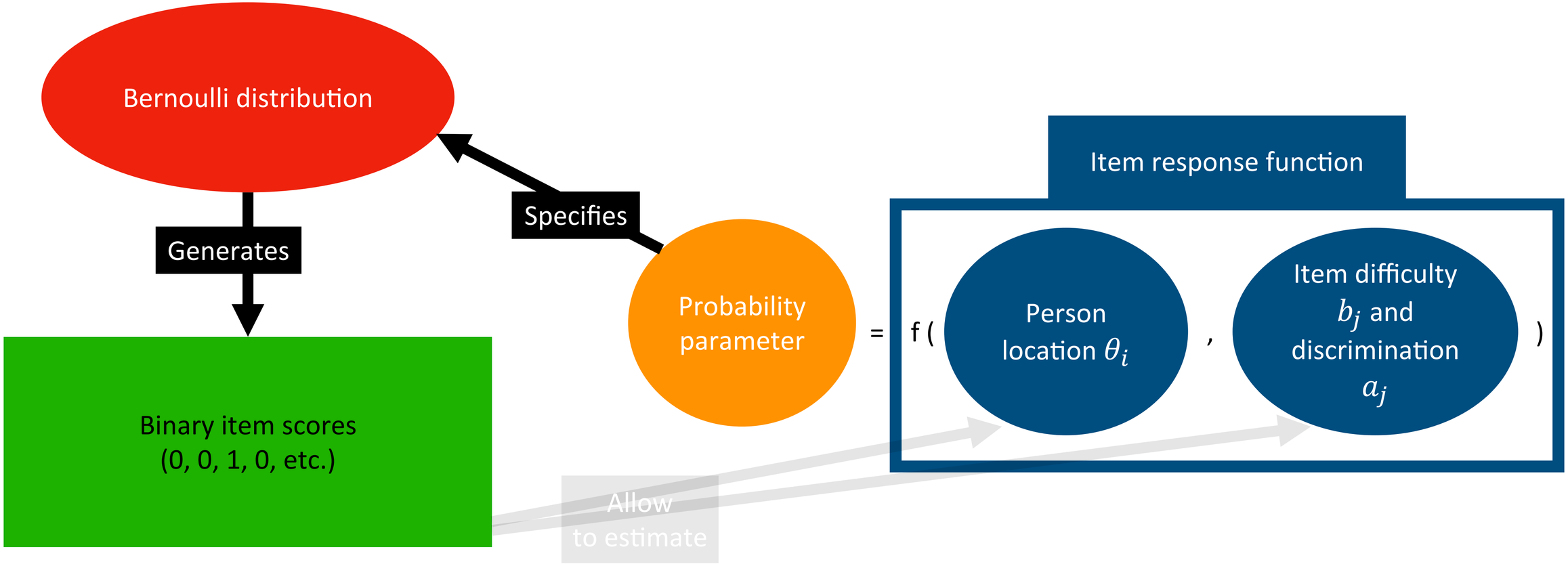
Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$

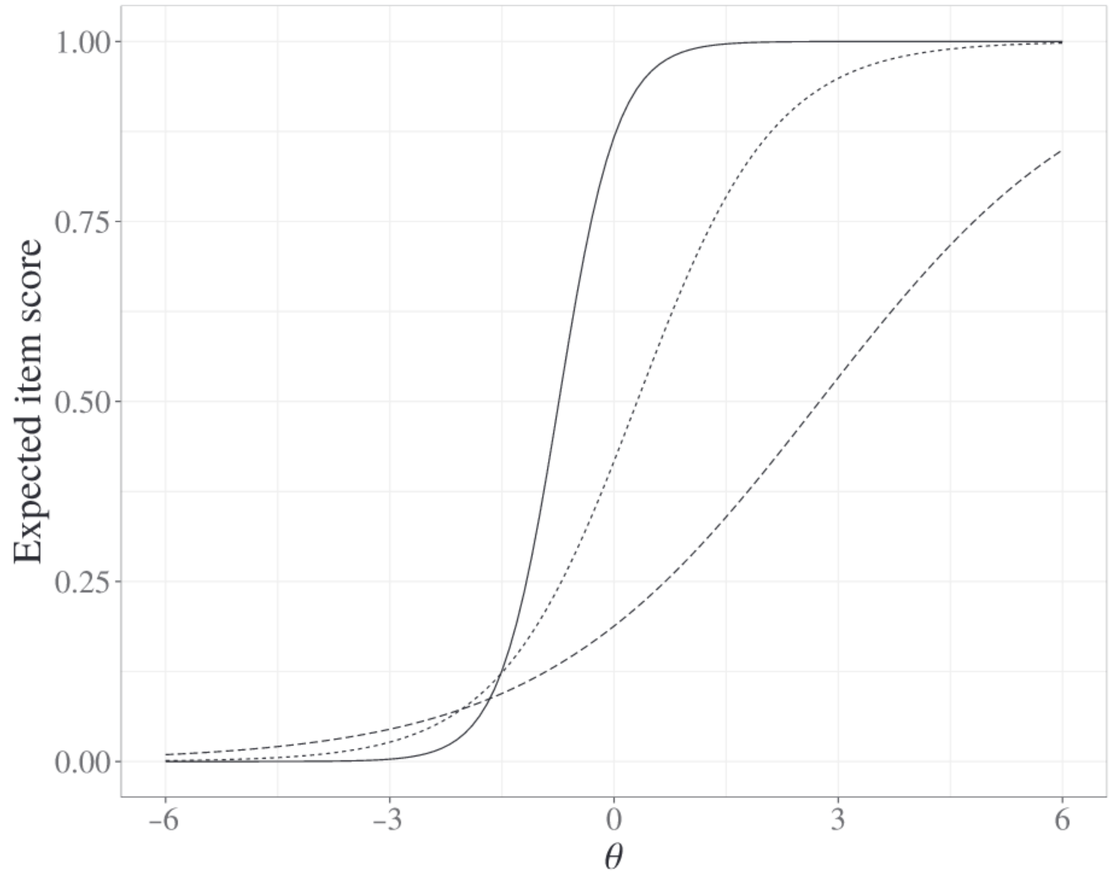


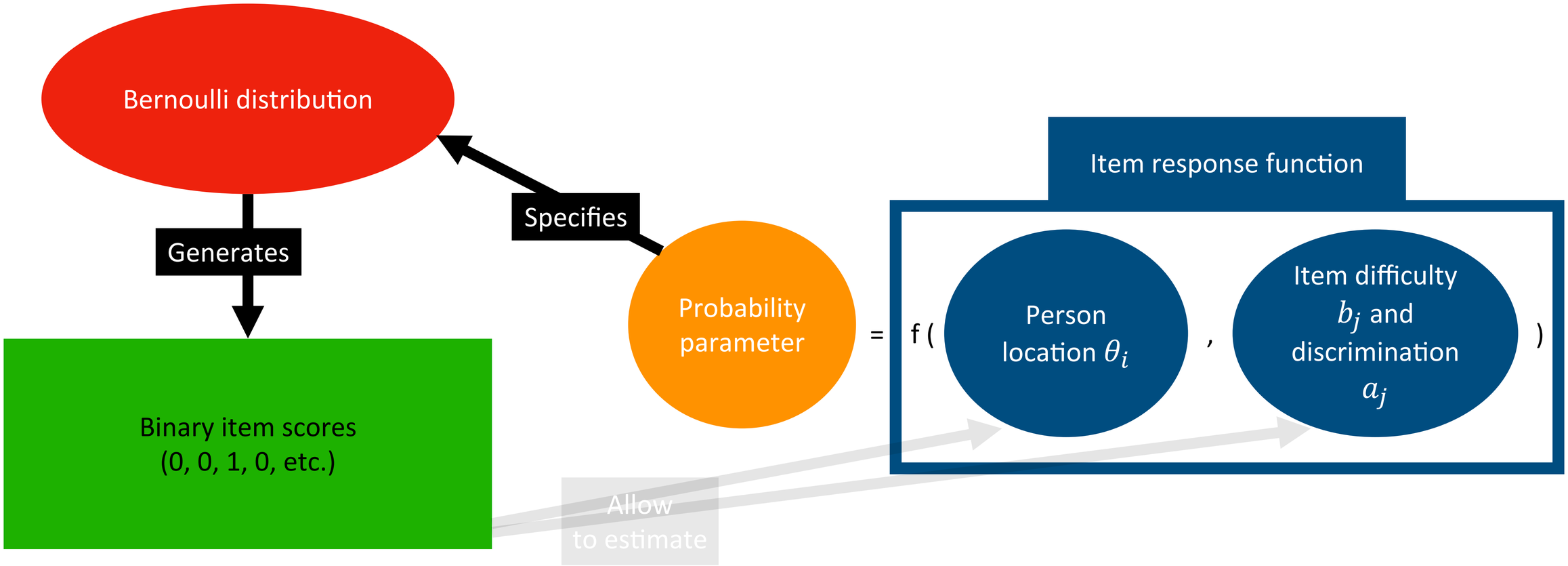
Figure 12 Expected scores of three items modeled with a 2-parameter logistic model

Item response function

f (Person location θ_i , Item difficulty b_j and discrimination a_j)

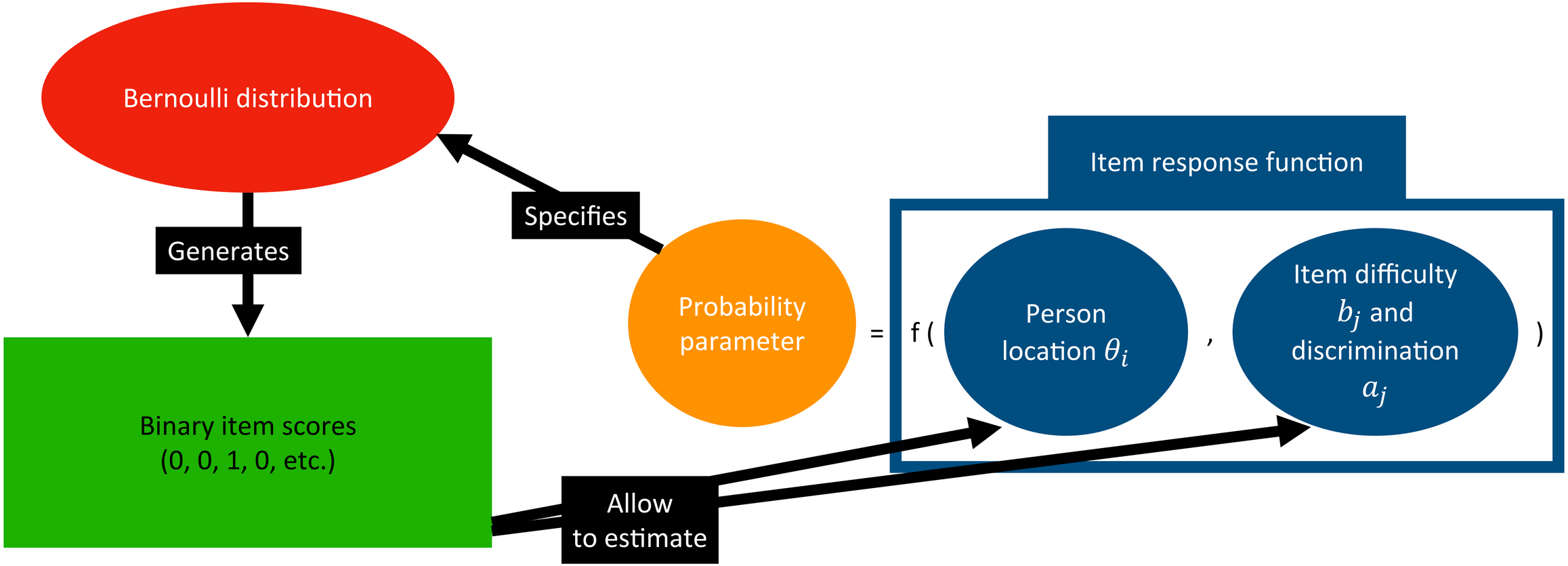
Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



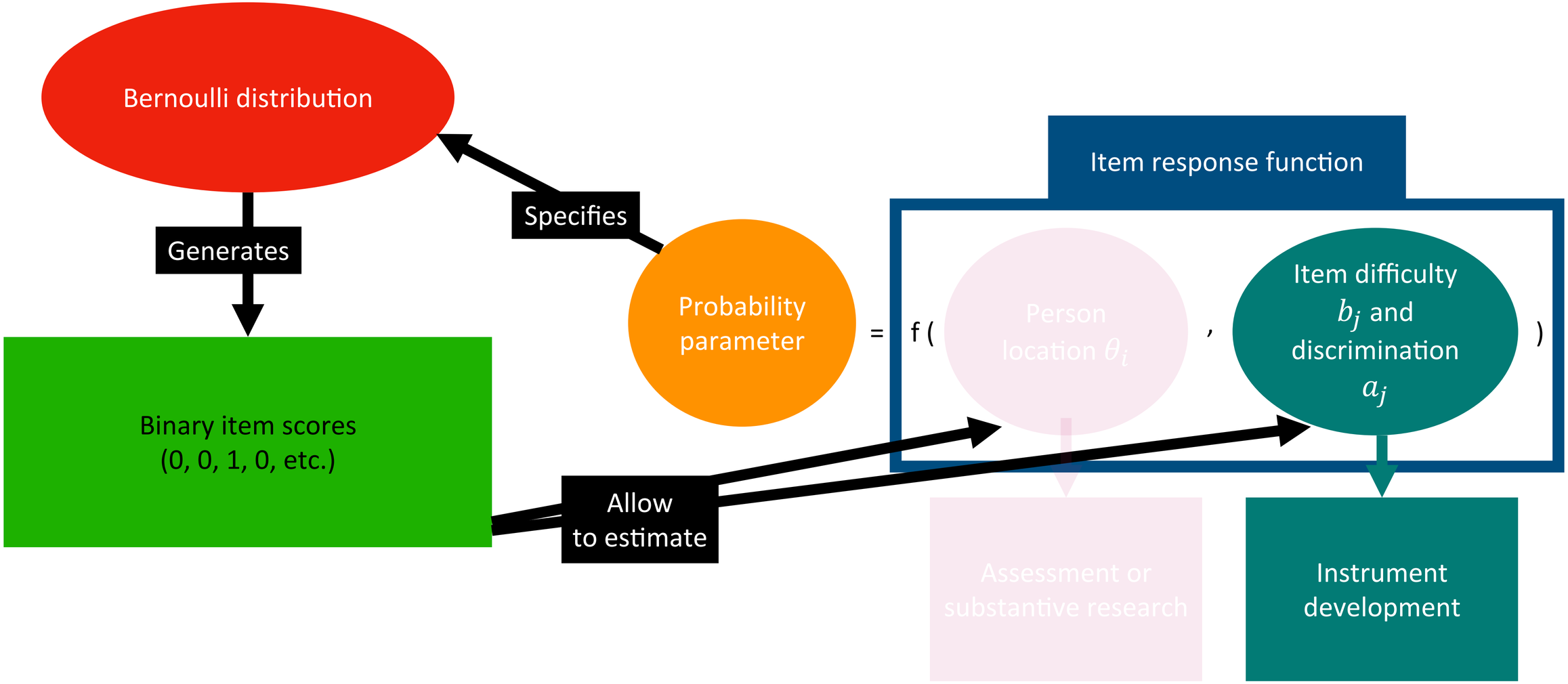
Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



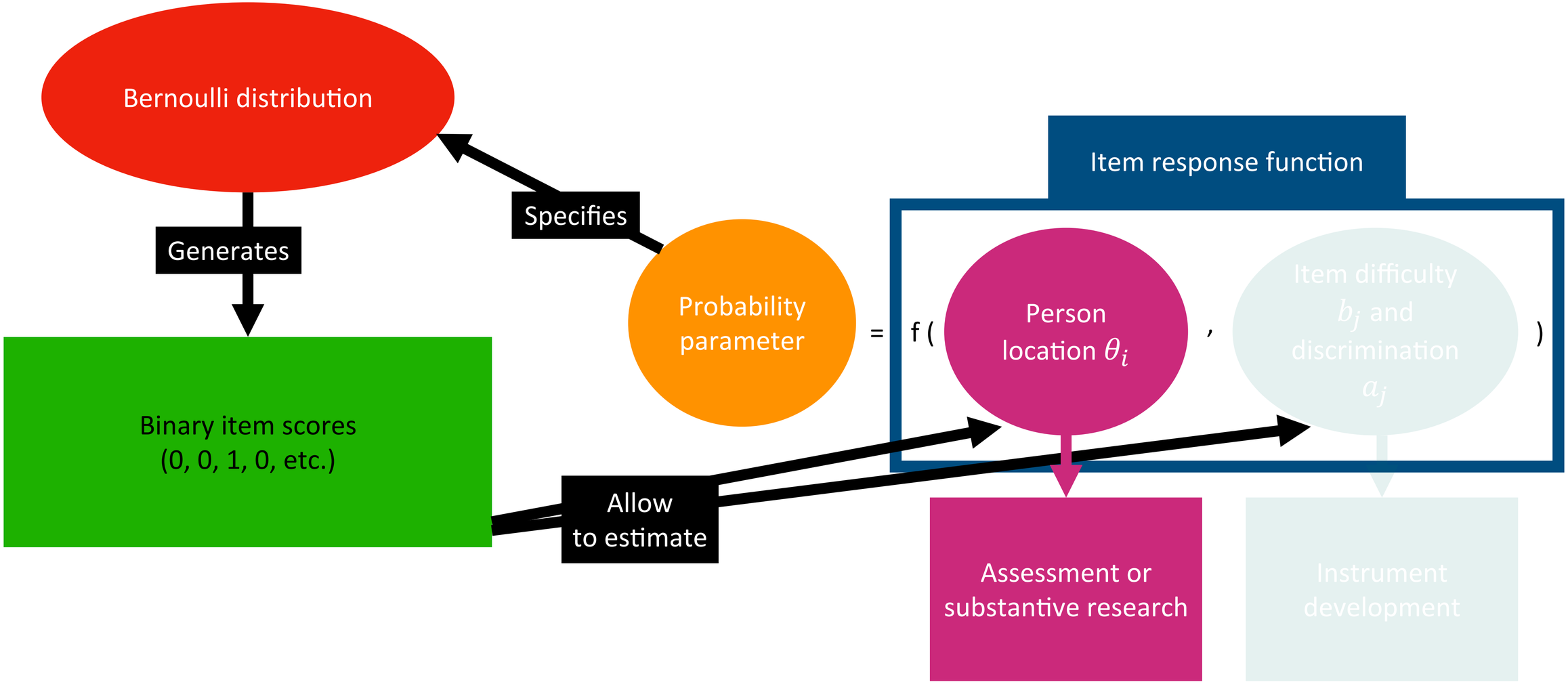
Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



Anatomy of an IRT model

Example : 2-parameter logistic $X_{ij} \sim Bernoulli\left(logit^{-1}(a_j\theta_i + b_j)\right)$



Notes and generalities about IRT
(useful for later)

- Person attributes are noted θ.
- θ is assumed to be **drawn from a standard normal ("z") distribution** (in general).
- **Reliability/standard errors** of measurement are studied as a function of θ (and of the items taken by that person, if variable).
- Since IRT works at the item response level:
  - It handles **missing data** by design → allows matrix sampling, computer-adaptive testing, etc.
  - It allows **different response formats** within a test.

Accommodating non-standard formats: Generalized models

Modern test theory as an approach

Table 1
Overview of Item Response Models Subsumed Under Generalized Linear Item Response Theory

Item format	Item response distribution	Transformation $[g(\tau_{ij})]$	Specification of Formula 1	Model
Dichotomous	Binomial	Logit: $\ln\{\tau_{ij}/(1-\tau_{ij})\}$	$b_j + a_j t_i$	Birnbaum's two-parameter model
			$b_j + a t_i$	Rasch's one-parameter model
			$b + a t_i$	Binomial error model
			b_j for $t_i = 0$; $b_j + a_j$ for $t_i = 1$	Macready and Dayton's (1980) state model for mastery testing
			$b_j + a_j t_i + c_{1j} t_{1i} + \cdots + c_{pj} t_{pi}$	Reckase's (1985) multidimensional model with $(p+1)$ latent traits
		Cumulative normal (probit): $\Phi^{-1}(\tau_{ij})$	$b_j + a_j t_i$	Lord's normal ogive model
		Complementary log–log: $\ln\{-\ln(1-\tau_{ij})\}$	$b_j + a t_i$	Goldstein's (1980) model
Continuous	Normal with homogeneous item response variance σ_j^2	Identity: τ_{ij}	$b_j + a_j t_i$	Jöreskog's (1971) model for congeneric item responses
			$b + a t_i$, $\sigma_j^2 = \sigma^2$	Lord and Novick's (1968) model for parallel item responses
			$b + a t_i$	Lord and Novick's (1968) model for tau-equivalent item responses
			$b_j + a t_i$	Lord and Novick's (1968) model for essentially tau-equivalent item responses
			$b_j + a_j t_i + c_{1j} t_i^2 + \cdots + c_{pj} t_i^{p+1}$	McDonald's (1982) polynomial model
		τ_{ikl}	$b_{kl} + a_{kl} t_i + c_{1kl} t_{ki} + c_{2kl} t_{li} + c_{3kl} t_{kli}$	Mellenbergh, Kelderman, Stijlen, and Zondag's (1979) model for items from the combination of the kth element of the first facet and the lth element of the second facet of a facet design
Ungraded options	Multinomial	Logit of kth category and fixed mth category: $\ln(\tau_{ijk}/\tau_{ijm})$	$b_{jk} + a_{jk} t_i$	Bock's (1972) model for nominal options
Graded options	Multinomial	Cumulative logit of kth category: $\ln\{(\tau_{ijk} + \cdots + \tau_{ijm})/(\tau_{ij1} + \cdots + \tau_{ijk-1})\}$	$b_{jk} + a_j t_i$	Samejima's (1969) model for cumulative proportions
		Adjacent kth and $(k-1)$th category logit: $\ln(\tau_{ijk}/\tau_{ijk-1})$	$b_{jk} + a t_i$	Masters's (1982) partial credit model
			$\delta_i + \gamma_k + a t_i$	Andrich's (1978) rating scale model with location (δ_i) and threshold (γ_k) parameters
Response time	Exponential	Inverse: $1/\tau_{ij}$	$b_j + a t_i$	Rasch's (1960) model for reading speed

Mellenberg (1994)

Modern test theory as an approach

Table 1 Overview of item response models available for different item types

Item type	Example distribution	Example link	Example inverse link	Example models
Continuous unbounded	Gaussian	Identity	Identity	Congeneric, parallel
Binary	Bernoulli	Logit	Logistic	1PL, 2PL, 3PL
Count (discrete with lower bound)	Poisson	Logarithm	Exponential	RPCM 2PPCM
Response time (continuous with lower bound)	Log-normal	Logarithm	Exponential	LNIRT models
Visual analog scale (continuous with lower and upper bound)	Beta	Logit	Logistic	BRM-1, BRM-2

Note: 1PL: 1-parameter logistic; 2PL: 2-parameter logistic; 3PL: 3-parameter logistic; RPCM: Rasch Poisson counts model; 2PPCM: 2-parameter Poisson counts model; LNIRT: log-normal item response theory; BRM-1: beta response model 1; BRM-2: beta response model 2.

Myszkowski (2024)

Modern test theory as an approach

Table 1 Overview of item response models available for different item types

Item type	Example distribution	Example link	Example inverse link	Example models
Continuous unbounded	Gaussian	Identity	Identity	Congeneric, parallel
Binary	Bernoulli	Logit		
Count (discrete with lower bound)	Poisson	Logarithm	E	
Response time (continuous with lower bound)	Log-normal	Logarithm	Exponential	LNIRT models
Visual analog scale (continuous with lower and upper bound)	Beta	Logit	Logistic	BRM-1, BRM-2

Classical test theory models

Note: 1PL: 1-parameter logistic; 2PL: 2-parameter logistic; 3PL: 3-parameter logistic; RPCM: Rasch Poisson counts model; 2PPCM: 2-parameter Poisson counts model; LNIRT: log-normal item response theory; BRM-1: beta response model 1; BRM-2: beta response model 2.

Myszkowski (2024)

Modern test theory as an approach

Table 1 Overview of item response models available for different item types

Item type	Example distribution	Example link	Example inverse link	Example models
Continuous unbounded	Gaussian	Identity	Identity	Congeneric, parallel
Binary	Bernoulli	Logit	Logistic	1PL, 2PL, 3PL
Count (discrete with lower bound)	Poisson	Logarithm	Expo	
Response time (continuous with lower bound)	Log-normal	Logarithm	Exponential	LNIRT models
Visual analog scale (continuous with lower and upper bound)	Beta	Logit	Logistic	BRM-1, BRM-2

"Common" IRT models

Note: 1PL: 1-parameter logistic; 2PL: 2-parameter logistic; 3PL: 3-parameter logistic; RPCM: Rasch Poisson counts model; 2PPCM: 2-parameter Poisson counts model; LNIRT: log-normal item response theory; BRM-1: beta response model 1; BRM-2: beta response model 2.

Myszkowski (2024)

Modern test theory as an approach

Table 1 Overview of item response models available for different item types

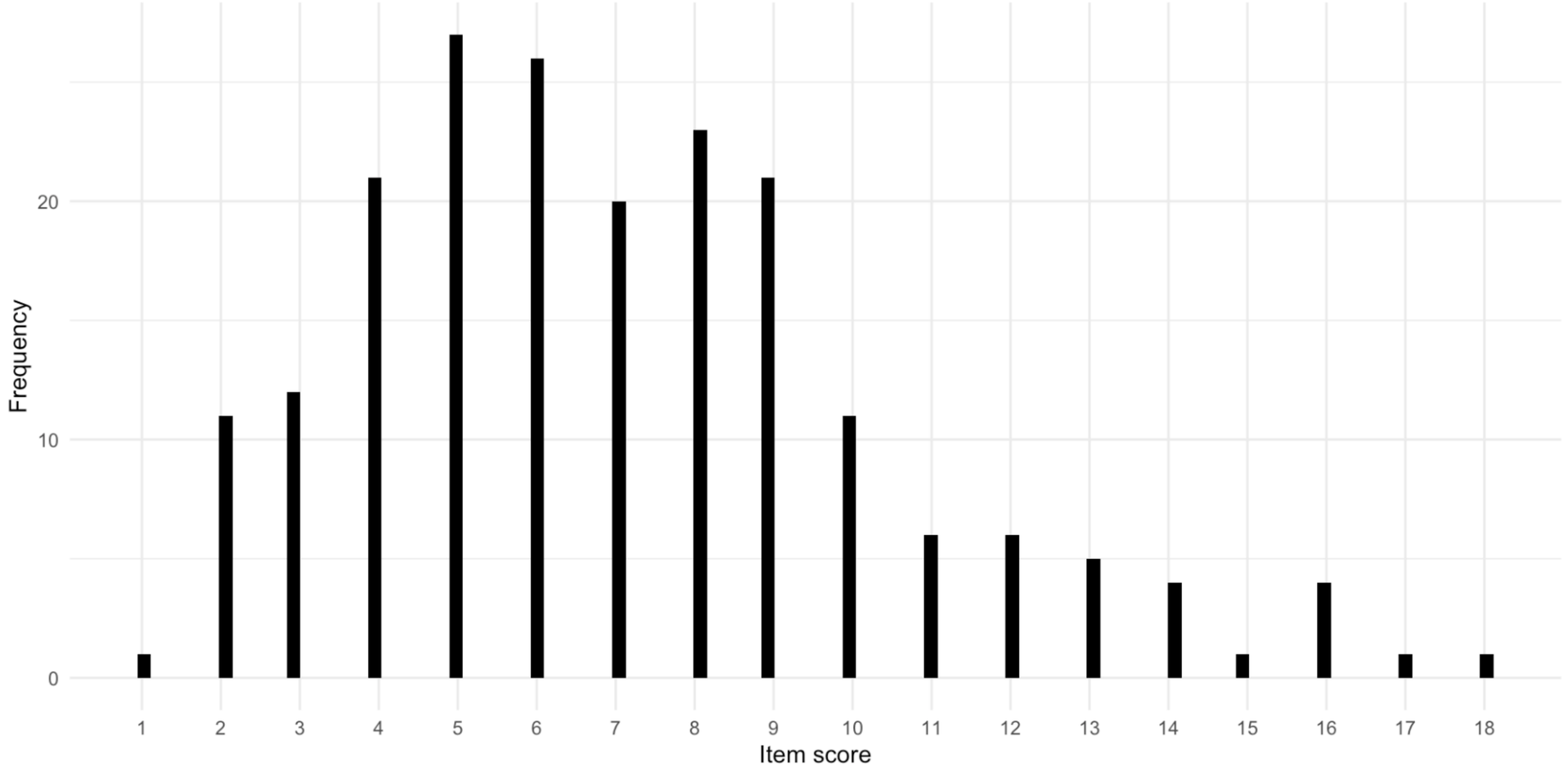
Item type	Example distribution	Example link	Example inverse link	Example models
Continuous unbounded	Gaussian	Identity	Identity	Congeneric, parallel
Binary	Bernoulli	Logit	Logistic	1PL, 2PL, 3PL
Count (discrete with lower bound)	Poisson	Logarithm	Exponential	RPCM, 2PPCM
Response time (continuous with lower bound)	Log-normal	Logarithm	Expo	…and more !
Visual analog scale (continuous with lower and upper bound)	Beta	Logit	Logistic	BRM-1, BRM-2

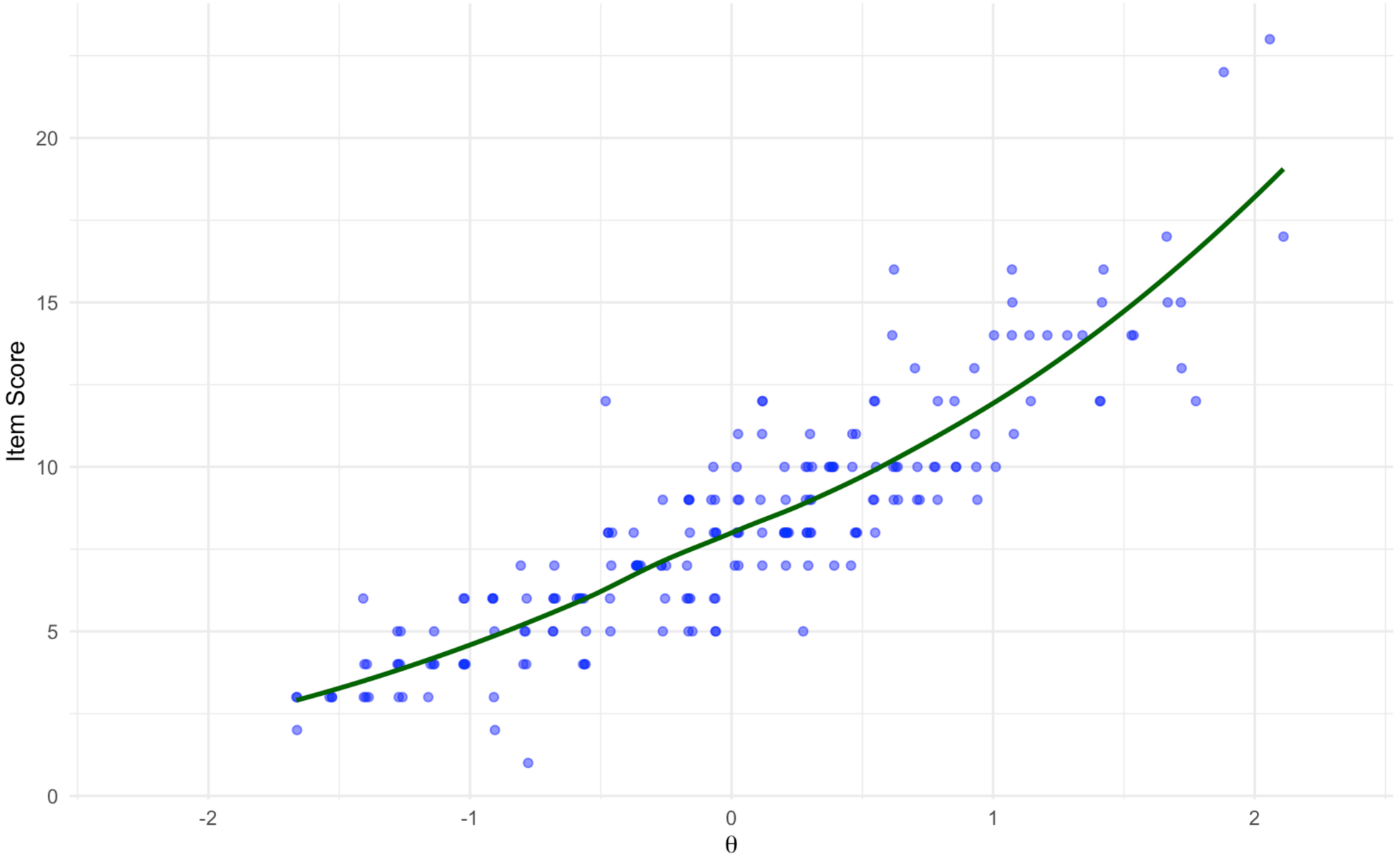
Note: 1PL: 1-parameter logistic; 2PL: 2-parameter logistic; 3PL: 3-parameter logistic; RPCM: Rasch Poisson counts model; 2PPCM: 2-parameter Poisson counts model; LNIRT: log-normal item response theory; BRM-1: beta response model 1; BRM-2: beta response model 2.

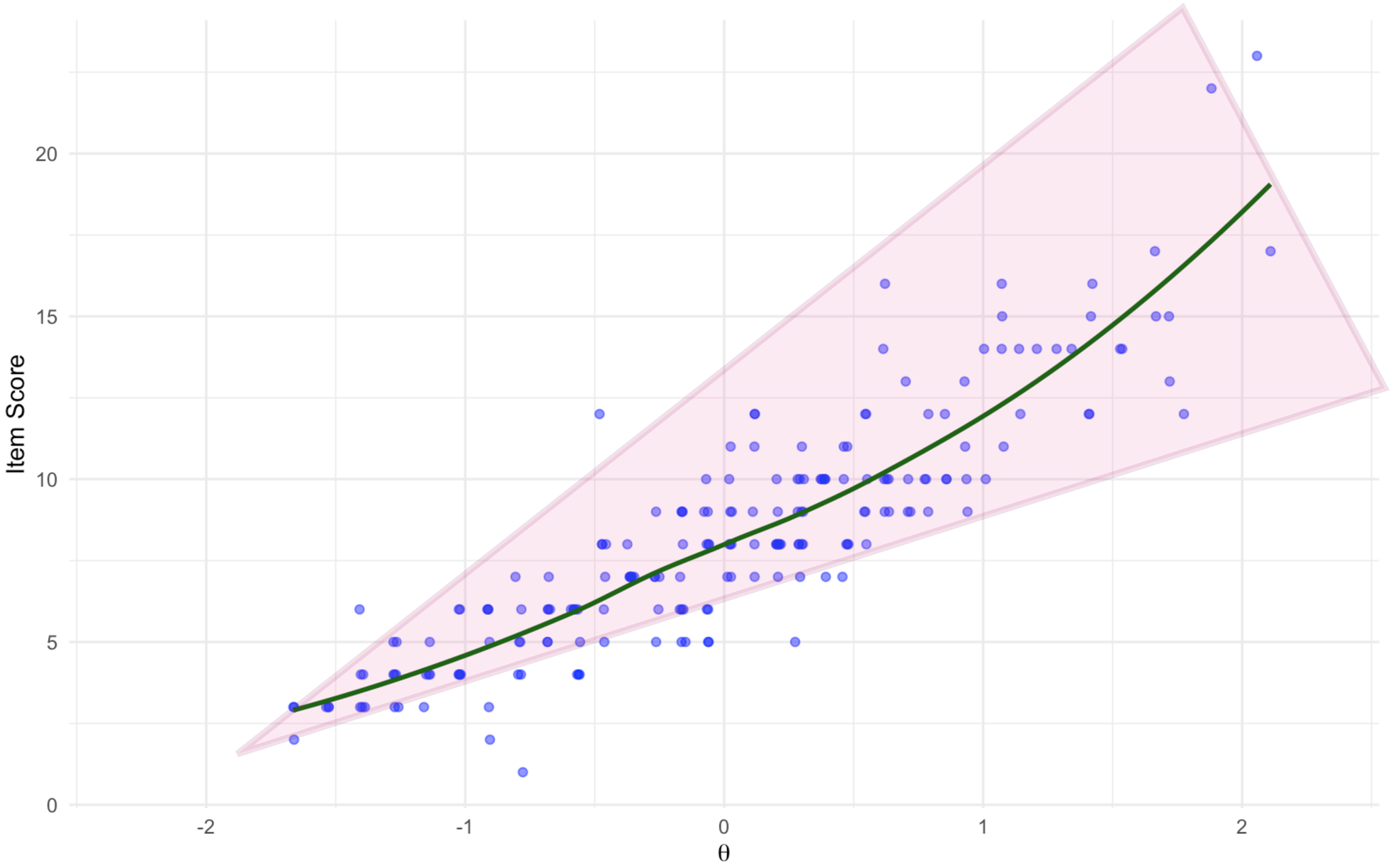
Myszkowski (2024)

Example : Modeling fluency scores
(Myszkowski & Storme, 2021, 2025)

- Development of the 2-parameter Poisson counts model (2PPCM; Myszkowski & Storme, 2021) for fluency tests
  - …and its estimation in a Bayesian multilevel regression framework in R and Stan (Myszkowski & Storme, 2025)
- Re-analysis of divergent thinking fluency data (from Silvia et al. 2008; Forthmann et al., 2019)
  - Alternate uses (e.g., “uses of a brick” 🧱)
  - Instances (e.g., “things that make a noise” 🔊)
  - Consequences (e.g., “if people no longer had to sleep” 💤)

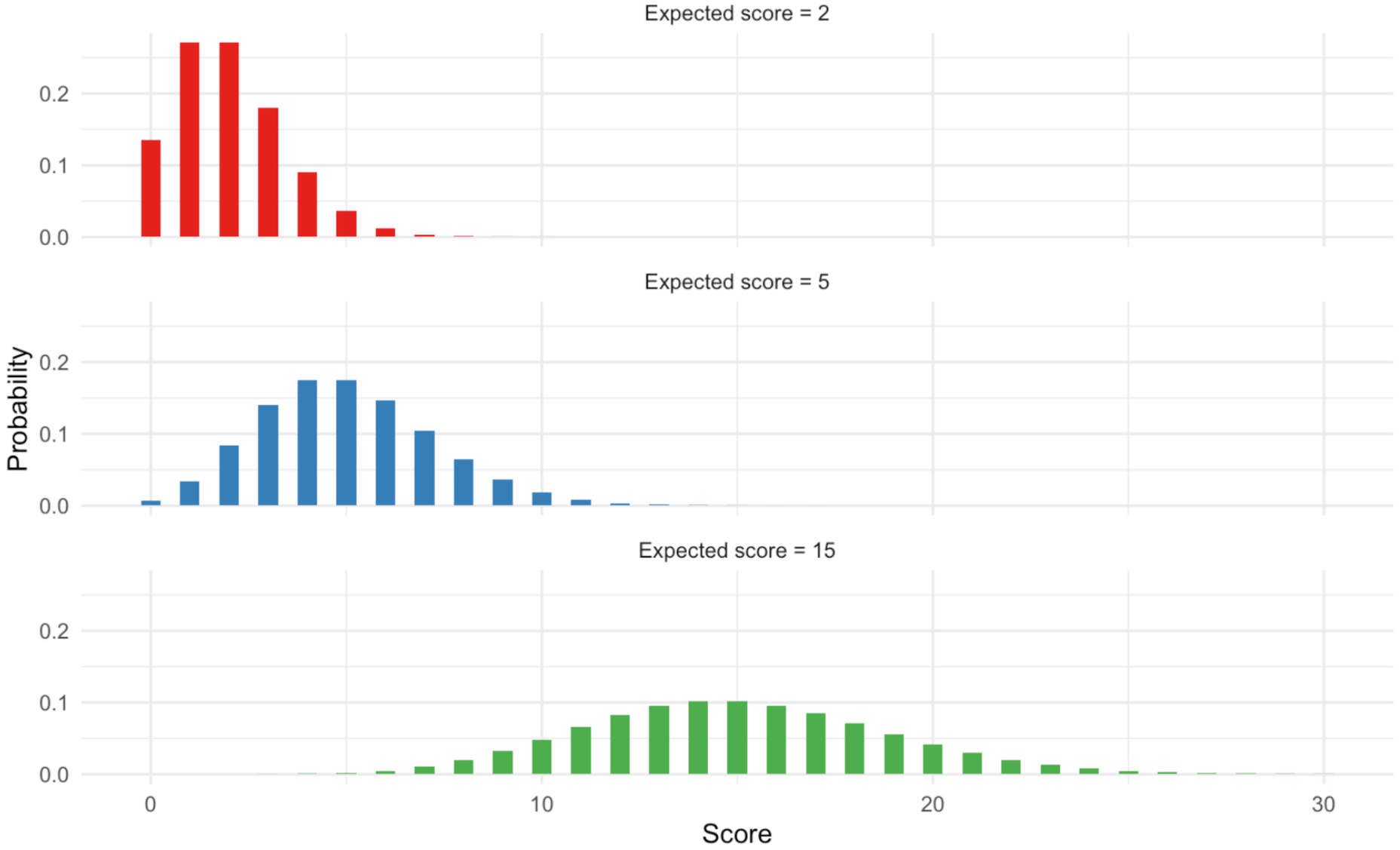

Frequency
0
10
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Item score


Item Score
20
15
10
5
0
-2
-1
0
1
2
θ


Item Score
20
15
10
5
0
-2
-1
0
1
2
θ

How do we account for these specificities?

- **Discrete** responses with increasing variance ?
  - **Poisson** distribution
- **Floor** effect at 0 ?
  - **Exponential** response function (i.e. logarithmic link)
- …and the typical components of an item response model:
  - **Item effects** : Difficulty and discrimination
  - **Person effects** : Person fluency



Myszkowski, 2025

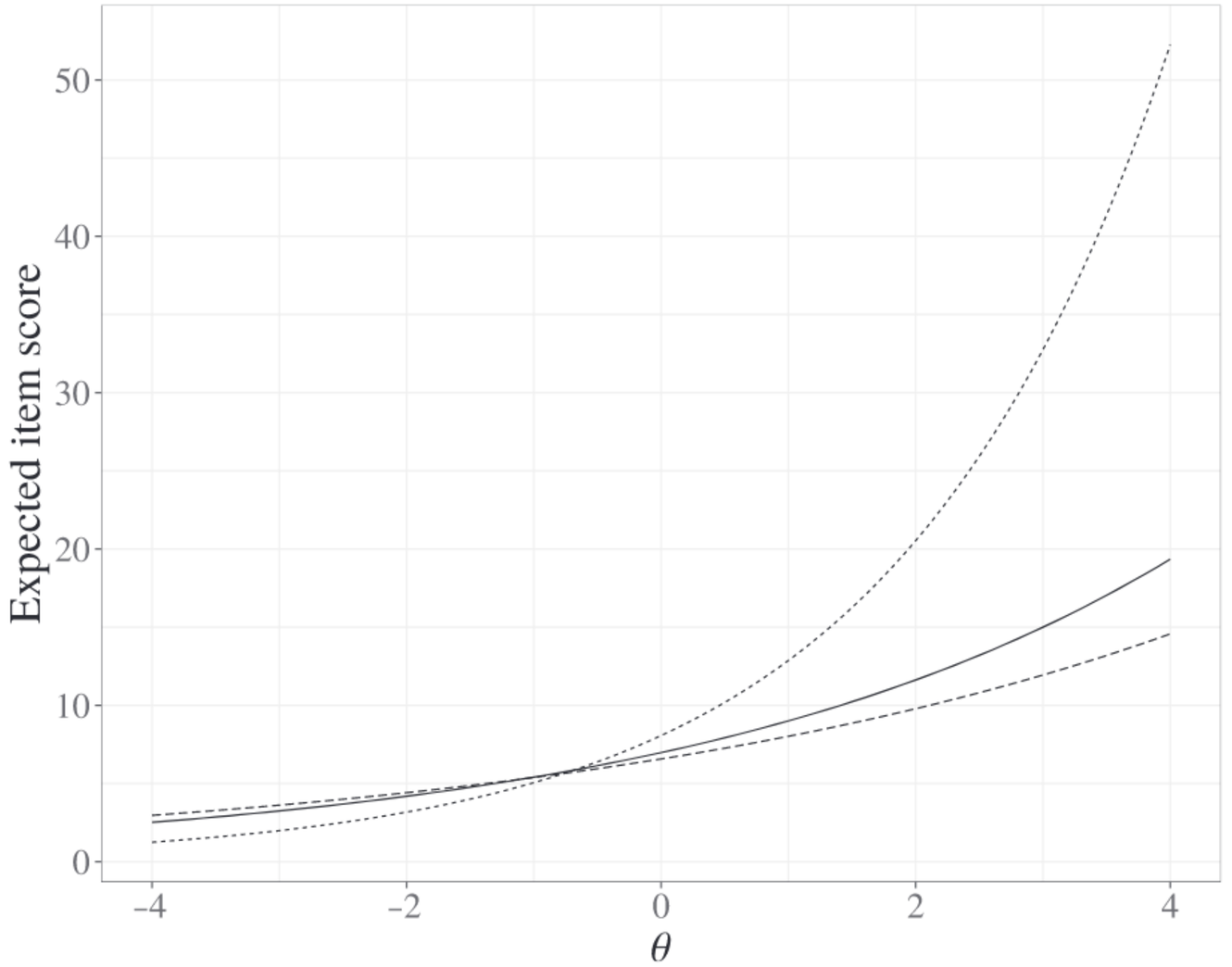
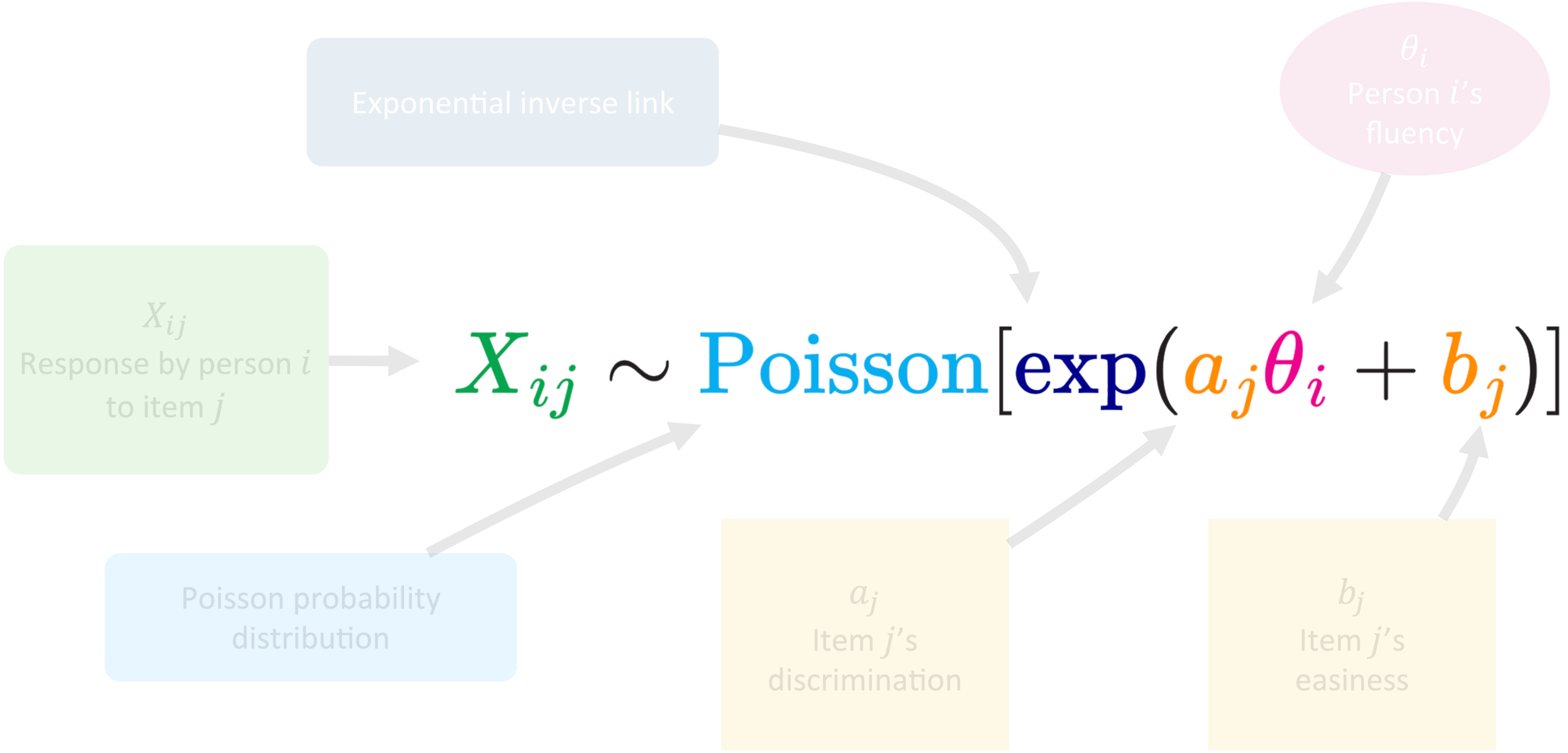


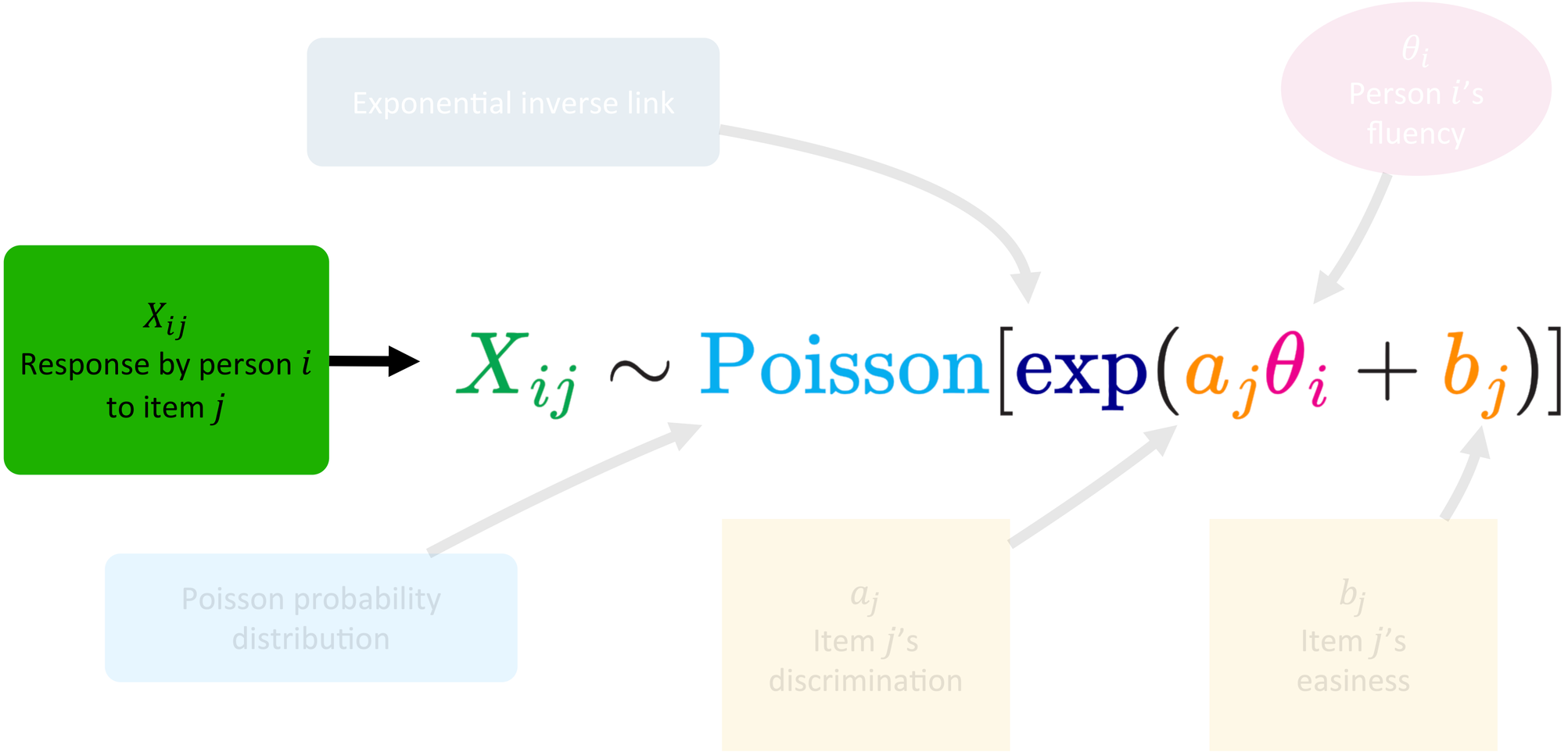
Figure 9 Expected scores of three items modeled with a 2-parameter log-linear model

Myszkowski, 2024

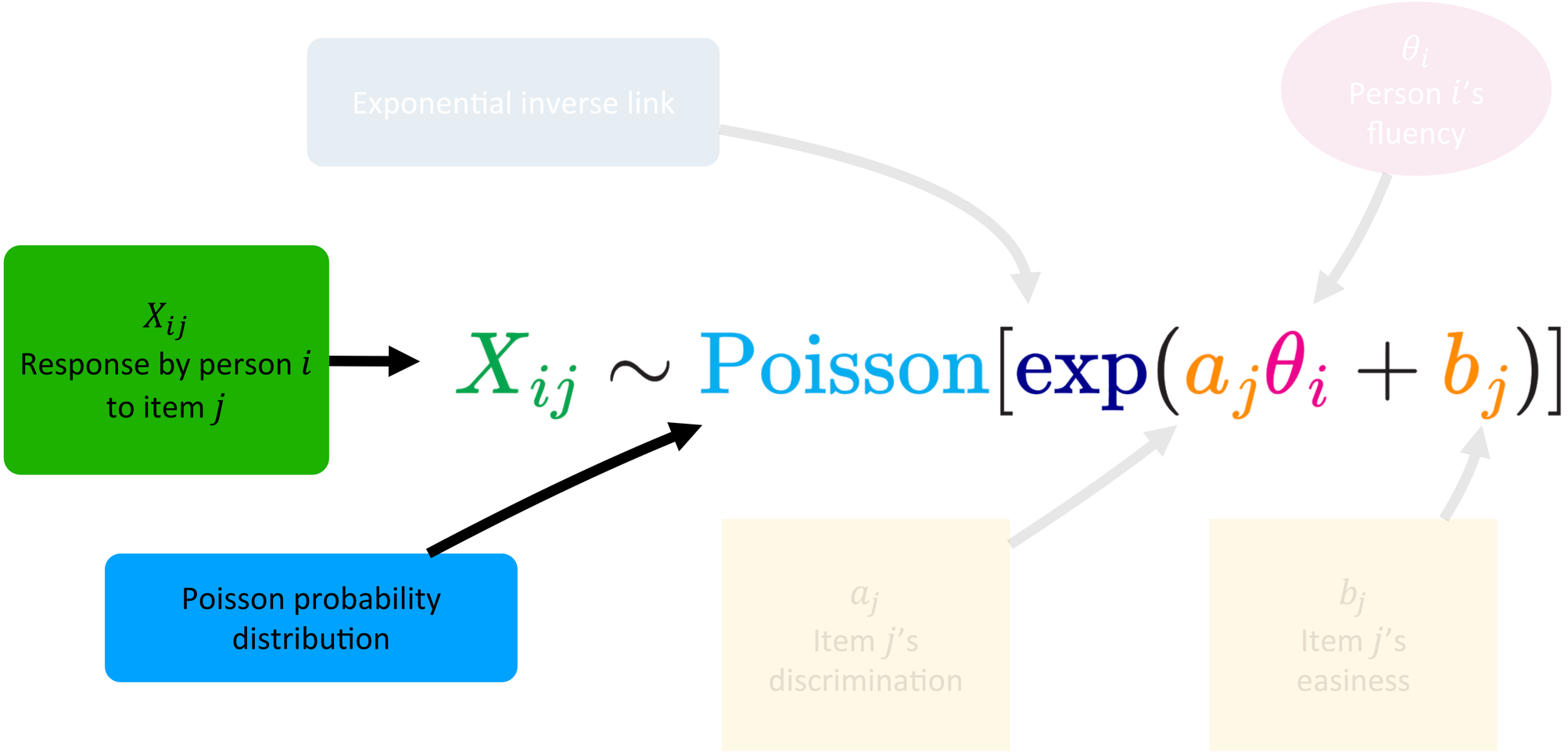
The 2-parameter Poisson counts model



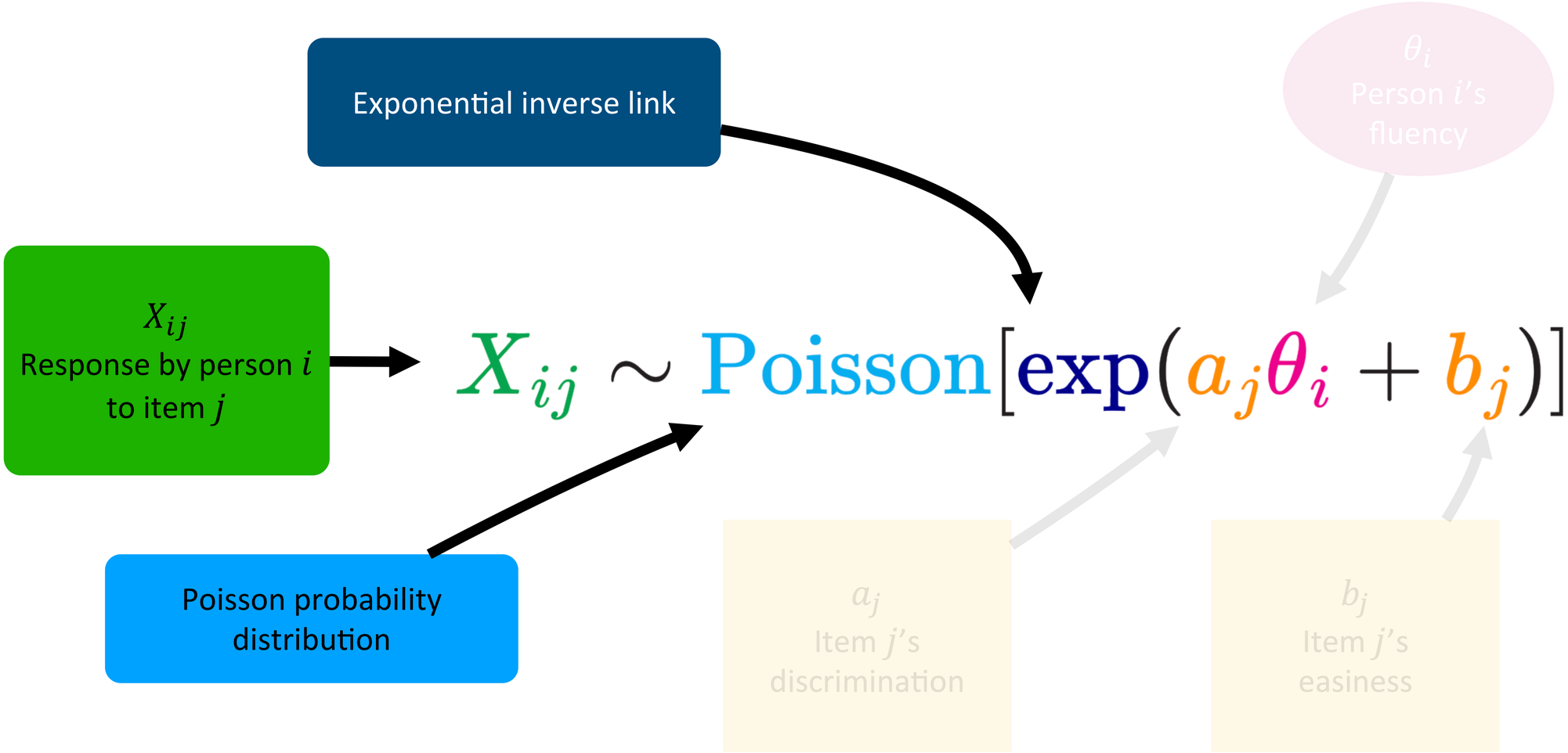
The 2-parameter Poisson counts model



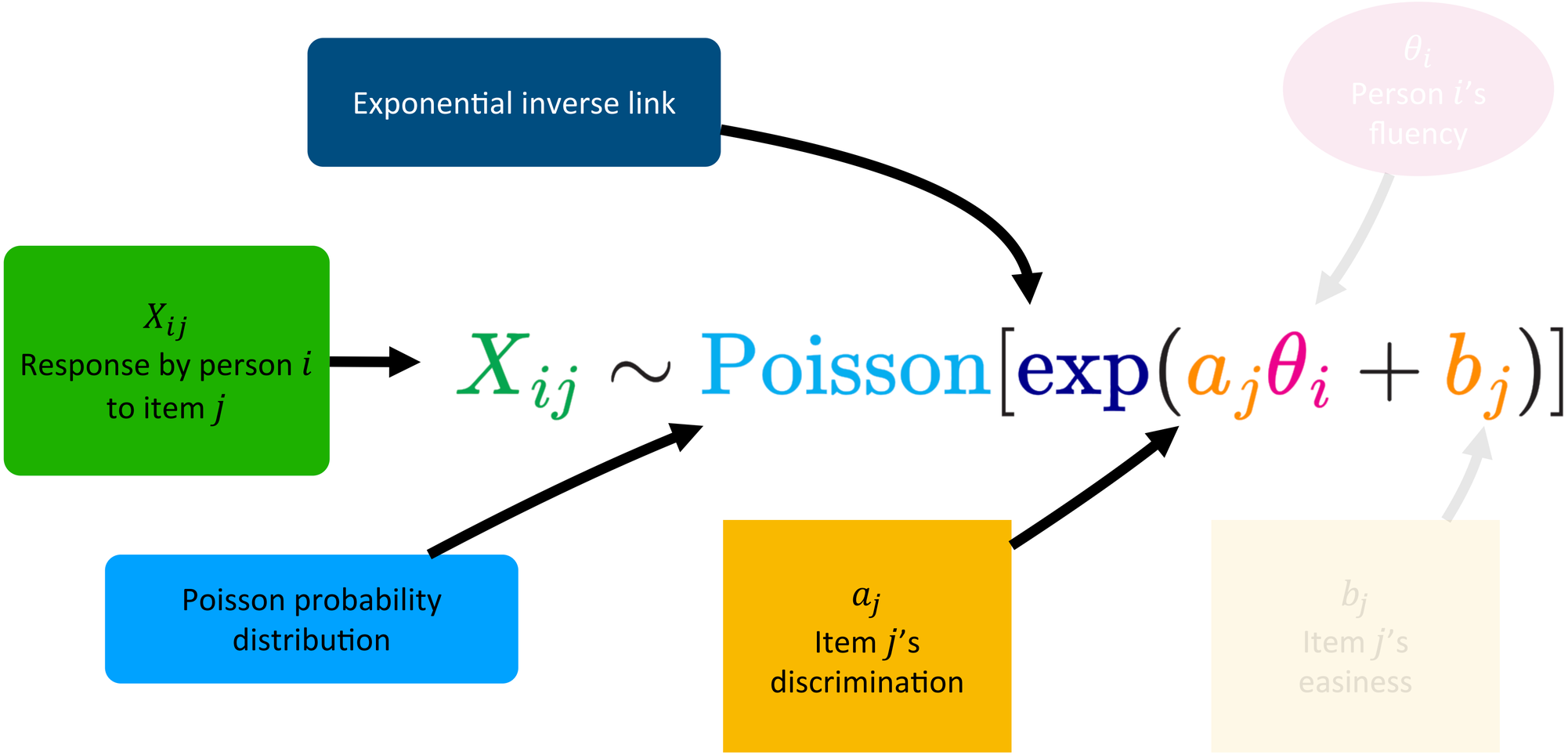
The 2-parameter Poisson counts model



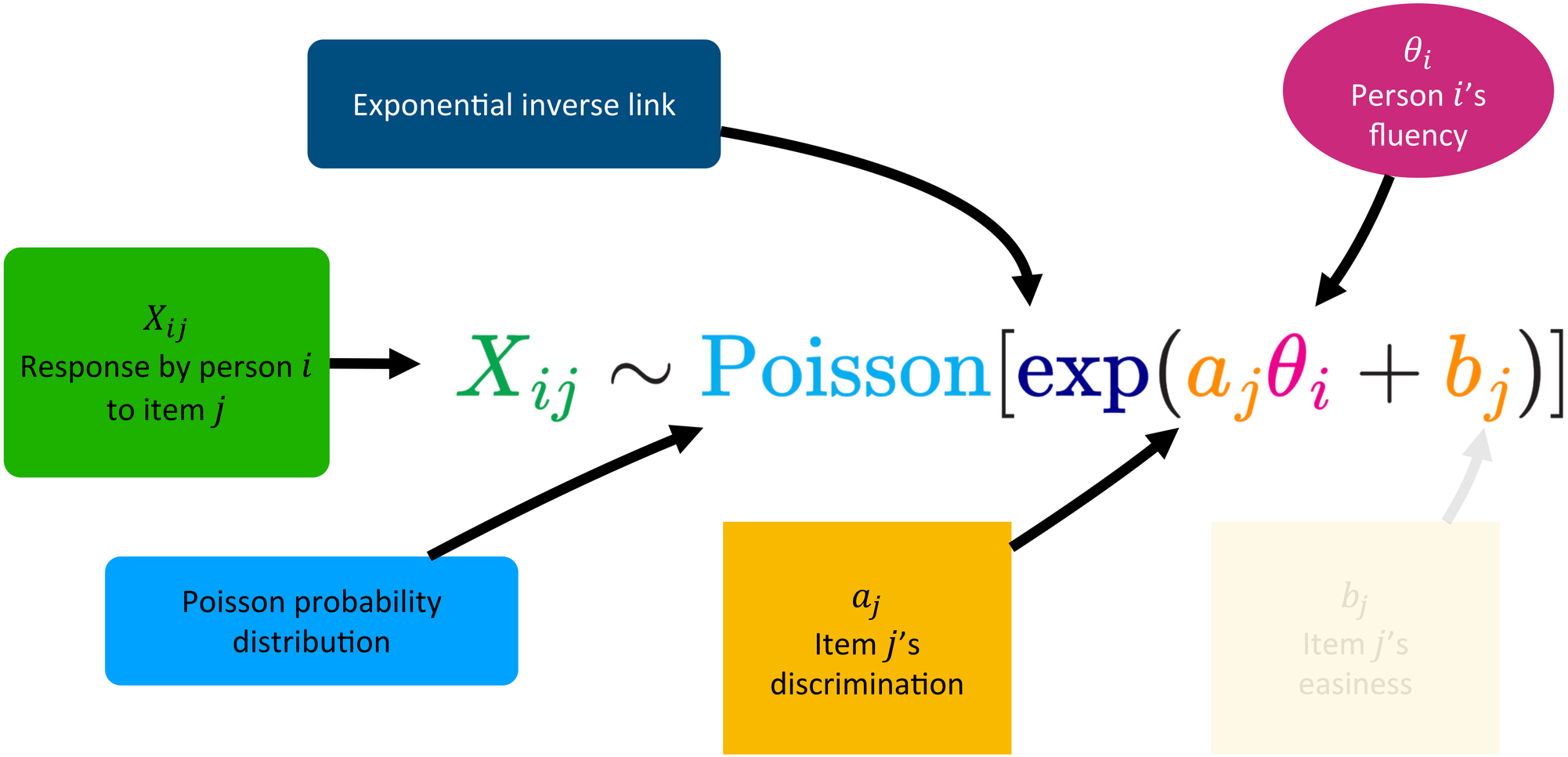
The 2-parameter Poisson counts model



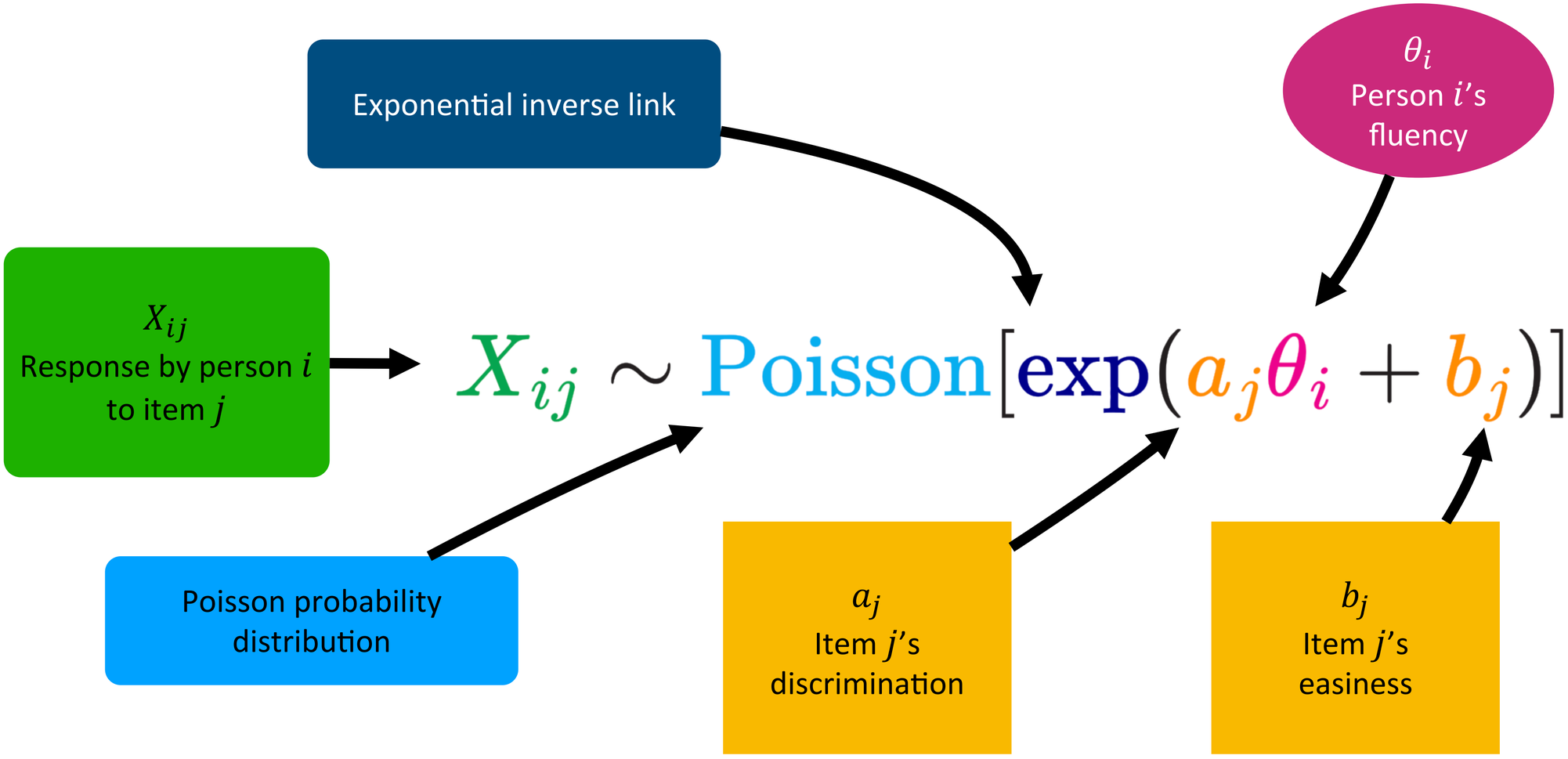
The 2-parameter Poisson counts model



The 2-parameter Poisson counts model



The 2-parameter Poisson counts model

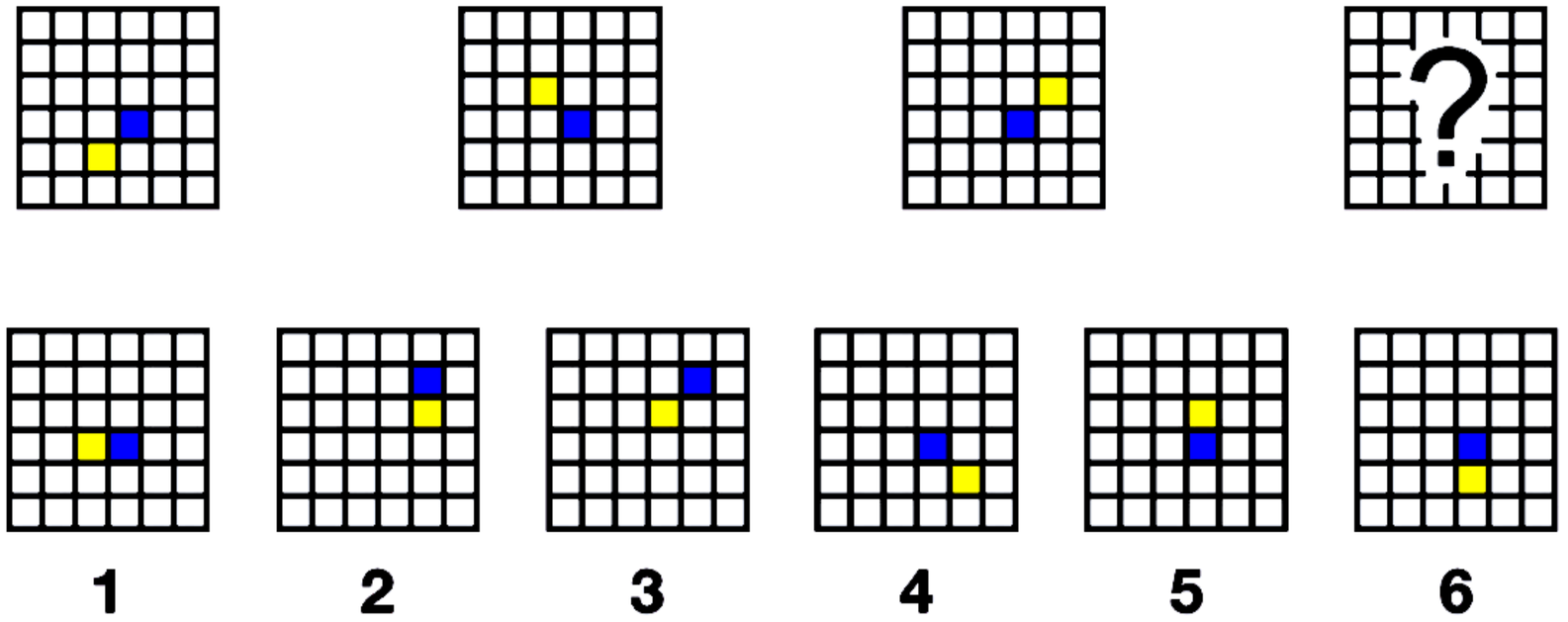


All the benefits of IRT

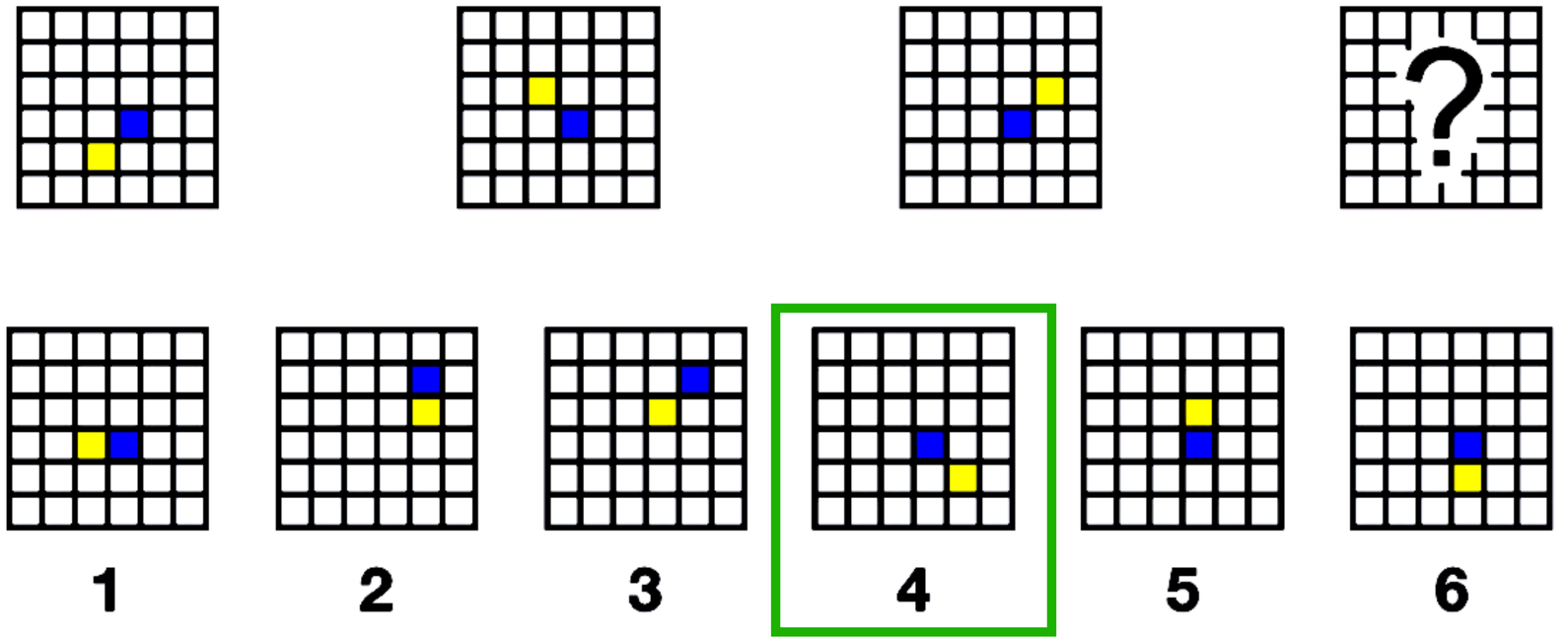
Unlocking the benefits of modern test theory for more tasks

- Better account of **reliability** (Myszkowski & Storme, 2025)
- Gain in **accuracy** from average scoring (Myszkowski, 2024)
- **Multidimensional** extensions (Myszkowski & Storme, 2021; Myszkowski, 2024)
- And other IRT things…
  - missing data, matrix sampling, measurement invariance testing, optimal test assembly, computer adaptive testing, etc.

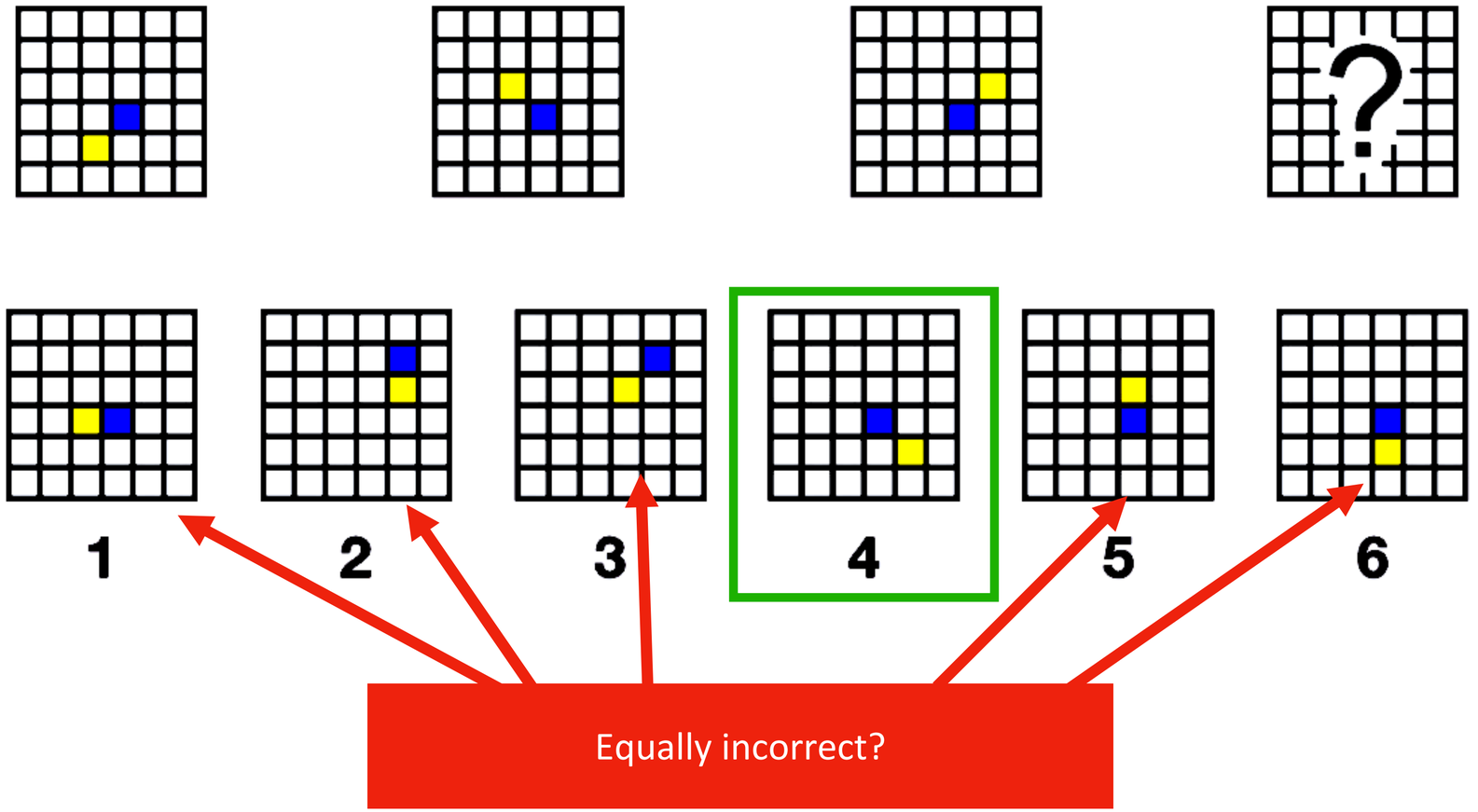
Boosting reliability with distractors: Nested logit models



Storme, Myszkowski, Baron & Bernard, 2020

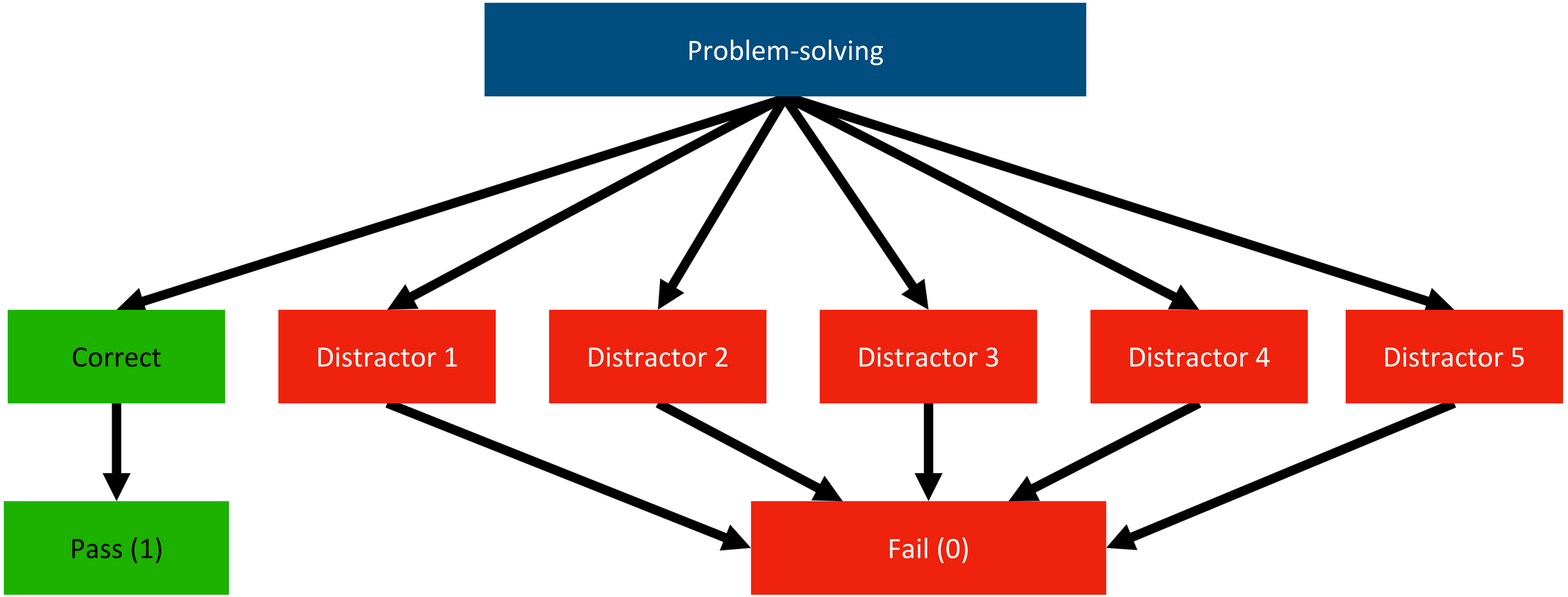


Storme, Myszkowski, Baron & Bernard, 2020

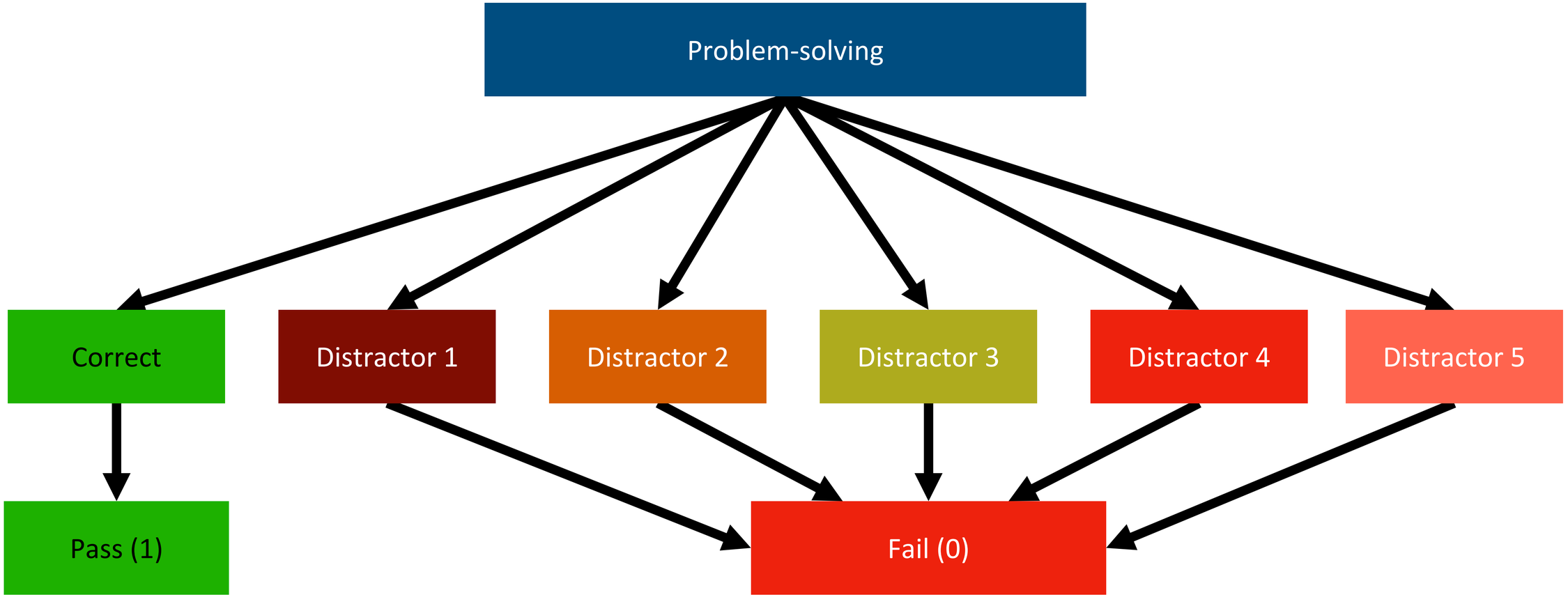


Storme, Myszkowski, Baron & Bernard, 2020

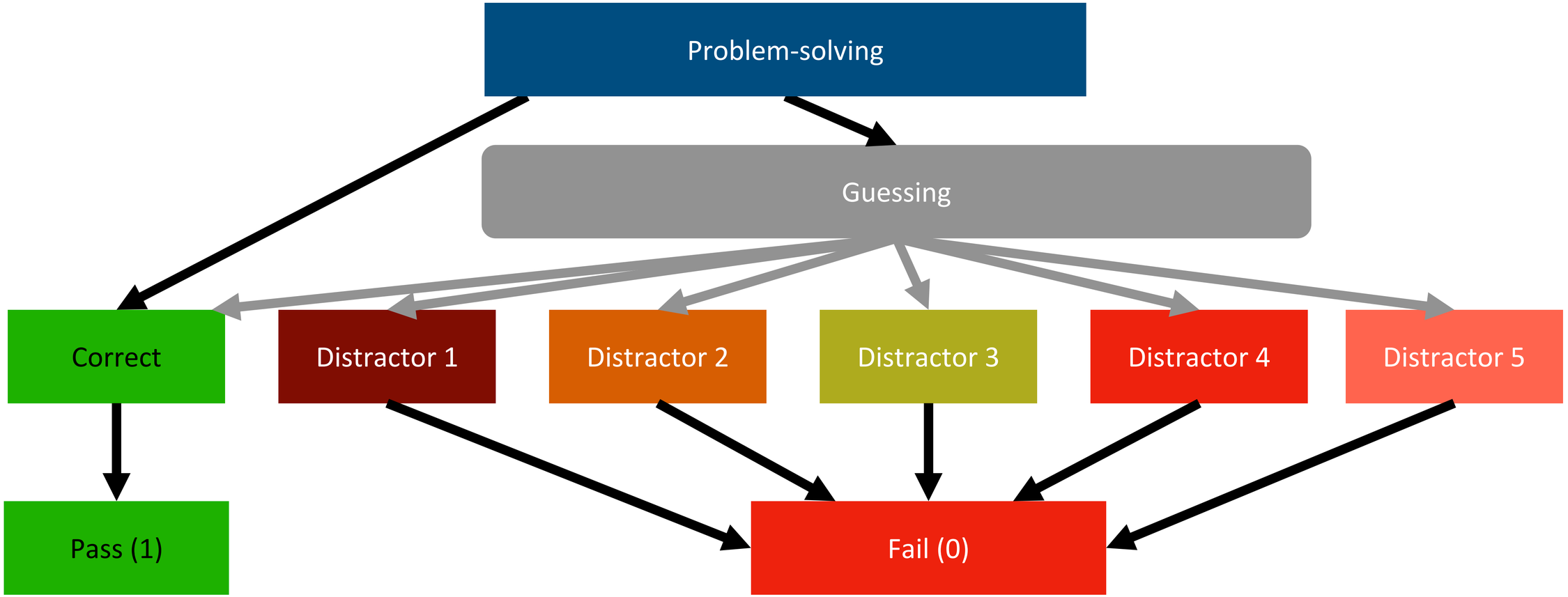
From categorical to binary



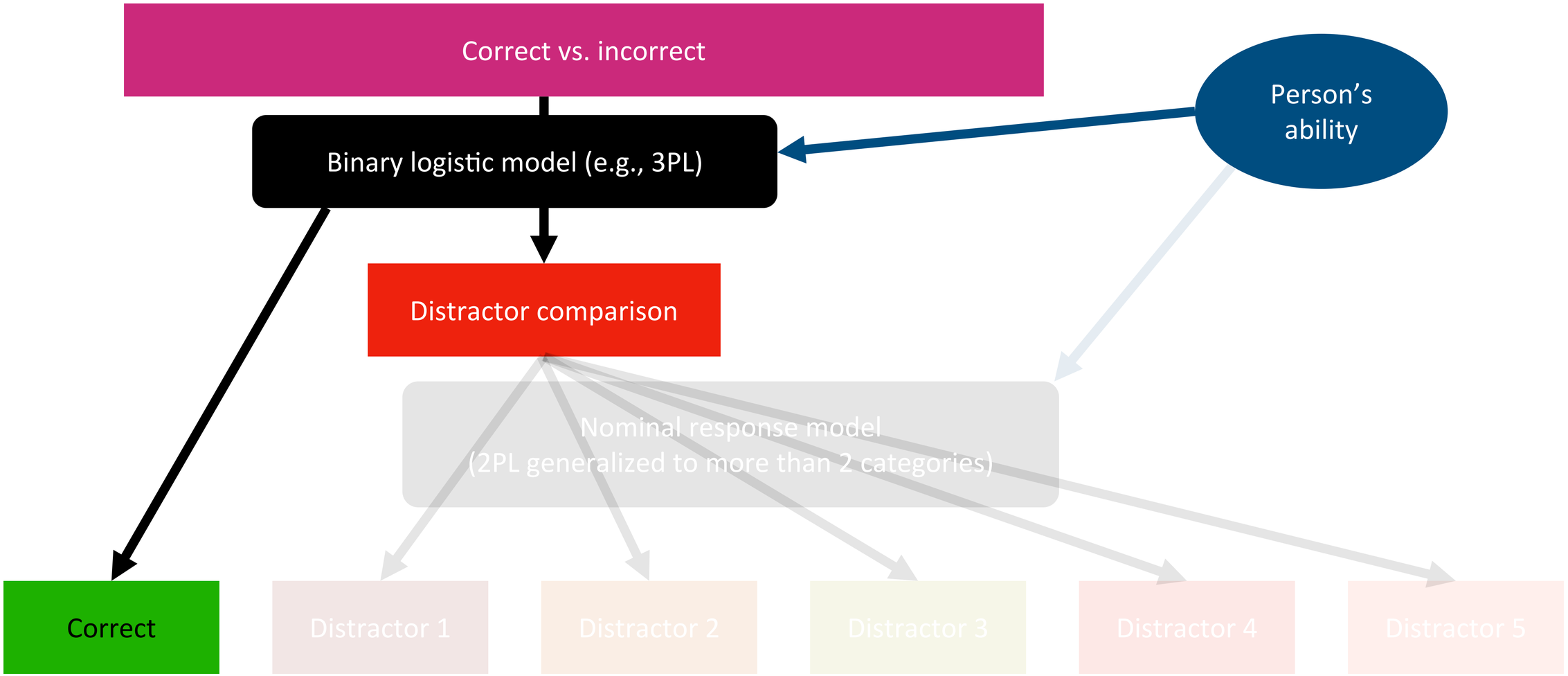
From categorical to binary



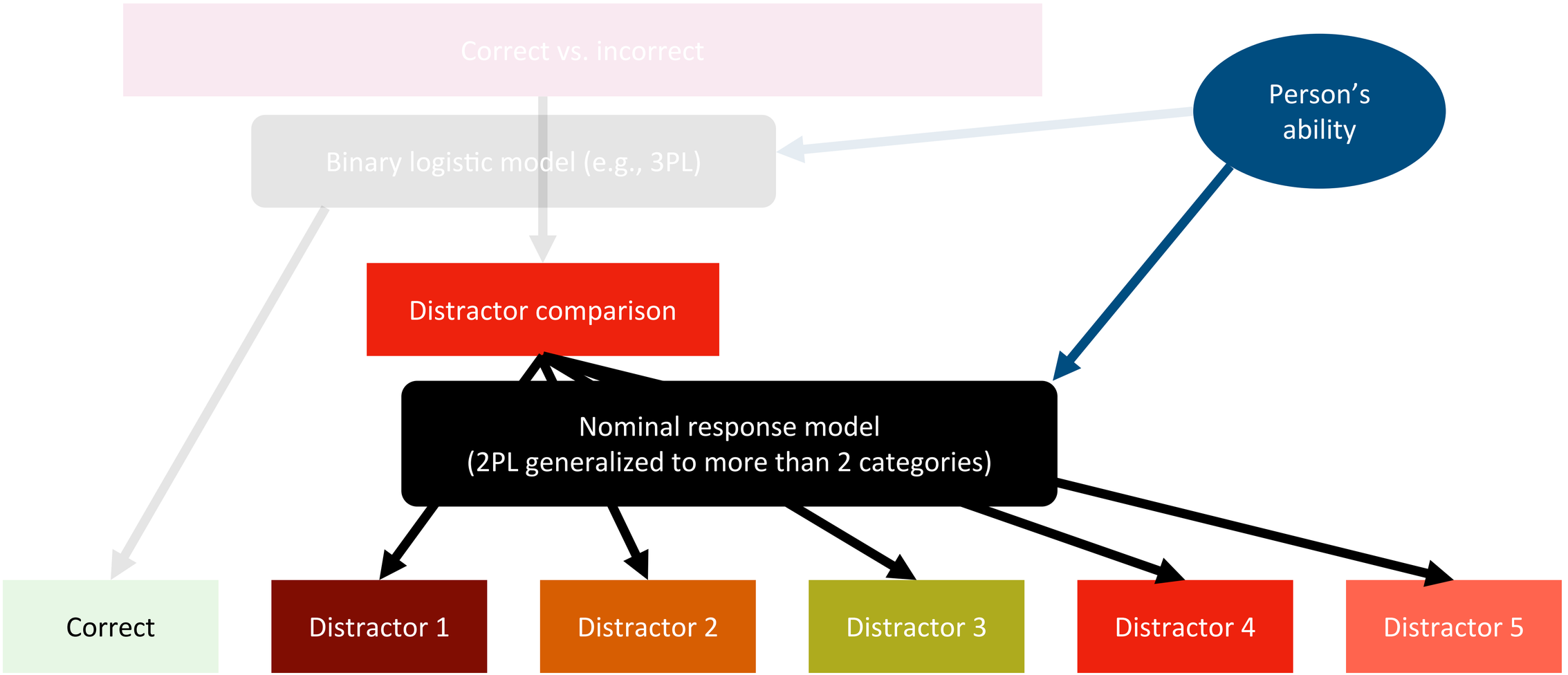
From categorical to binary



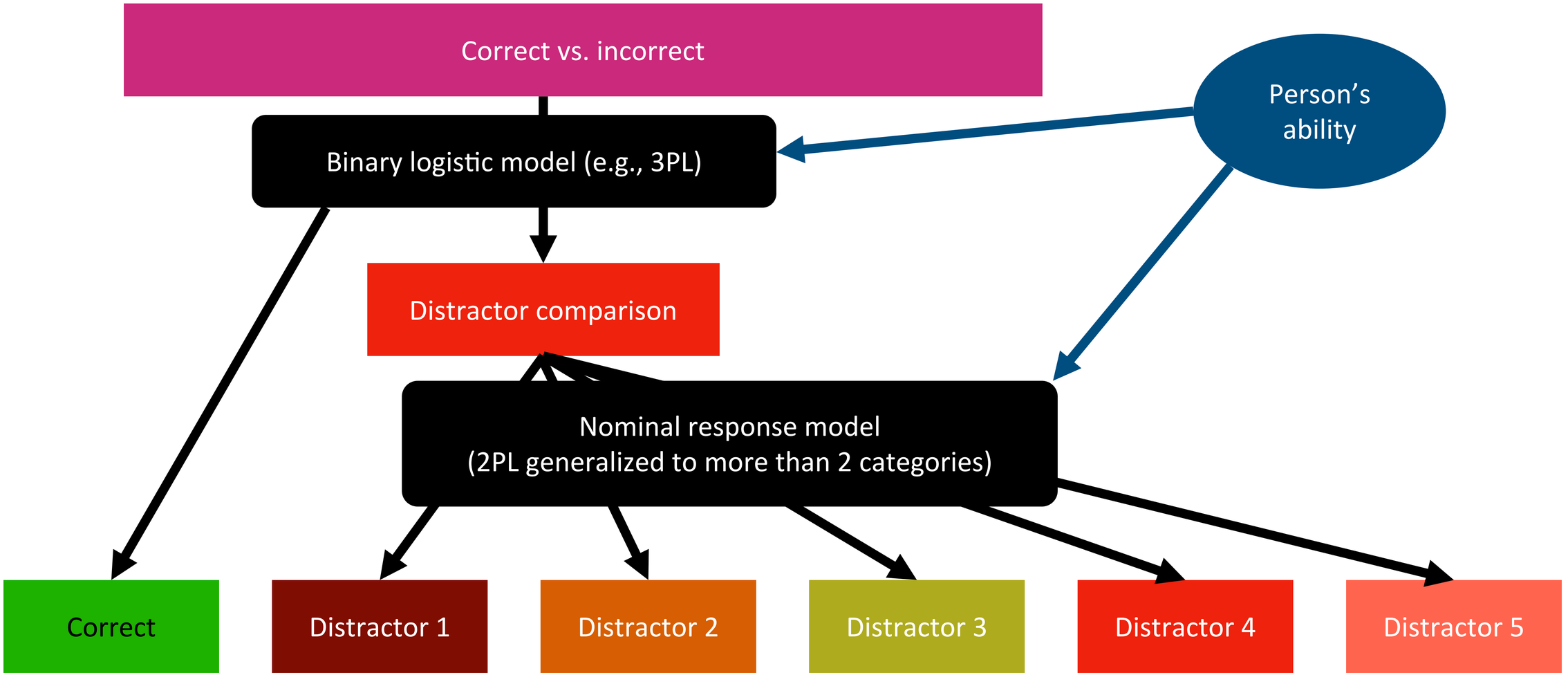
Nested logit models
(Suh & Bolt, 2010)



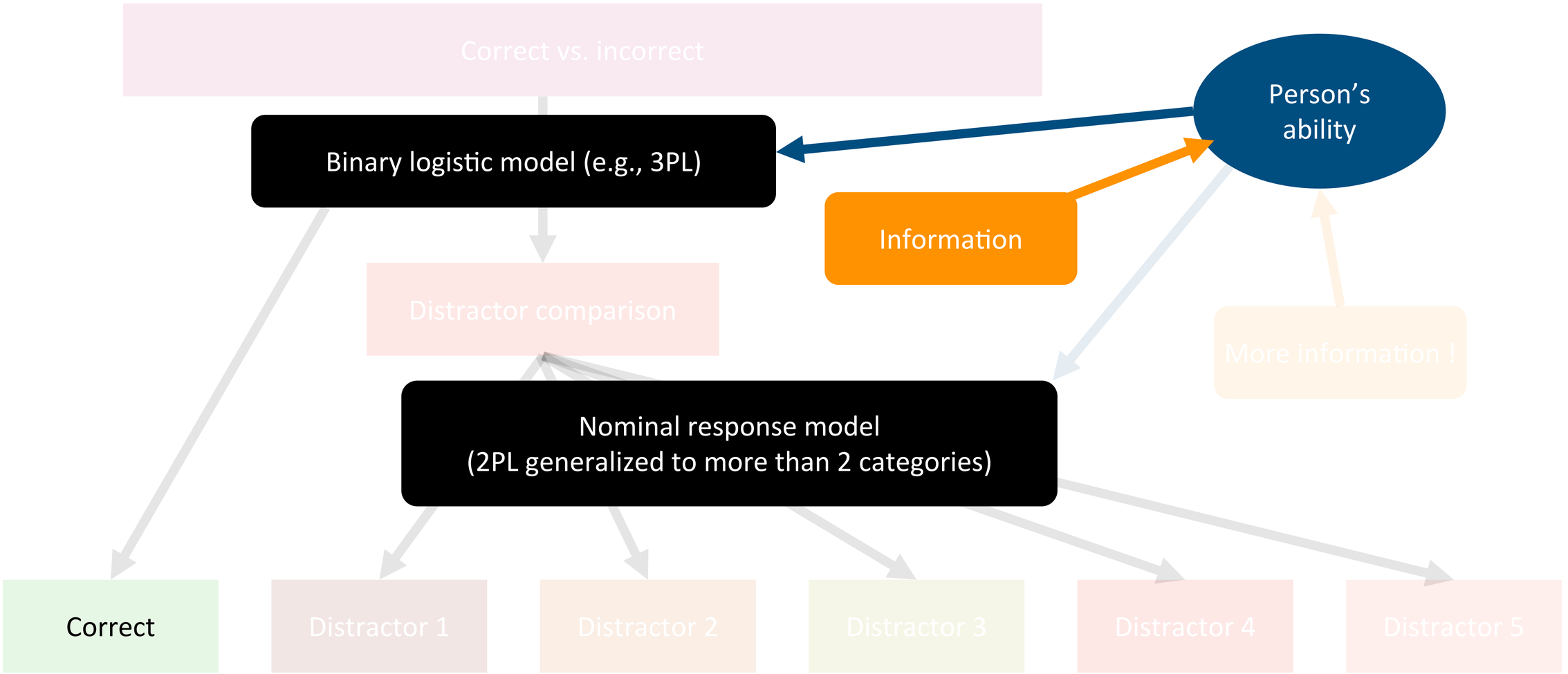
Nested logit models
(Suh & Bolt, 2010)



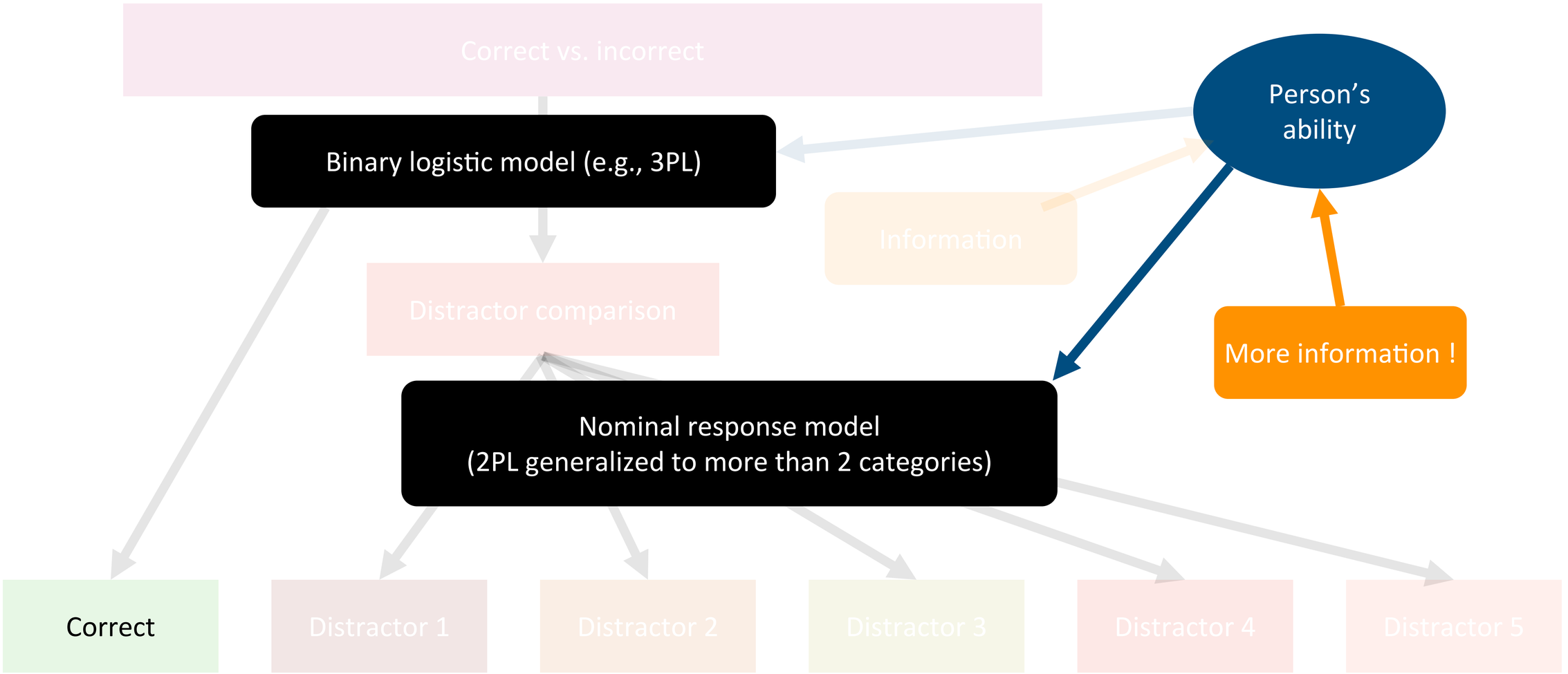
Nested logit models
(Suh & Bolt, 2010)



Nested logit models
(Suh & Bolt, 2010)



Nested logit models
(Suh & Bolt, 2010)



Example 1 : Figural series completion test
(Storme, Myszkowski, Baron & Bernard, 2020)

- Figural series completion test from an online assessment company
- 2949 voluntary adult respondents (enrolled on website)
- Binary IRT models + Categorical IRT models estimated in R package “mirt” (Chalmers, 2012)
- Significant ($p < .001$) boost in reliability across the sample from using Nested Logit Models over Binary IRT models ($\approx +.15$ for persons 2 SD below the mean)

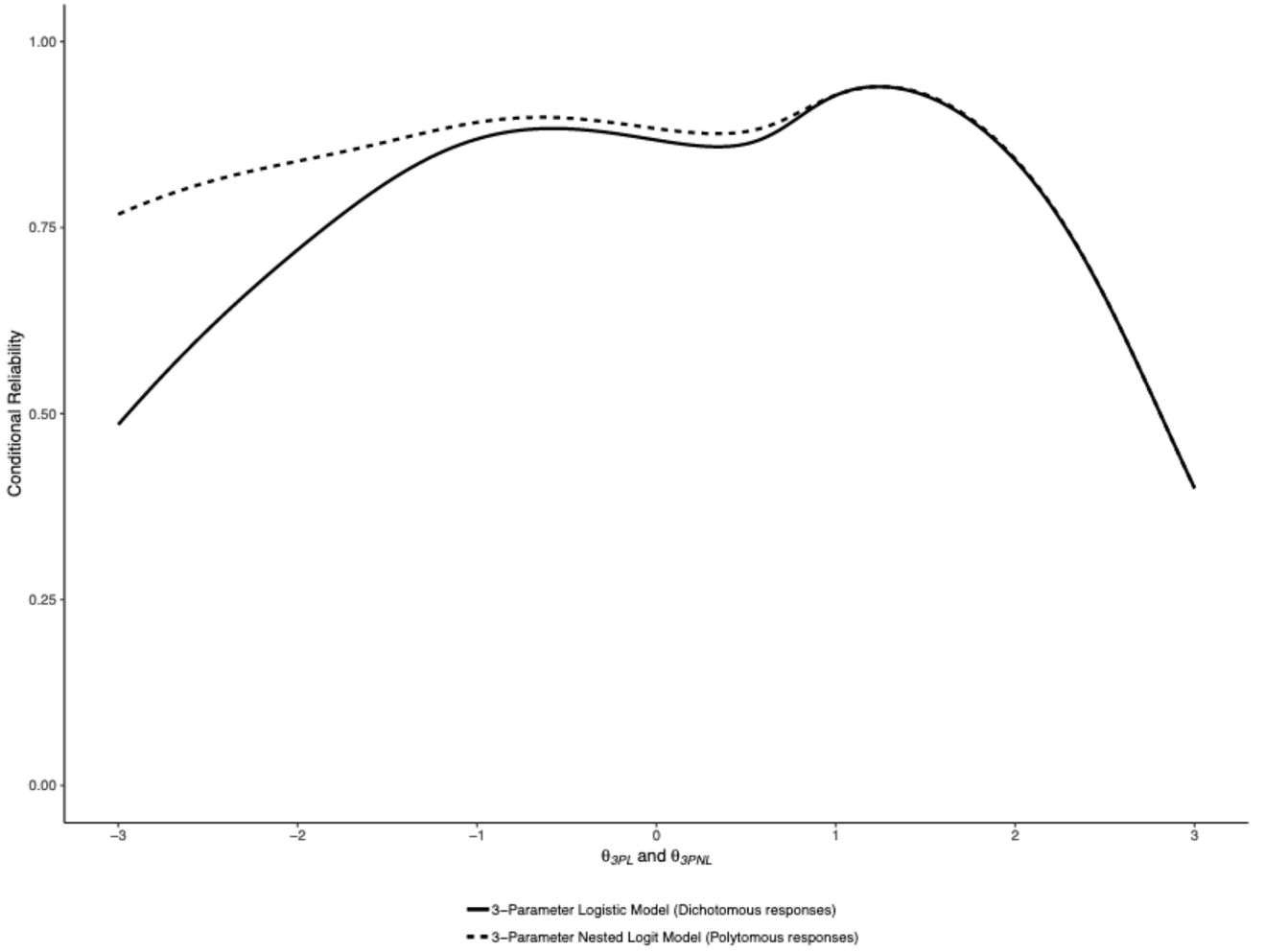
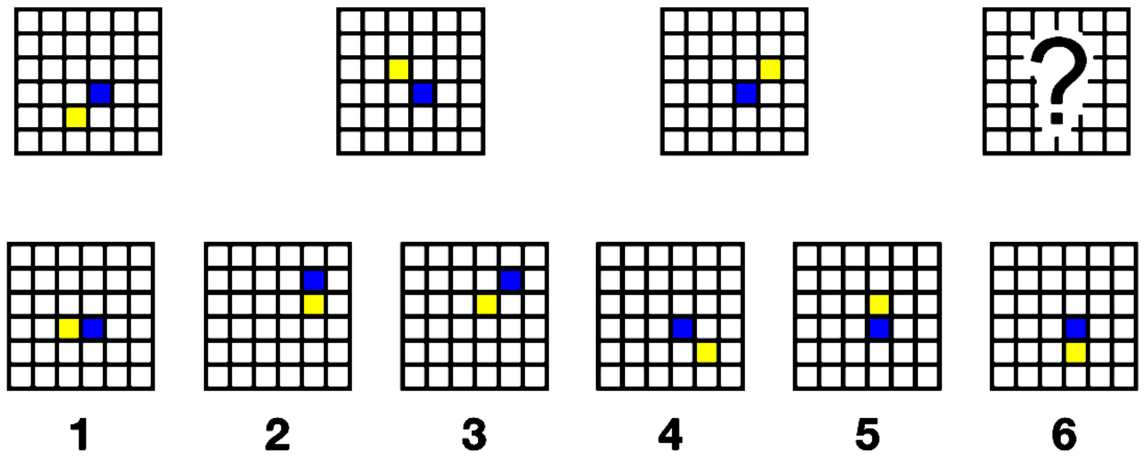
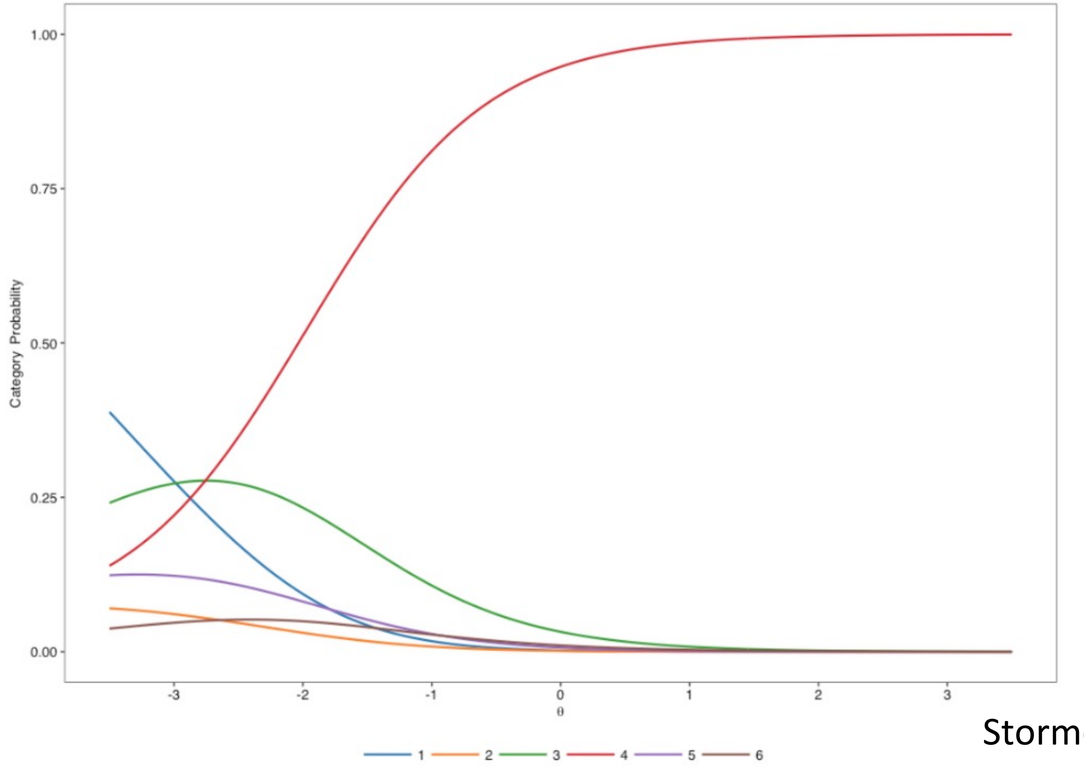


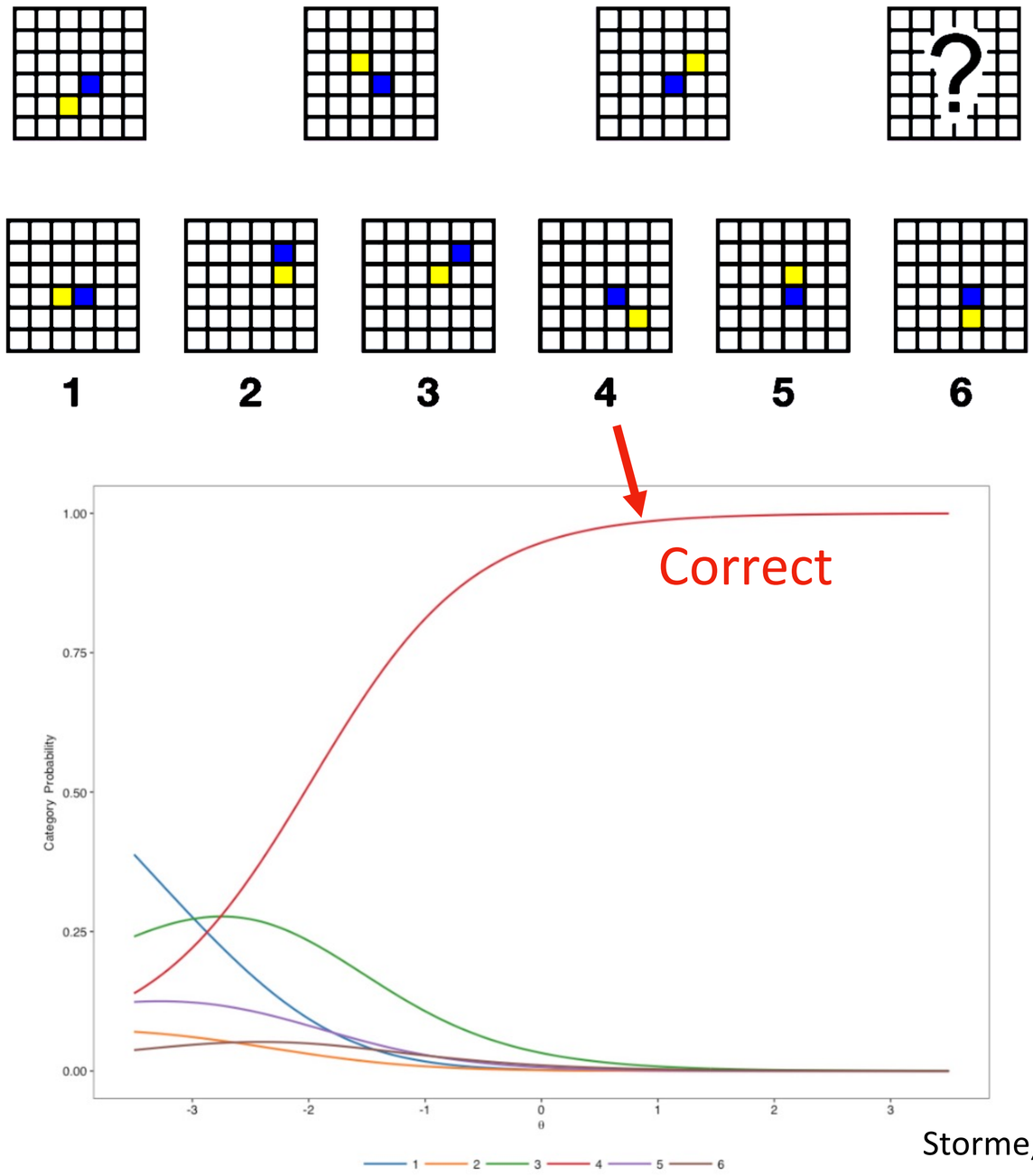
Figure 9. Comparison of the reliability functions of the 3-Parameter Logistic (3PL) and Nested Logit (3PNL) models.

Storme, Myszkowski, Baron & Bernard, 2020

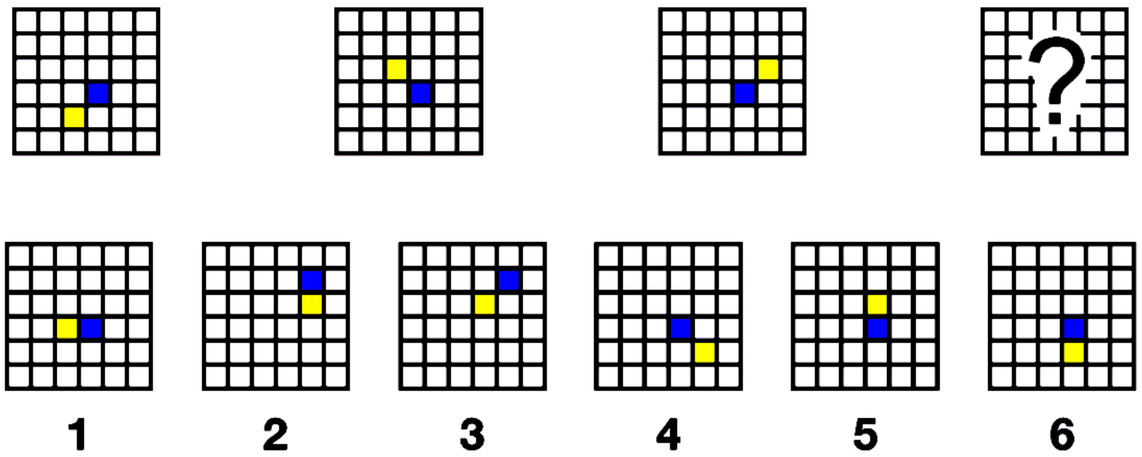




Storme, Myszkowski, Baron & Bernard, 2020

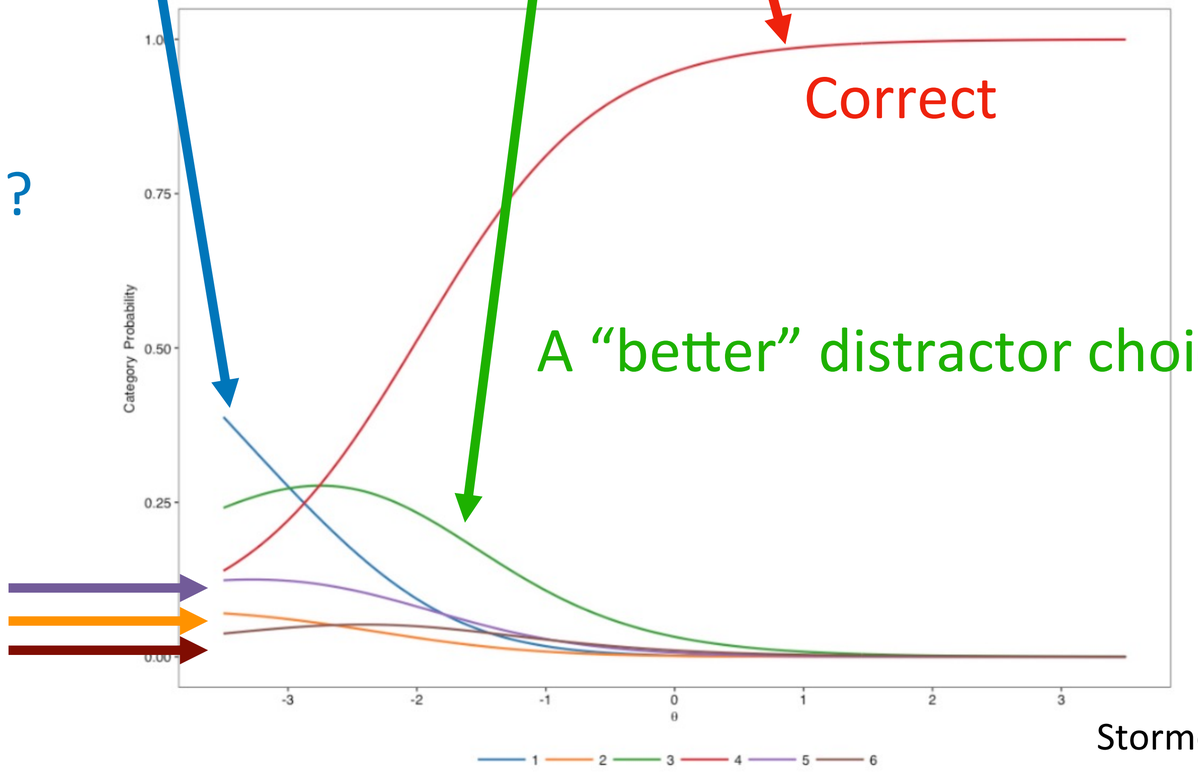


Storme, Myszkowski, Baron & Bernard, 2020



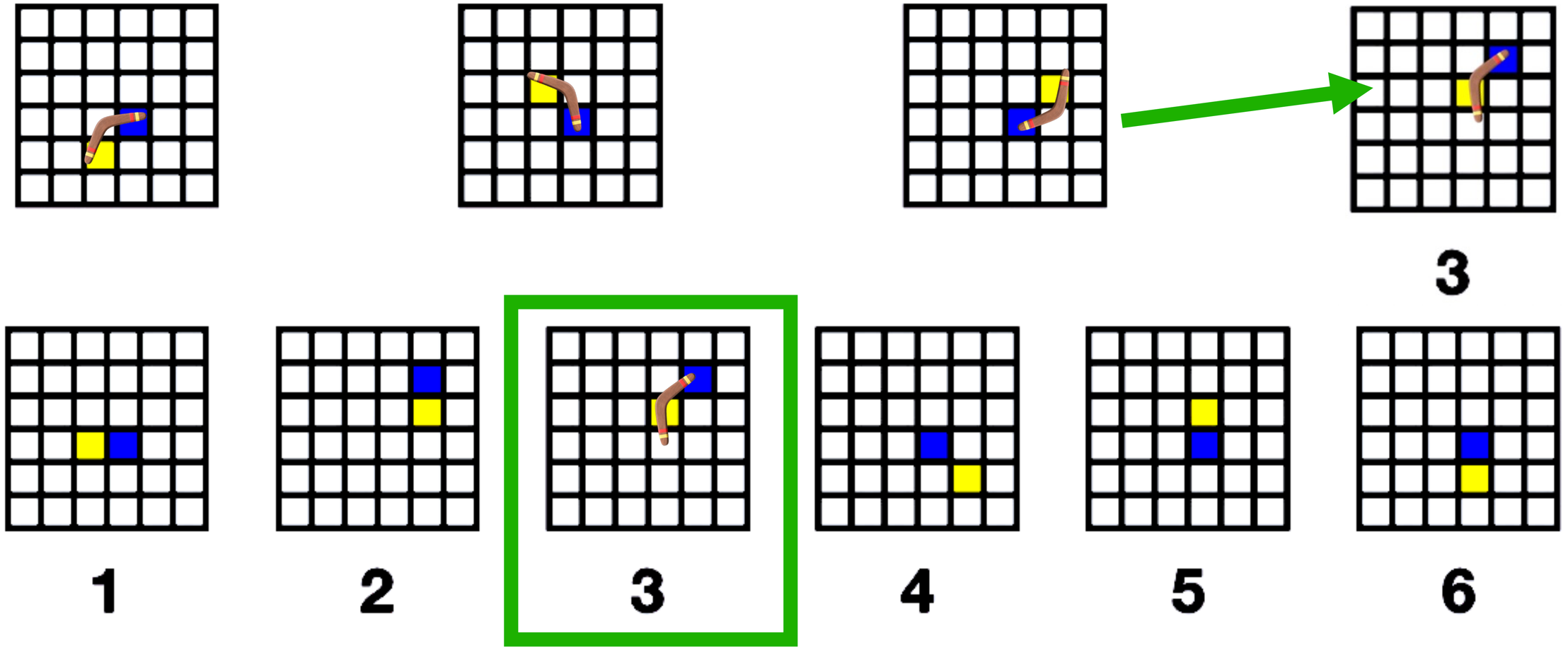
A “worse”

distractor choice ?



(Rarely picked)

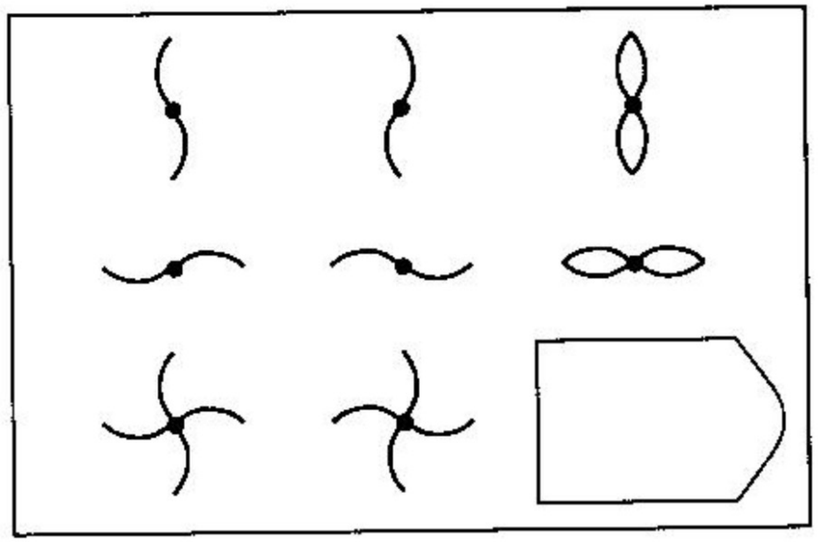
Storme, Myszkowski, Baron & Bernard, 2020

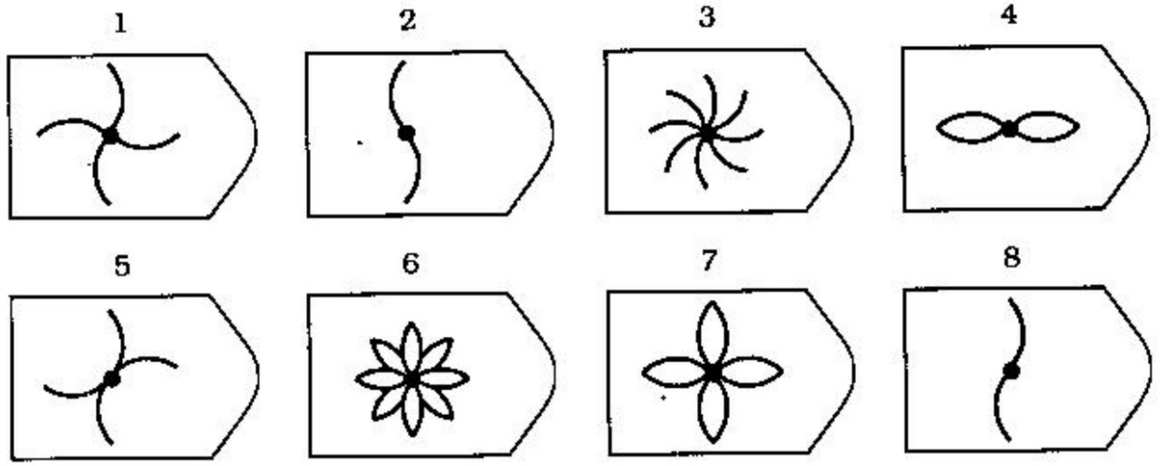


Storme, Myszkowski, Baron & Bernard, 2020

Example 2 : Progressive Matrices
(Myszkowski & Storme, 2018)

- Last Series of the Standard Progressive Matrices (Raven, 1941)
- 499 undergraduate students
- Binary IRT models + Categorical IRT models estimated in R package “mirt” (Chalmers, 2012)
- Significant ($p < .001$) boost in reliability across the sample from using Nested Logit Models over Binary IRT models ($\approx +.25$ for persons 2 SD below the mean)





Raven, 1941

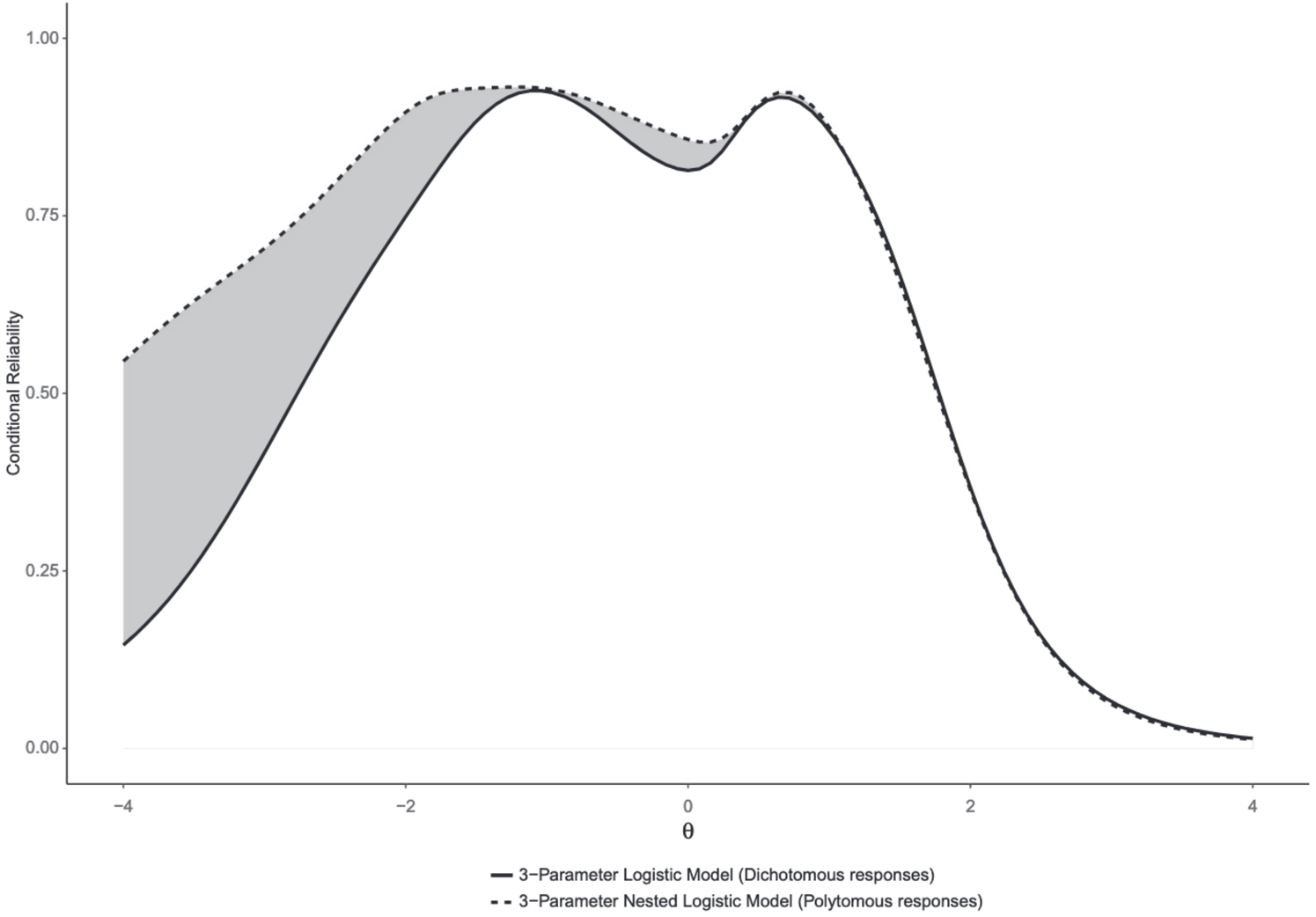


Fig. 3. Gain in reliability from recovering distractor response information.

Myszkowski & Storme, 2018

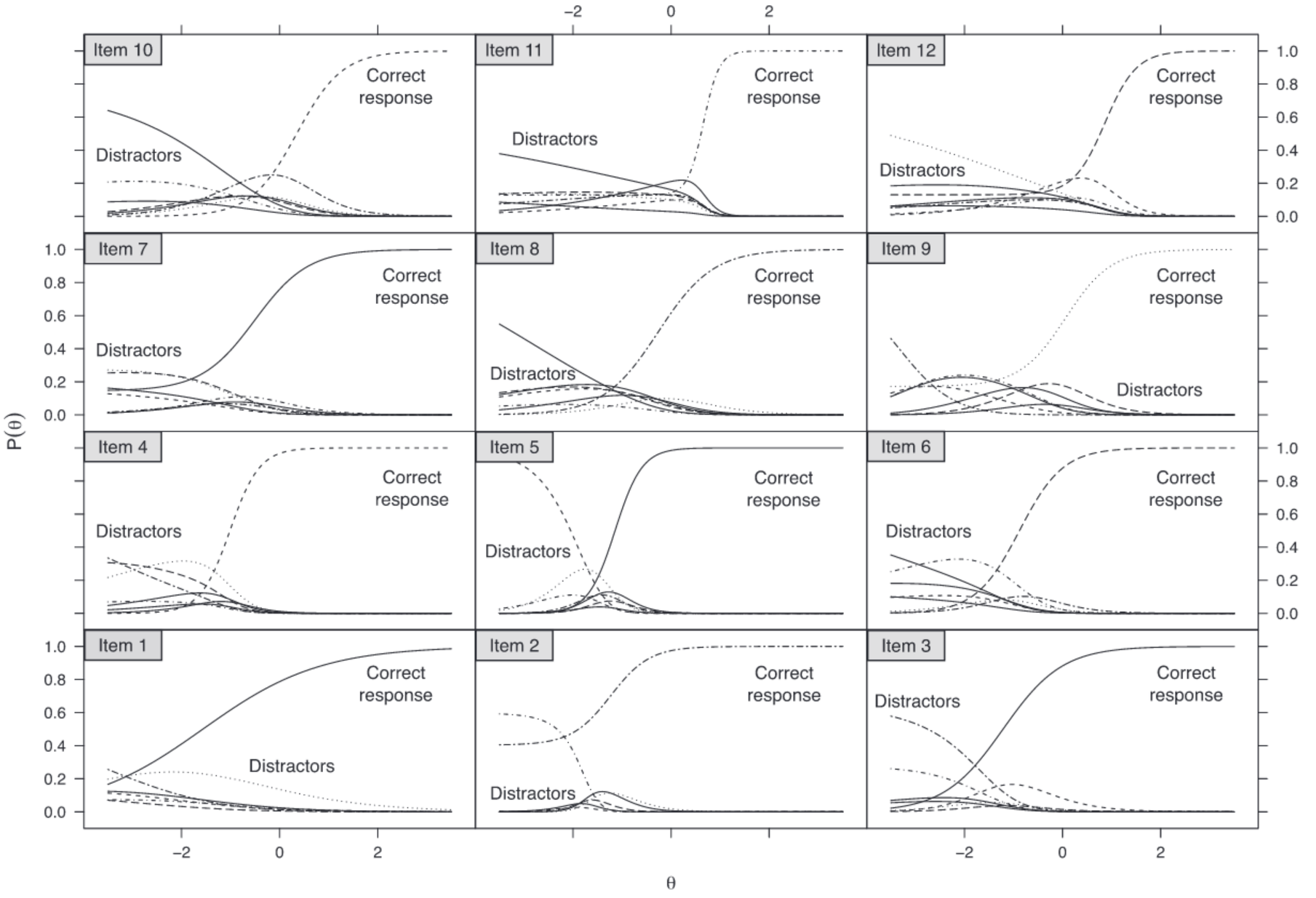
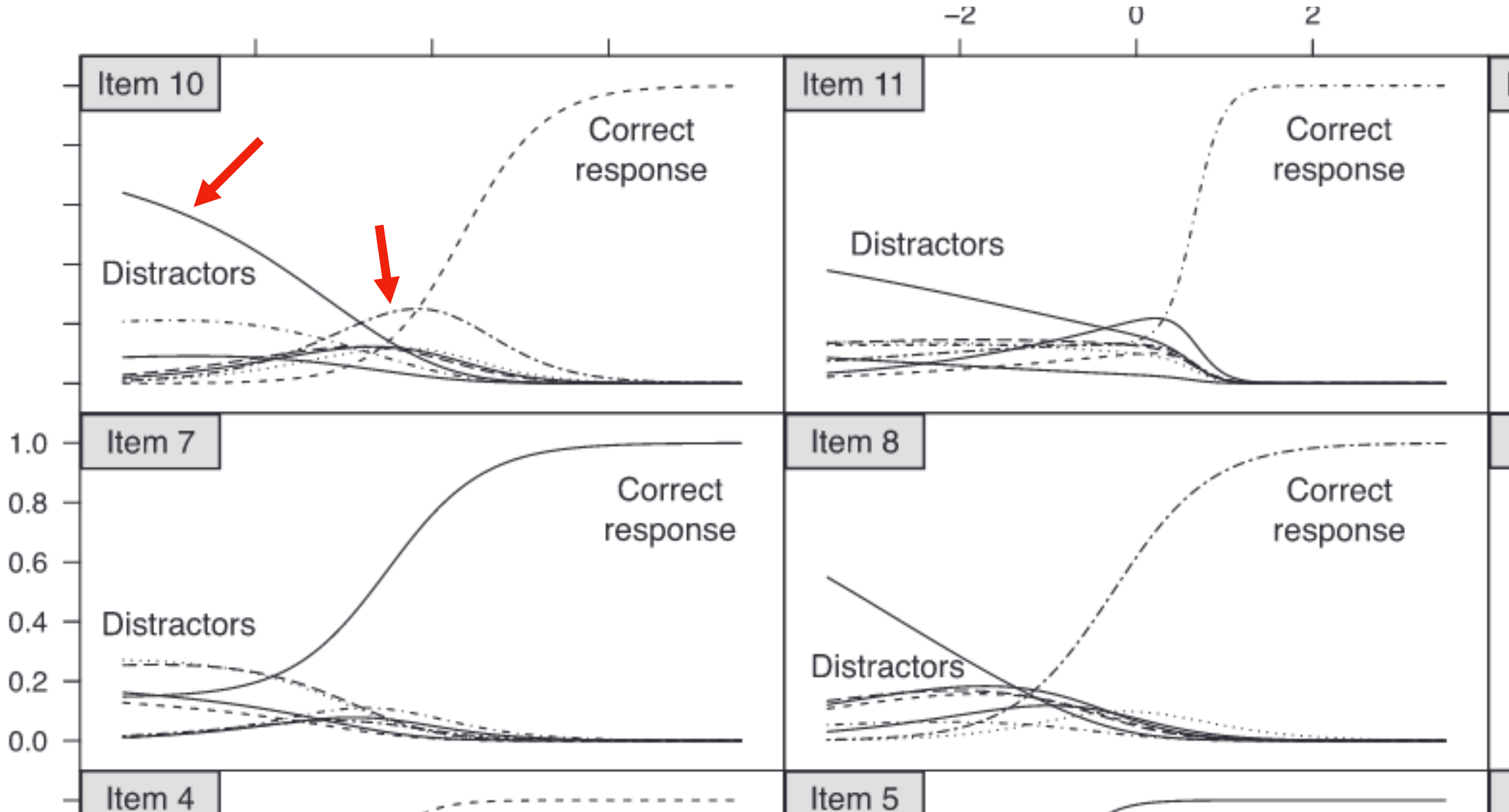


Fig. 2. Item characteristic curves of the best fitting (3PNL) model.

Myszkowski & Storme, 2018


−2
0
2
Item 10
Correct response
Distractors
Item 11
Correct response
Distractors
1.0
0.8
0.6
0.4
0.2
0.0
Item 7
Correct response
Distractors
Item 8
Correct response
Distractors
Item 4
Item 5

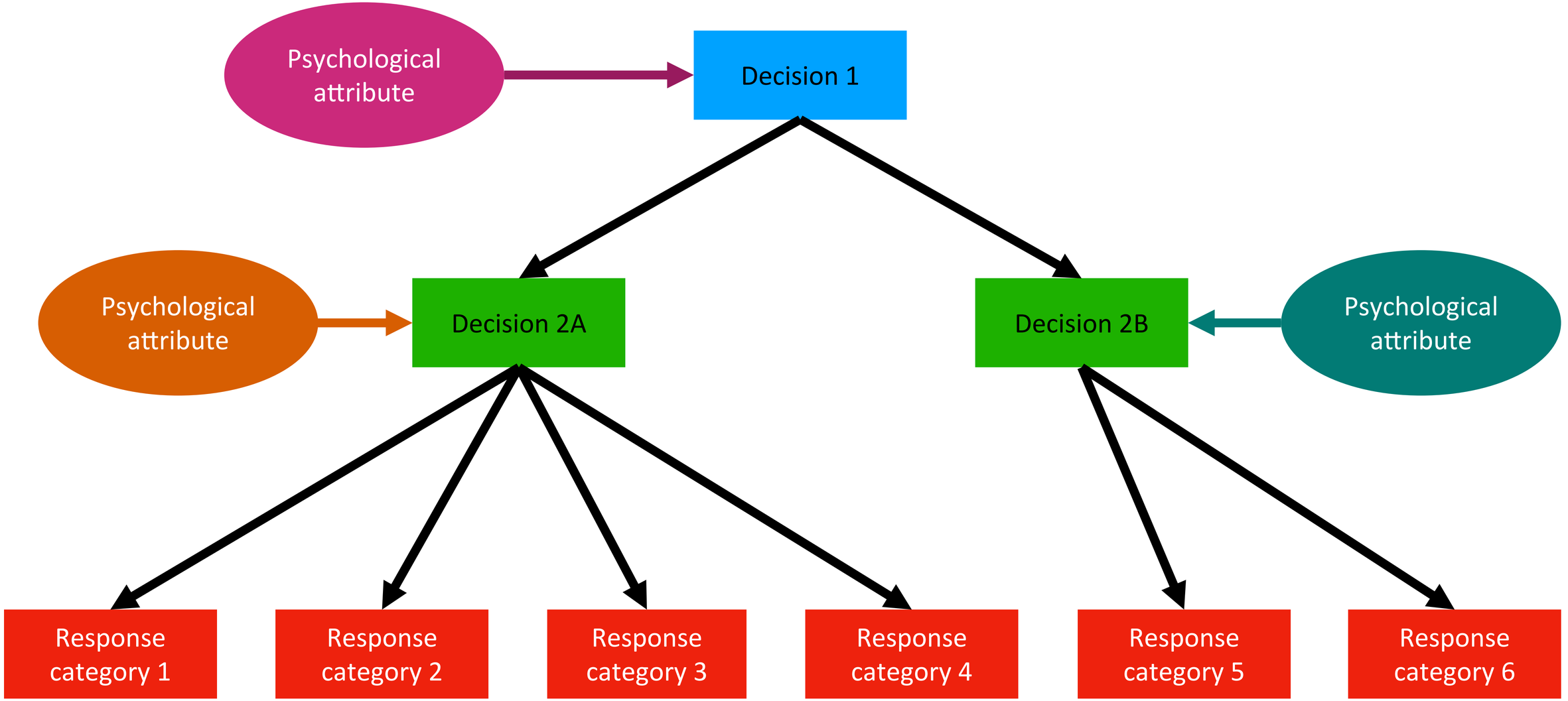
Extending response models

To maximize information

- By using distractor responses, we can…
  - Make tests more reliable at lower levels (great for shorter diagnostic tests)
  - Identify distractors that might deserve partial credit
  - Identify ambiguous items

Disentangling response processes: IRTree models

Breaking down test responses

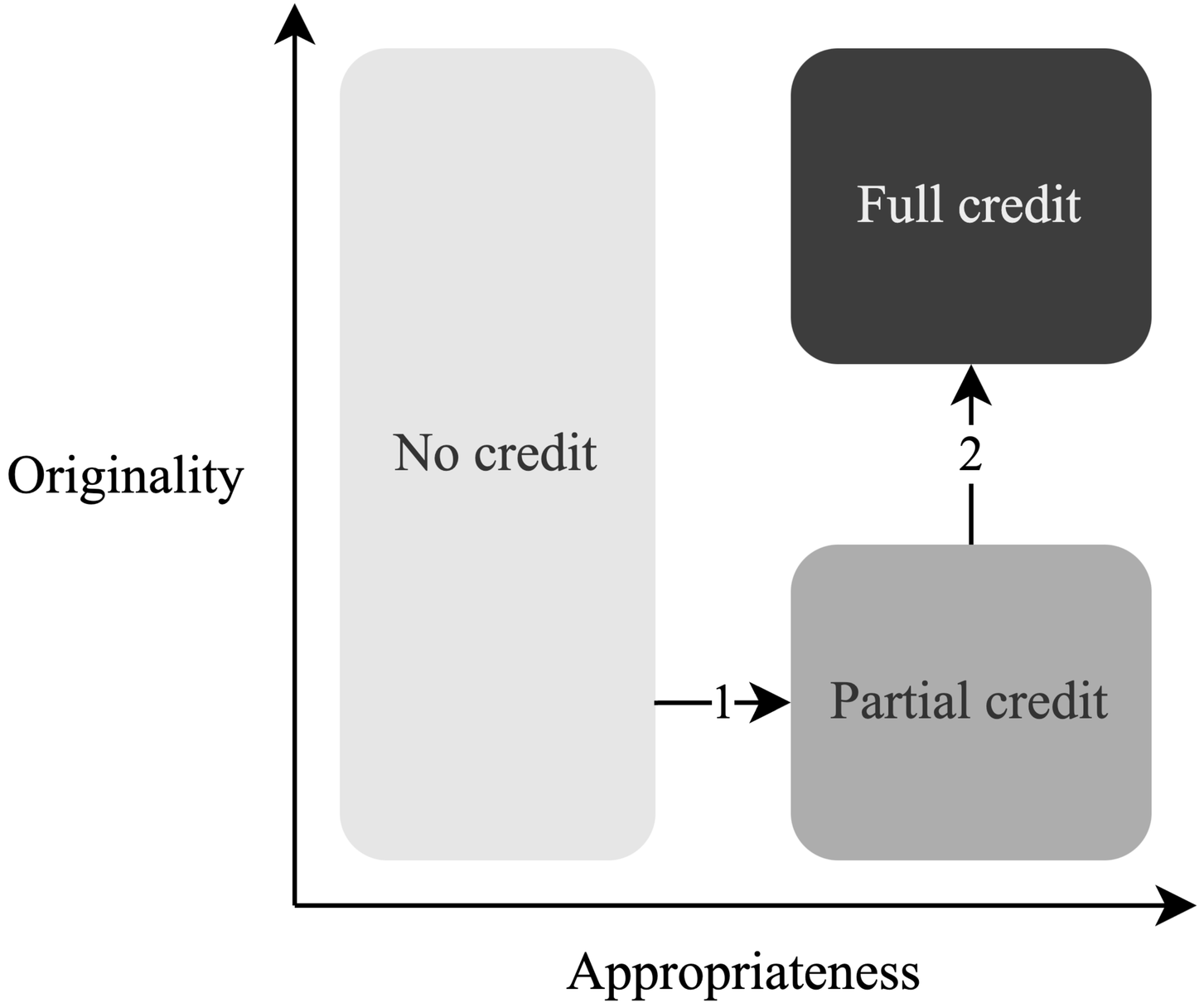


IRT + Decision tree = IRTree

(De Boeck & Partchev, 2012)

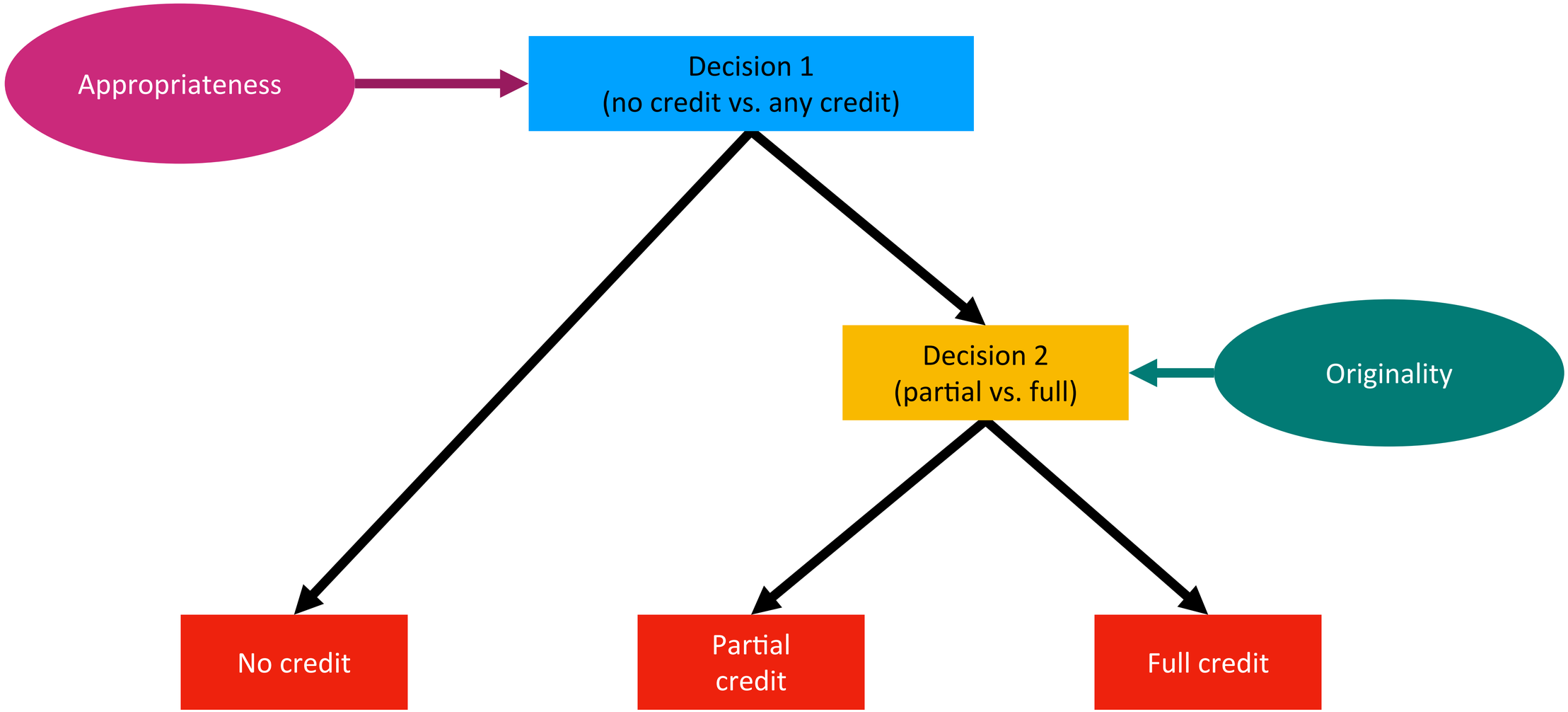
Example 1 : Disentangling ratings
(Myszkowski & Storme, 2025)

- IRT model proposition for OECD's Programme for International Student Assessment (PISA) Creative Thinking 2022 data
- Students produce a creative solution to open-ended problems (e.g., imagine a dialogue for a situation)
- Creativity rated :
  - 0 : No credit
  - 1: Partial credit
  - 2: Full credit
- Treated as ordinal in PISA scorings.

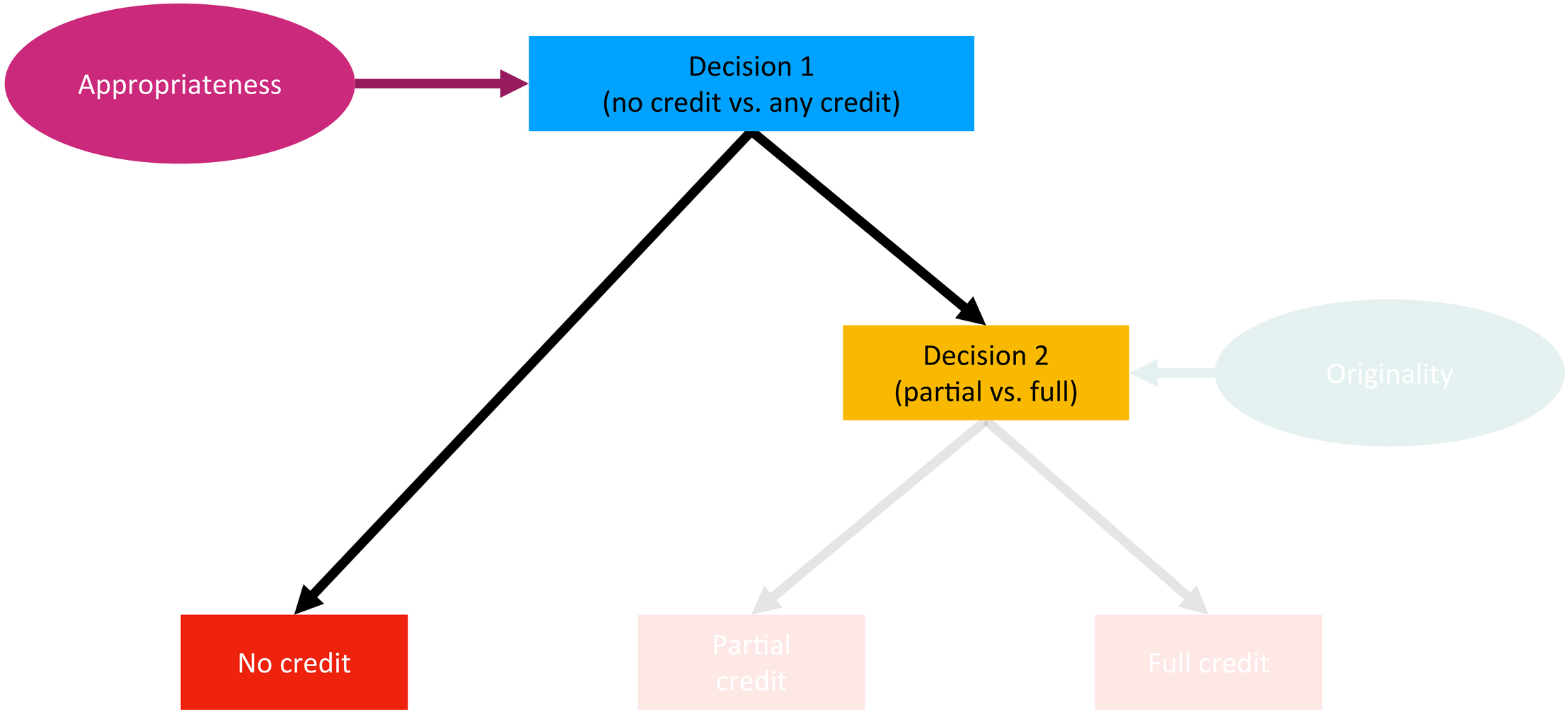


Myszkowski & Storme, 2025

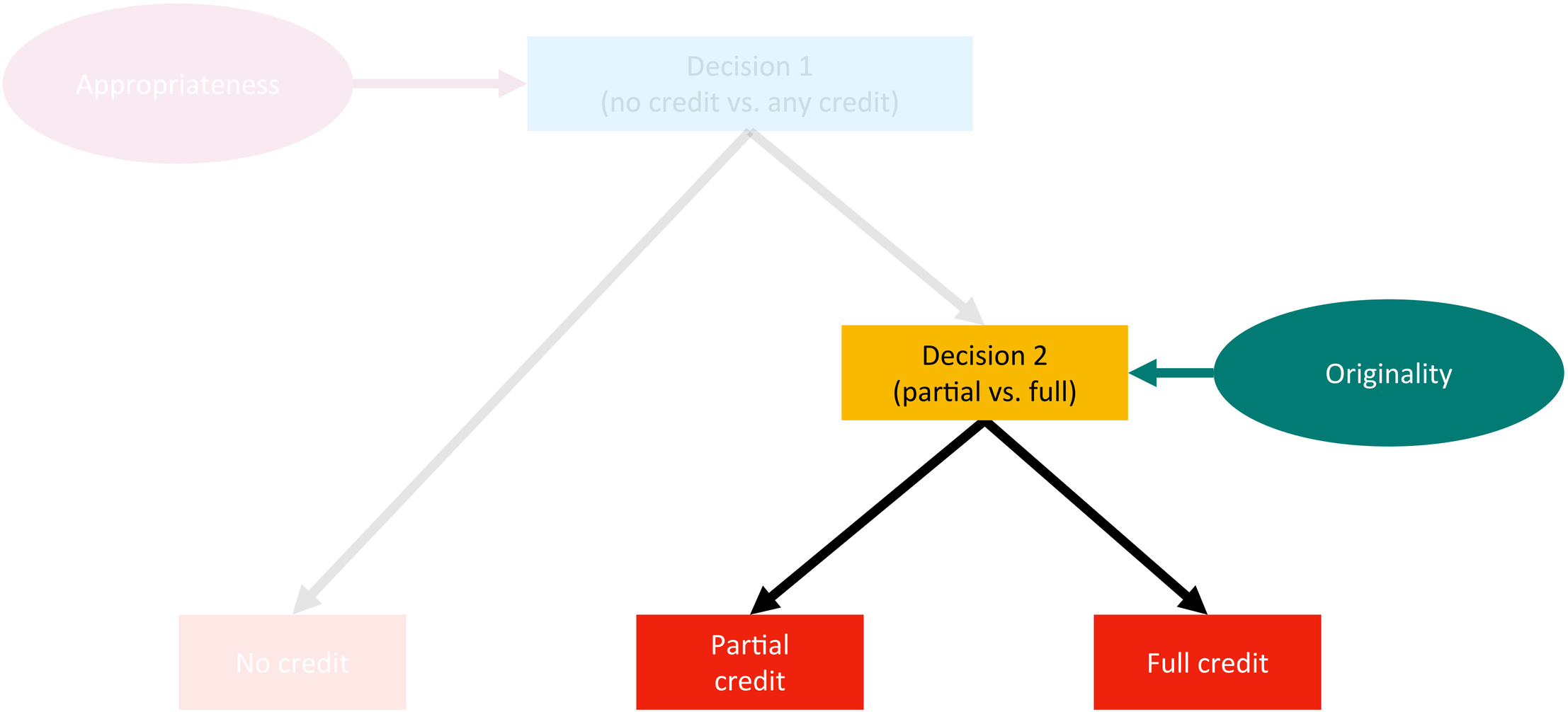
Disentangling creativity components



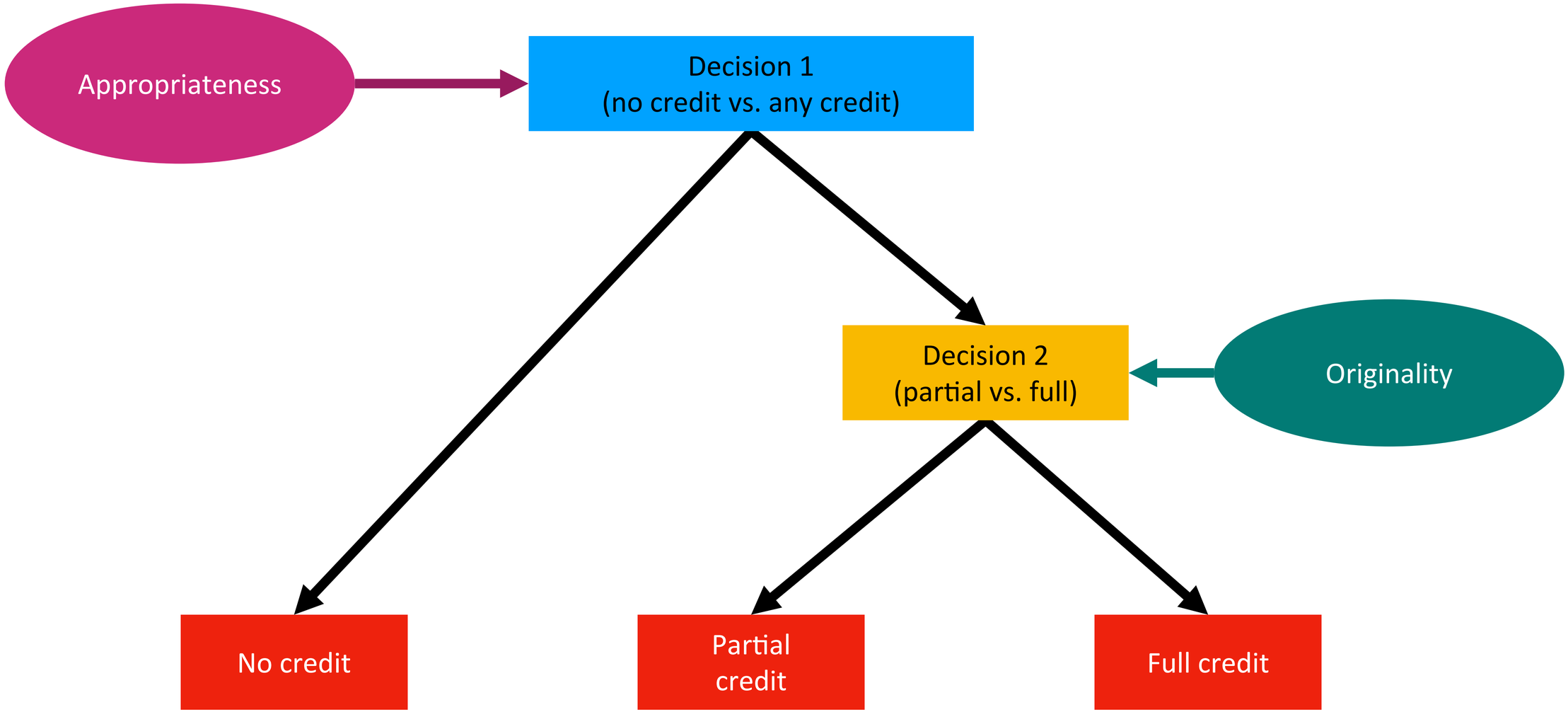
Disentangling creativity components



Disentangling creativity components



Disentangling creativity components



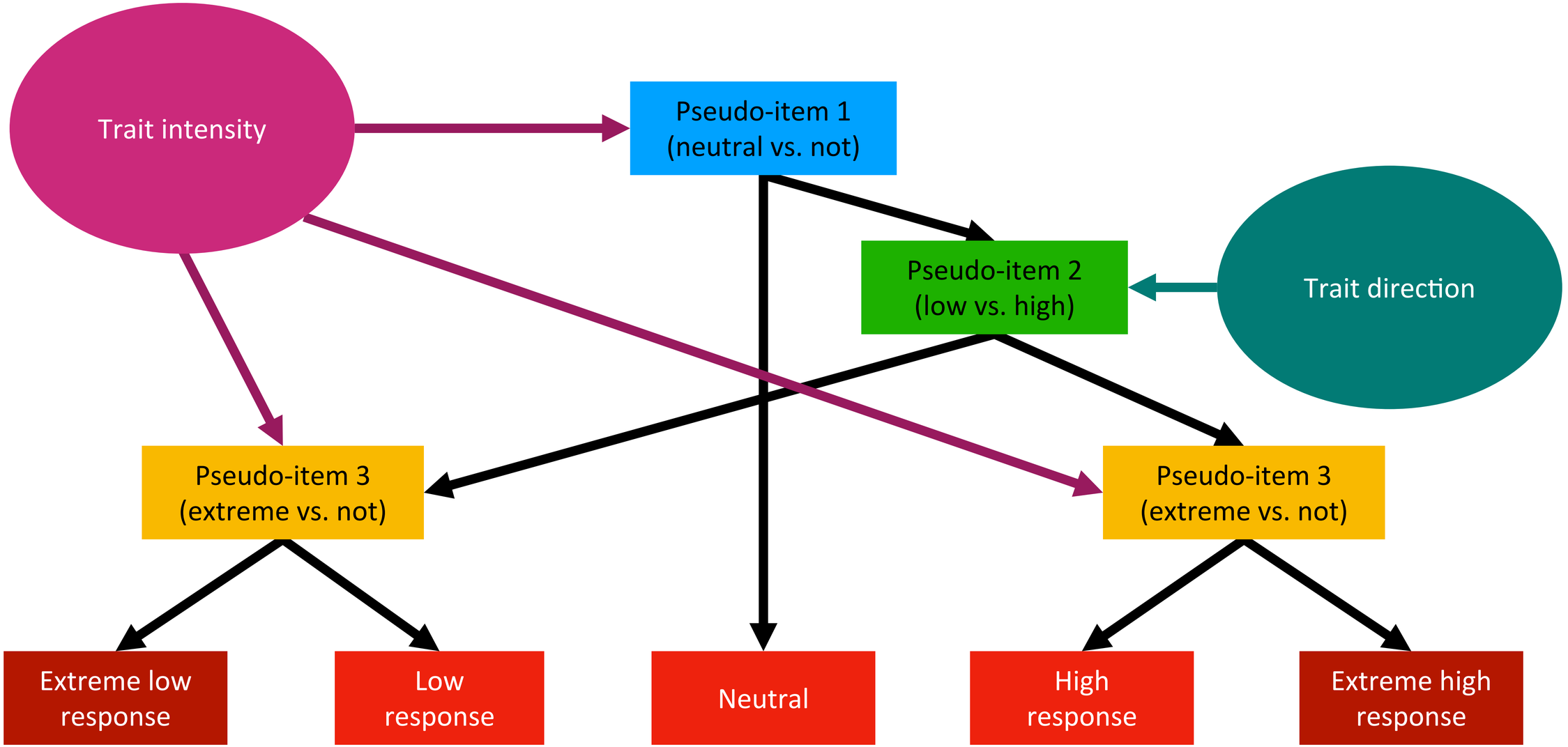
Advantages

- More realistic (we don’t pretend to have measured the originality of all products)
- Disentangle 2 traits underlying the ratings (and their structure)
- Better understanding of relations with other variables (e.g., reading ability might be better related to appropriateness than originality)

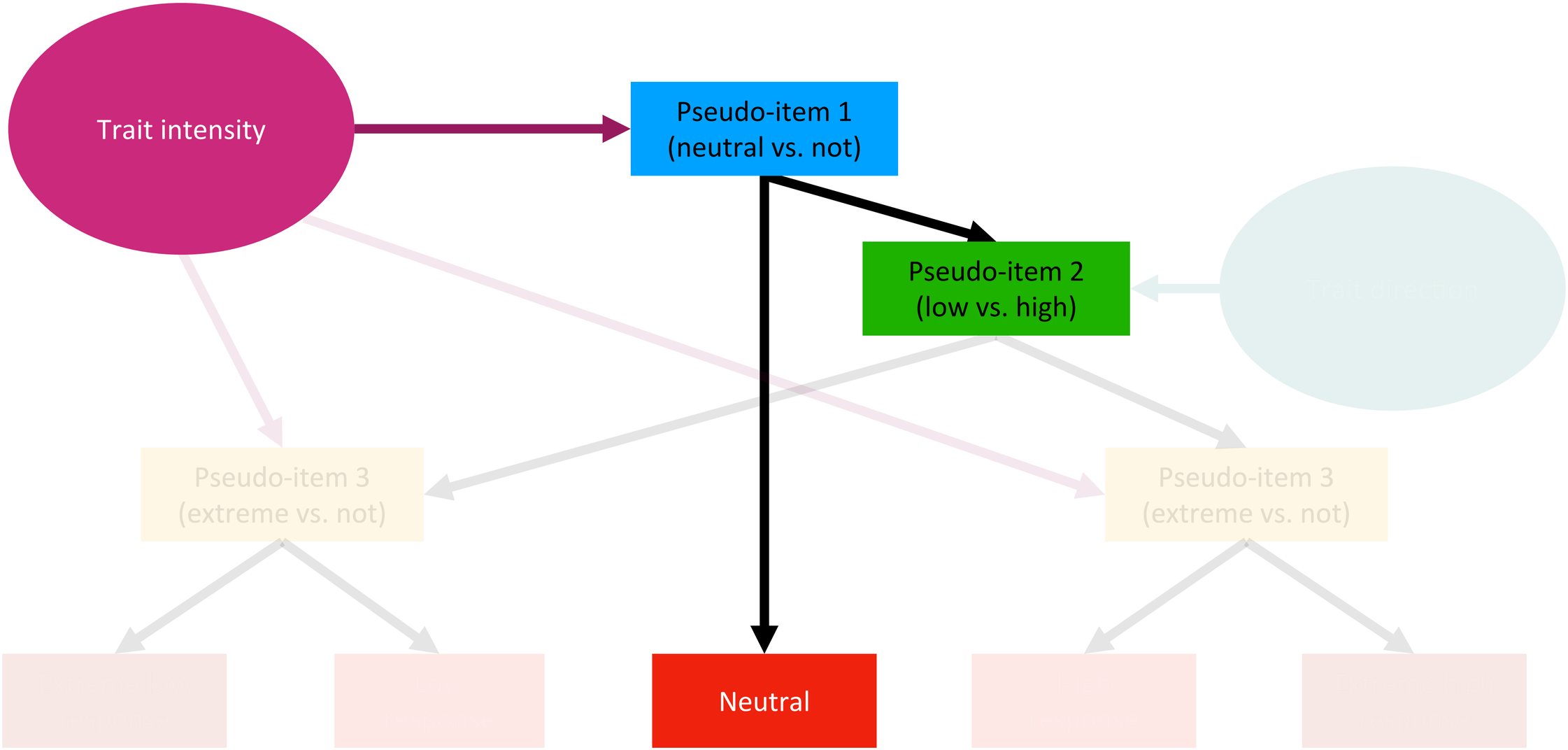
Example 2 : Capturing trait intensity
(Storme, Çelik & Myszkowski, 2020)

- 452 undergraduates
- Big Five Inventory (BFI, John & Srivastava, 1999)
  - Measure of trait levels and trait intensity with the trait variability tree model (TVTM; Lang et al., 2019)
  - Estimated in generalized linear multilevel regression (GLMM) framework with the R package “lme4” (Bates et al., 2015)
- Career adapt-abilities scale (CAAS; Johnston et al., 2013)

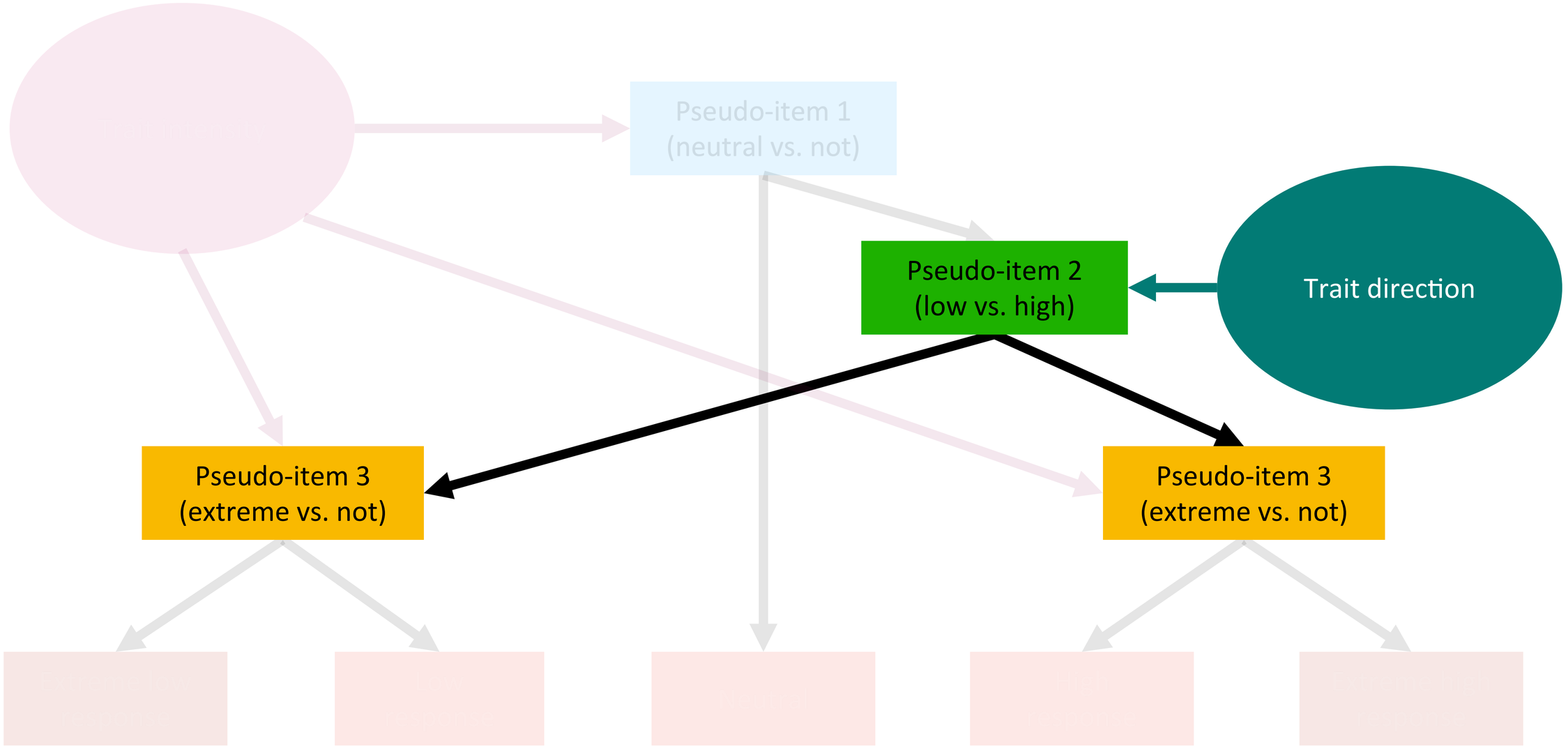
Disentangling intensity and direction



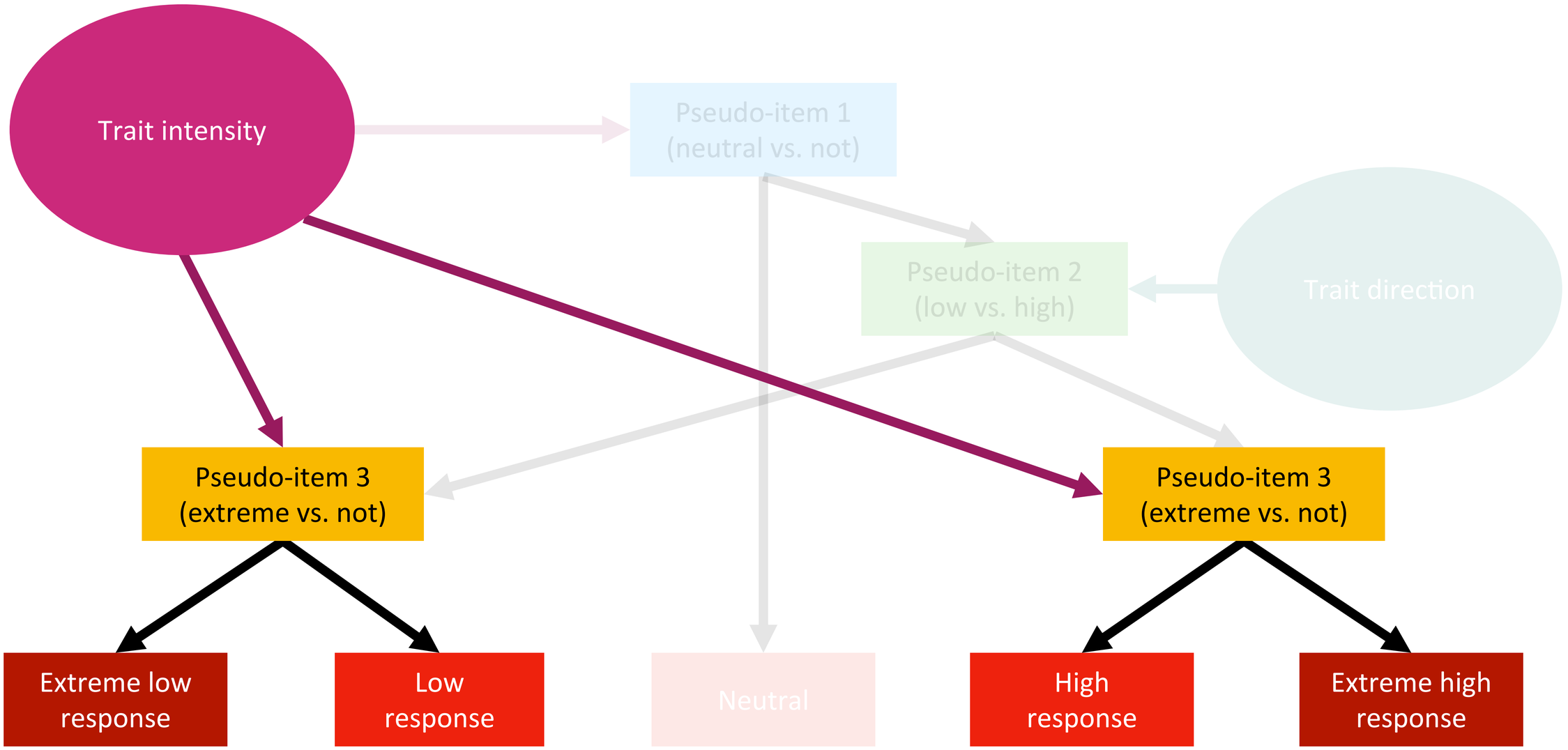
Disentangling intensity and direction



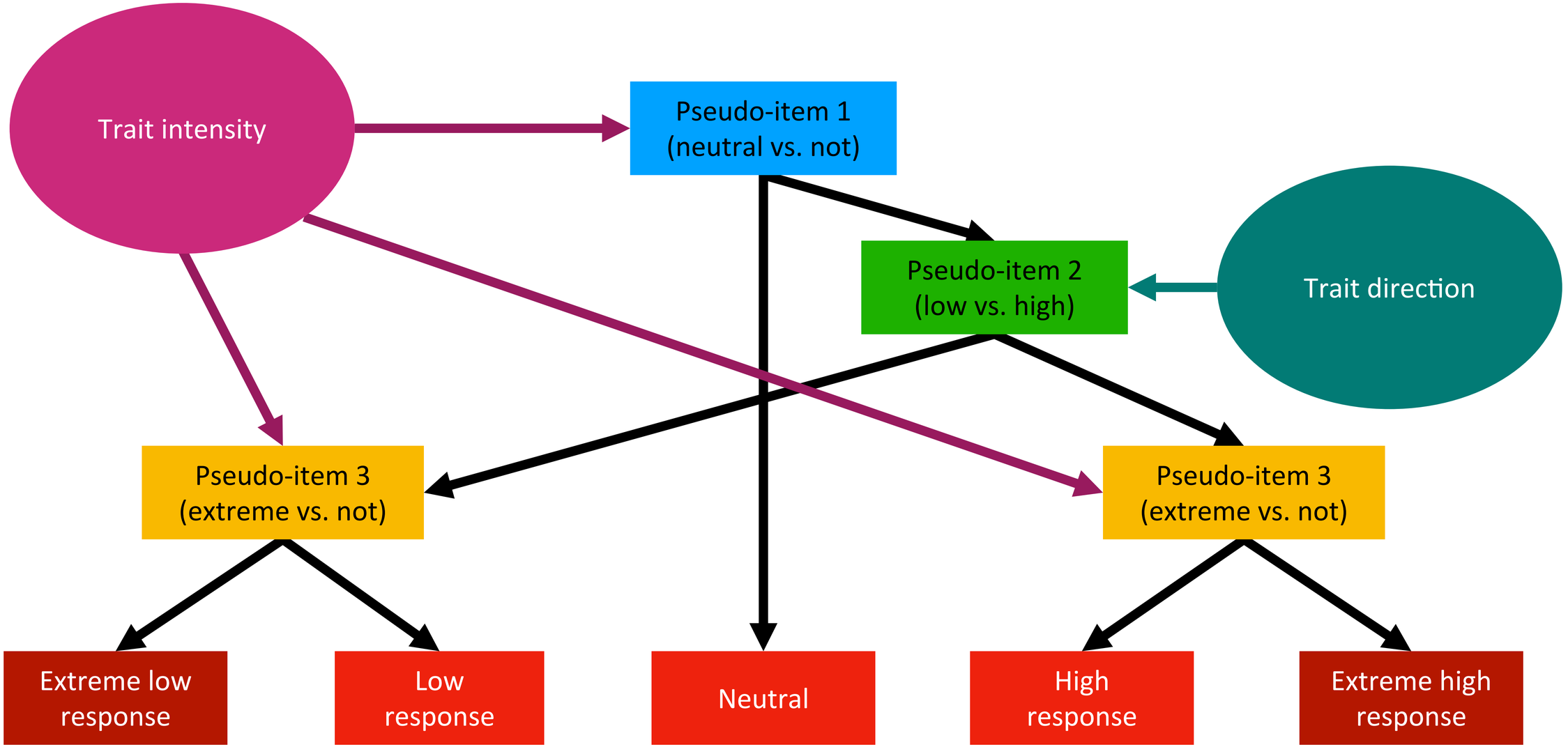
Disentangling intensity and direction



Disentangling intensity and direction



Disentangling intensity and direction



Trait intensity

- Unidimensional across Big Five traits and predicts:
  - career adaptability ($r = .31$) (Storme, Çelik & Myszkowski, 2020)
  - job performance ($r = .19$) (Lievens et al., 2018)
  - effectiveness in negotiations (Çelik, Storme & Myszkowski, 2022)
- Just extreme response tendency?
- A covert measure of:
  - confidence / decisiveness / emotional expressiveness
  - vs. caution / ambivalence / suppression of traits

Going beyond answers:
Joint hierarchical models

Test performance as an ensemble of behaviors

Test responses
(Pass/Fail, Likert scale, etc.)

Test performance as an ensemble of behaviors

Test behaviors

Physiological measures

Test responses
(Pass/Fail, Likert scale, etc.)

Skipping/returning to items

Pointer movements
and keystrokes

Eye movements

Response times

...

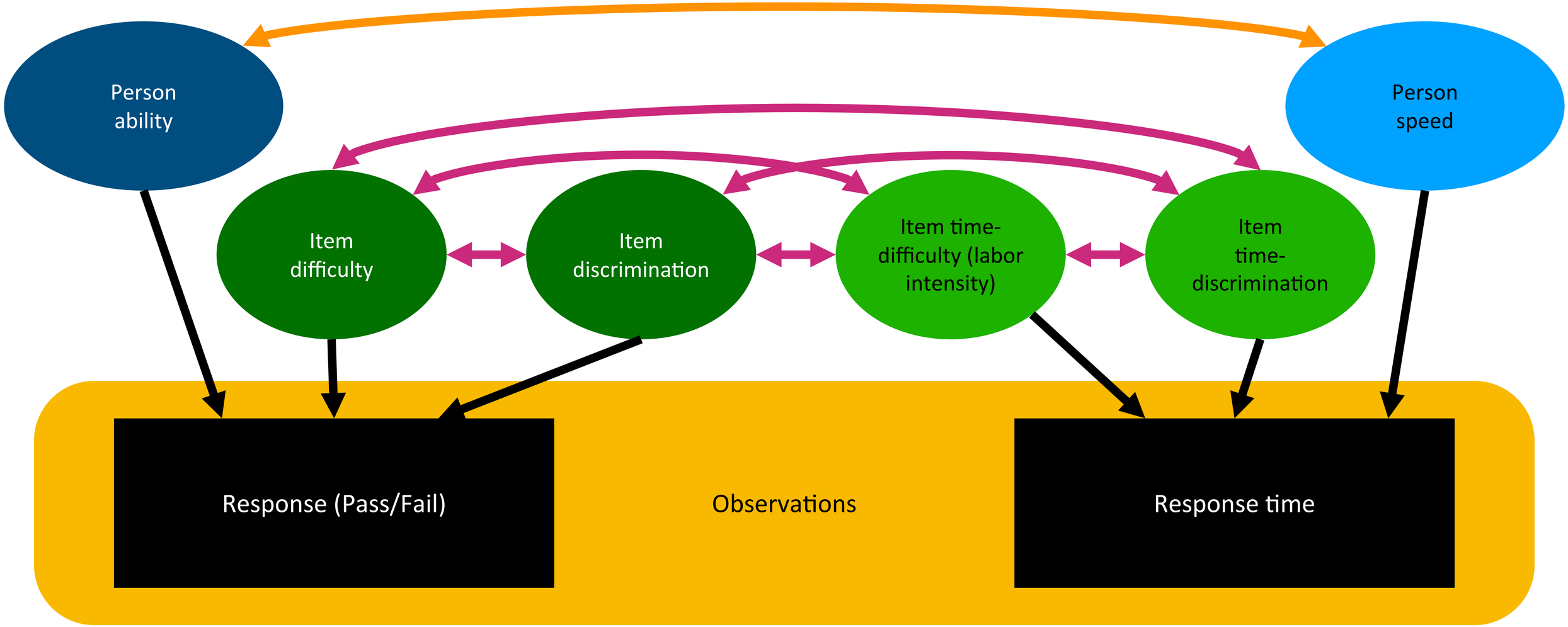
Example 1 : Speed-accuracy trade-off
(Myszkowski, 2019)

- Revised version (Myszkowski & Storme, 2017) of the Visual Aesthetic Sensitivity Test (Götz, 1985)
- 201 undergraduate students
- Collected responses and response times
- Non-speeded test
- Hypothesized that faster respondents are worse performers

X

C

Joint modeling approach
(van der Linden & Fox, 2016)



Joint modeling approach
(van der Linden & Fox, 2016)

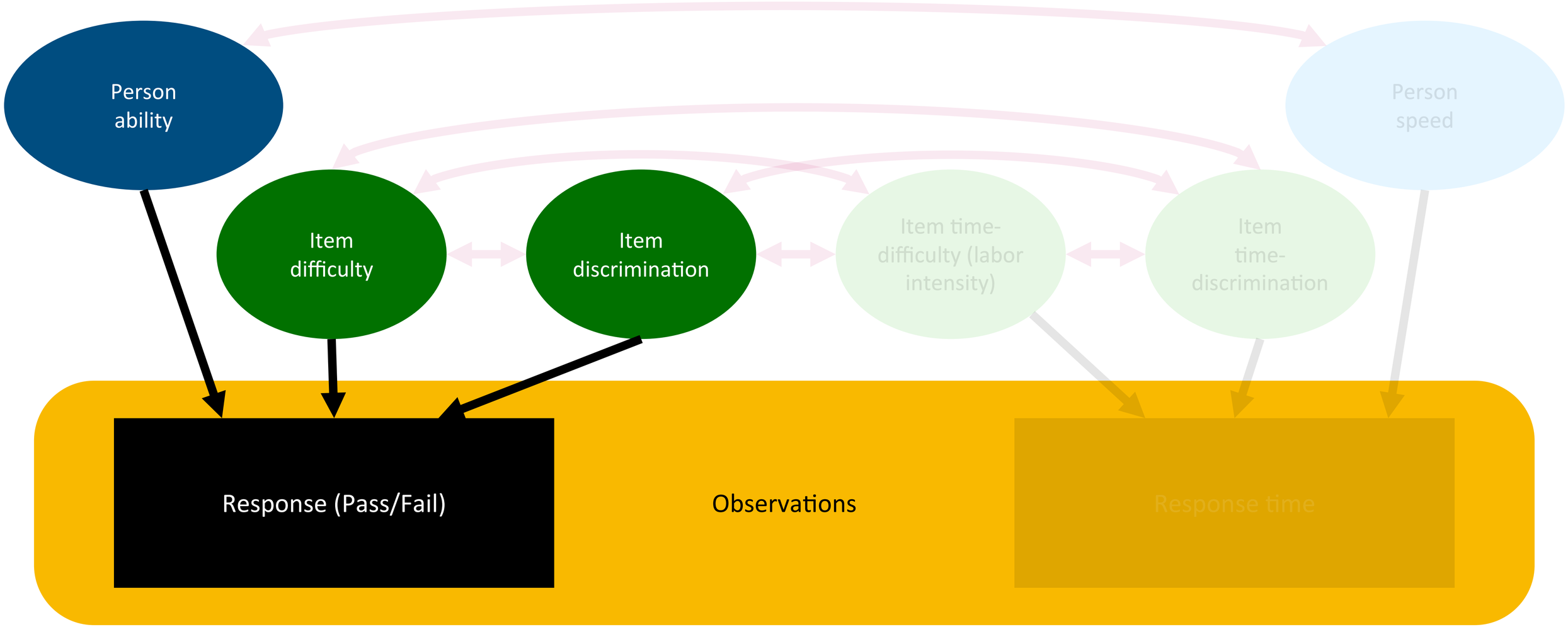


Table 1 Overview of item response models available for different item types

Item type	Example distribution	Example link	Example inverse	Example
Continuous unbounded	Gaussian	Identity	Ident	
Binary	Bernoulli	Logit	Logistic	1PL, 2PL, 3PL
Count (discrete with lower bound)	Poisson	Logarithm	Exponential	RPCM 2PPCM
Response time (continuous with lower bound)	Log-normal	Logarithm	Exponential	LNIRT models
Visual analog scale (continuous with lower and upper bound)	Beta	Logit	Logistic	BRM-1, BRM-2

Responses

Note: 1PL: 1-parameter logistic; 2PL: 2-parameter logistic; 3PL: 3-parameter logistic; RPCM: Rasch Poisson counts model; 2PPCM: 2-parameter Poisson counts model; LNIRT: log-normal item response theory; BRM-1: beta response model 1; BRM-2: beta response model 2.

Myszkowski (2024)

Joint modeling approach
(van der Linden & Fox, 2016)

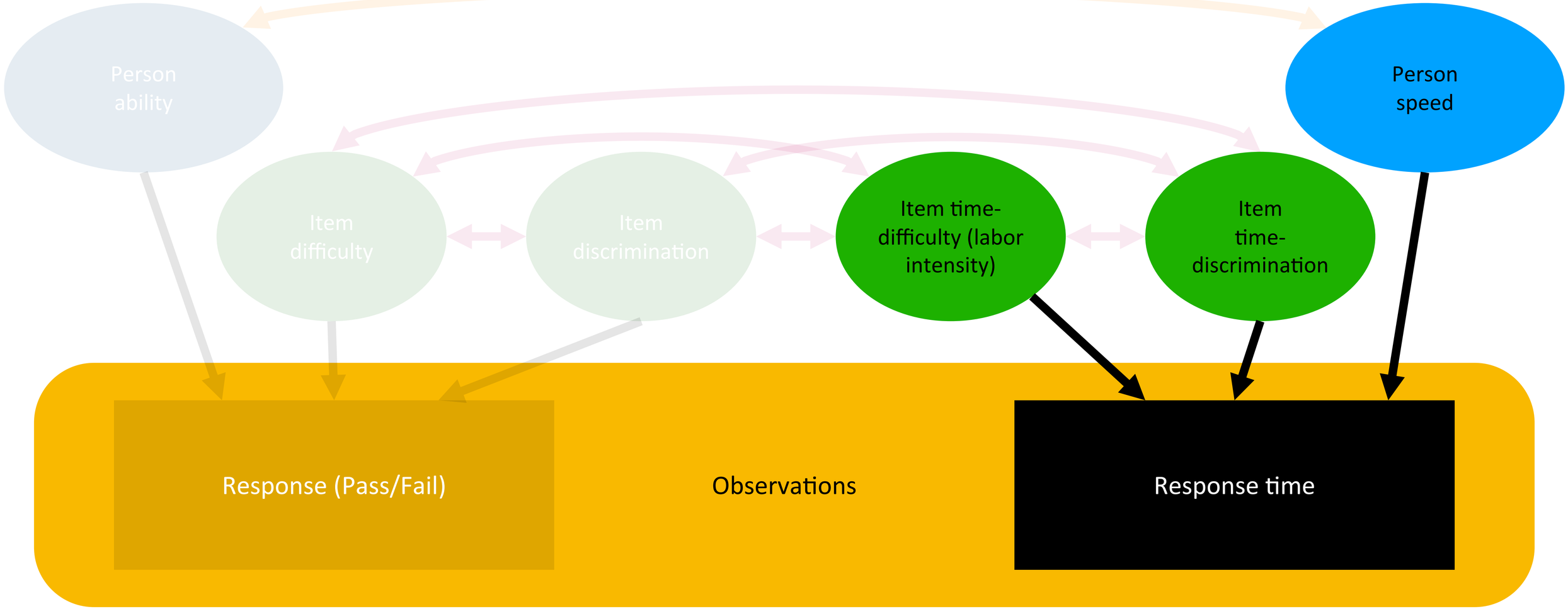


Table 1 Overview of item response models available for different item types

Item type	Example distribution	Example link	Example inverse link	Example models
Continuous unbounded	Gaussian	Identity	Identity	Congeneric, parallel
Binary	Bernoulli	Logit	Logistic	1PL, 2PL, 3PL
Count (discrete with lower bound)	Poisson	Logarithm	Exponential	RPCM 2PPCM
Response time (continuous with lower bound)	Log-normal	Logarithm	Exponential	LNIRT models
Visual analog scale (continuous with lower and upper bound)	Beta	Logit	Logistic	BRM-2

Note: 1PL: 1-parameter logistic; 2PL: 2-parameter logistic; 3PL: 3-parameter logistic; RPCM: Rasch Poisson counts model; 2PPCM: 2-parameter Poisson counts model; LNIRT: log-normal item response theory; BRM-1: beta response model 1; BRM-2: beta response model 2.

Response times

Myszkowski (2024)

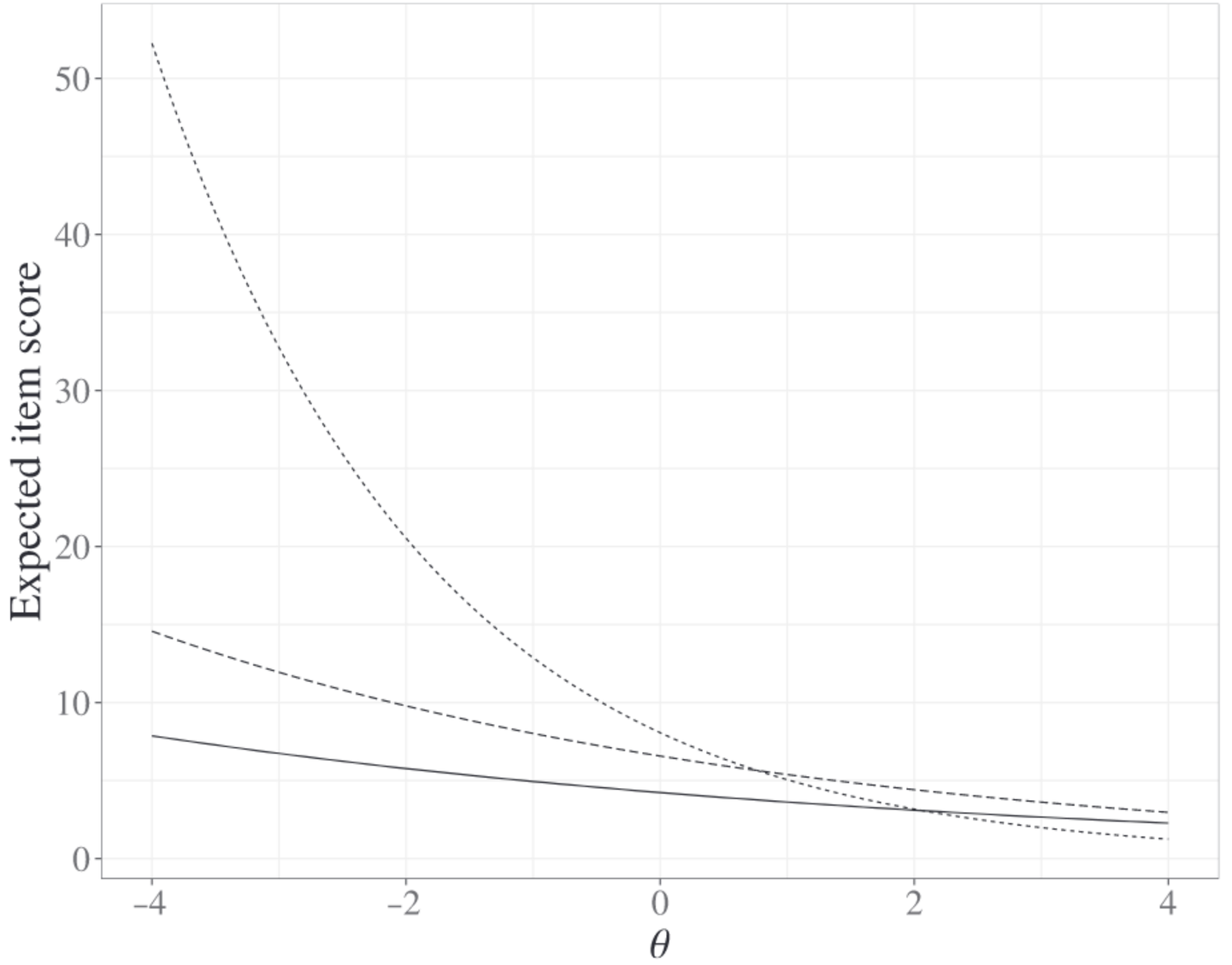
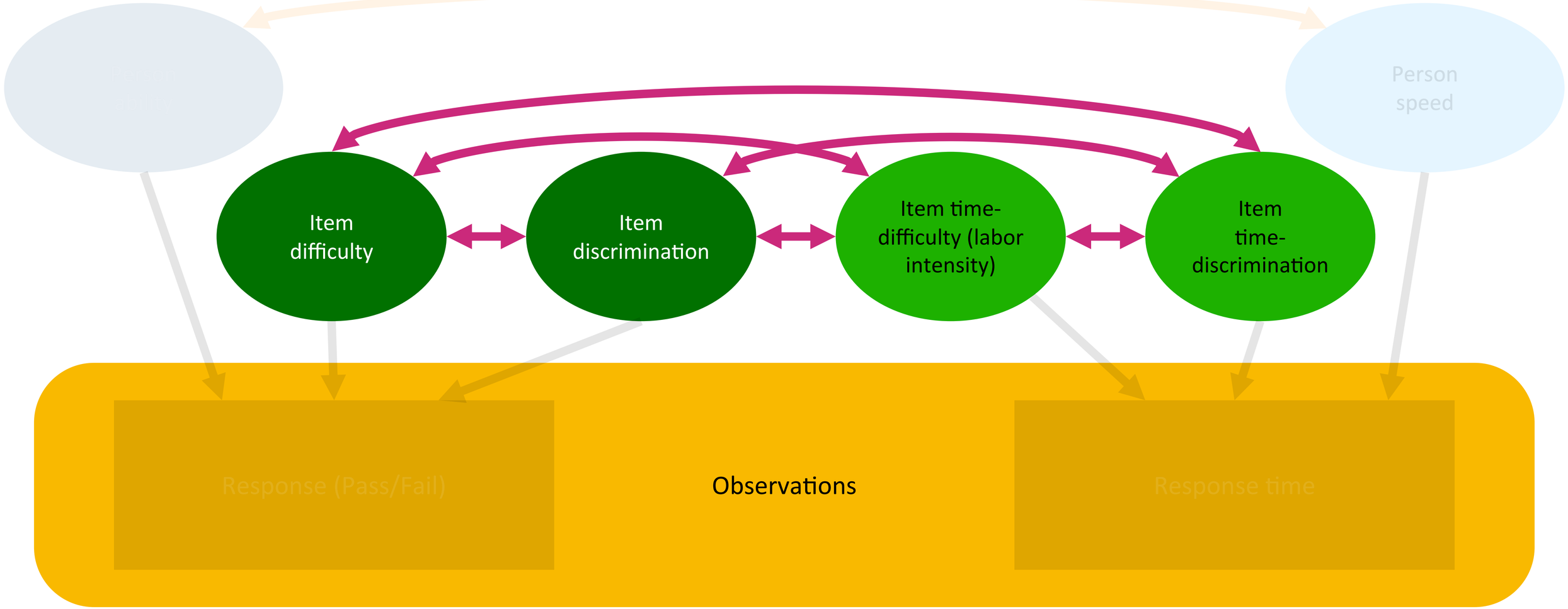


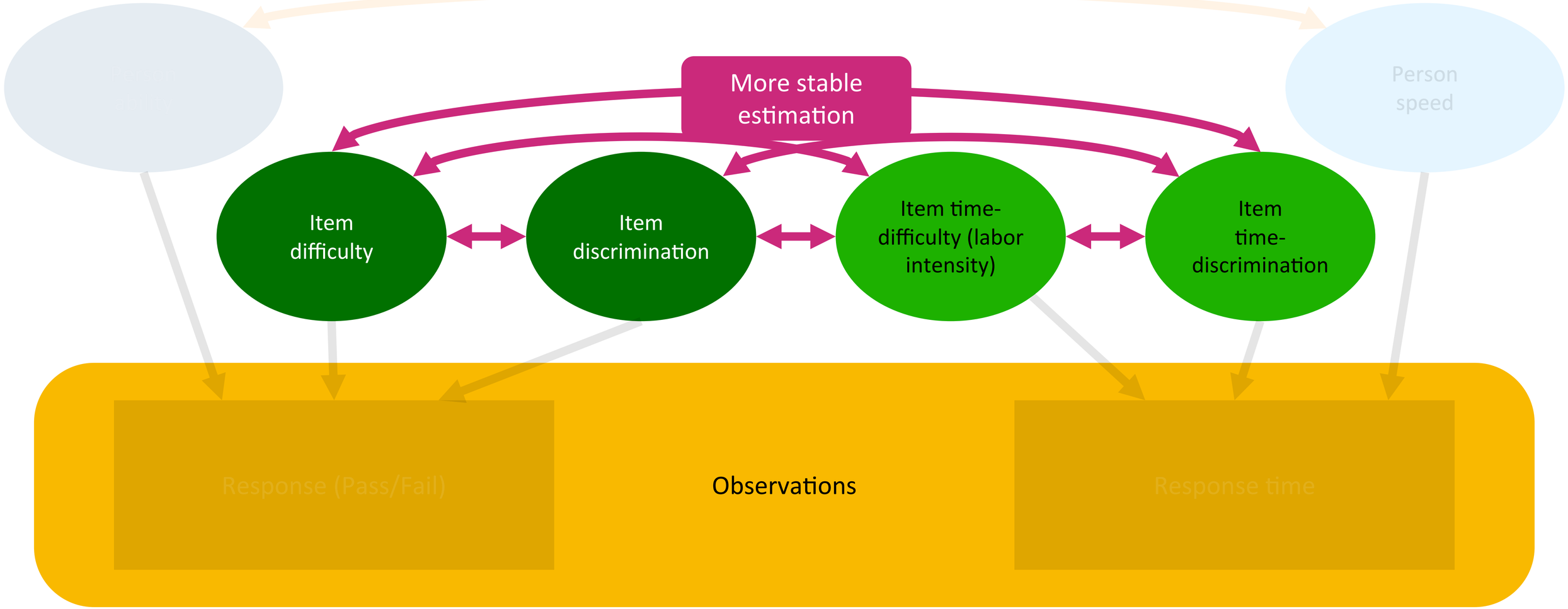
Figure 10 Expected scores of three items modeled with a 2-parameter log-linear model (where the slopes are negative)

Myszkowski, 2024

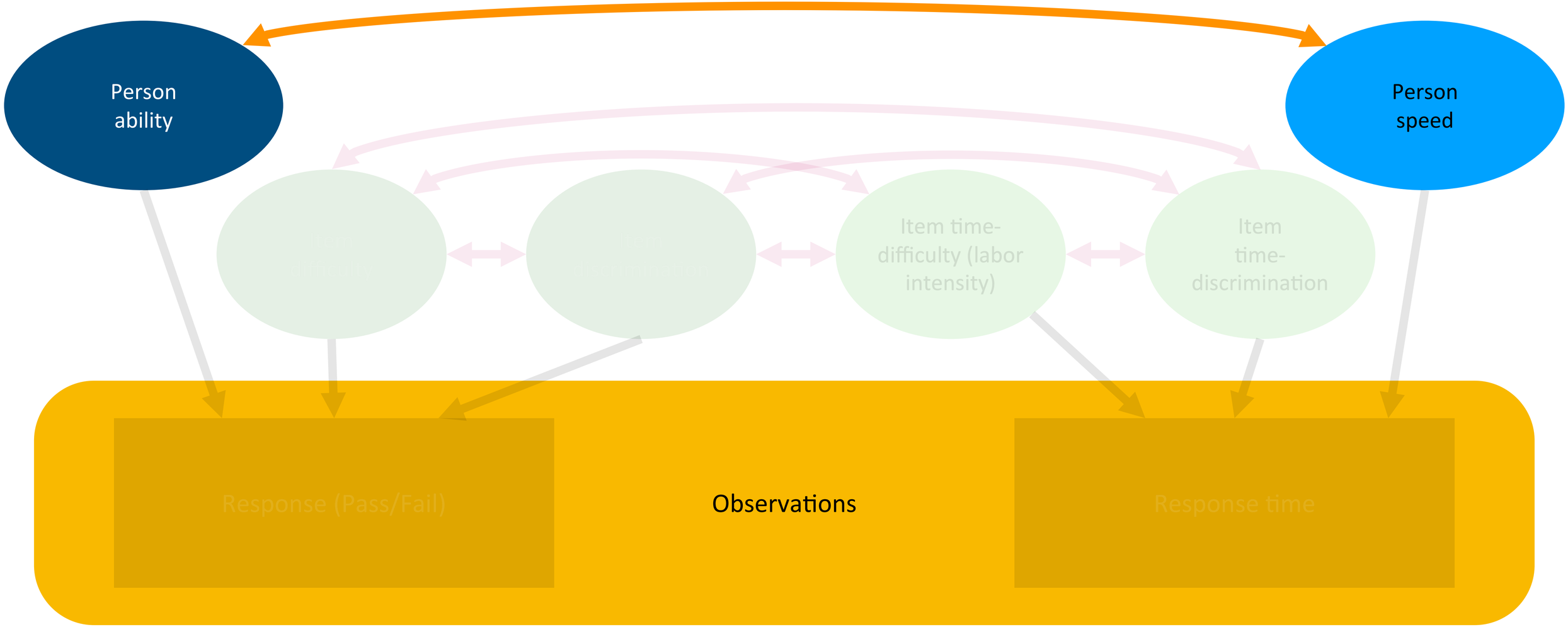
Joint modeling approach
(van der Linden & Fox, 2016)



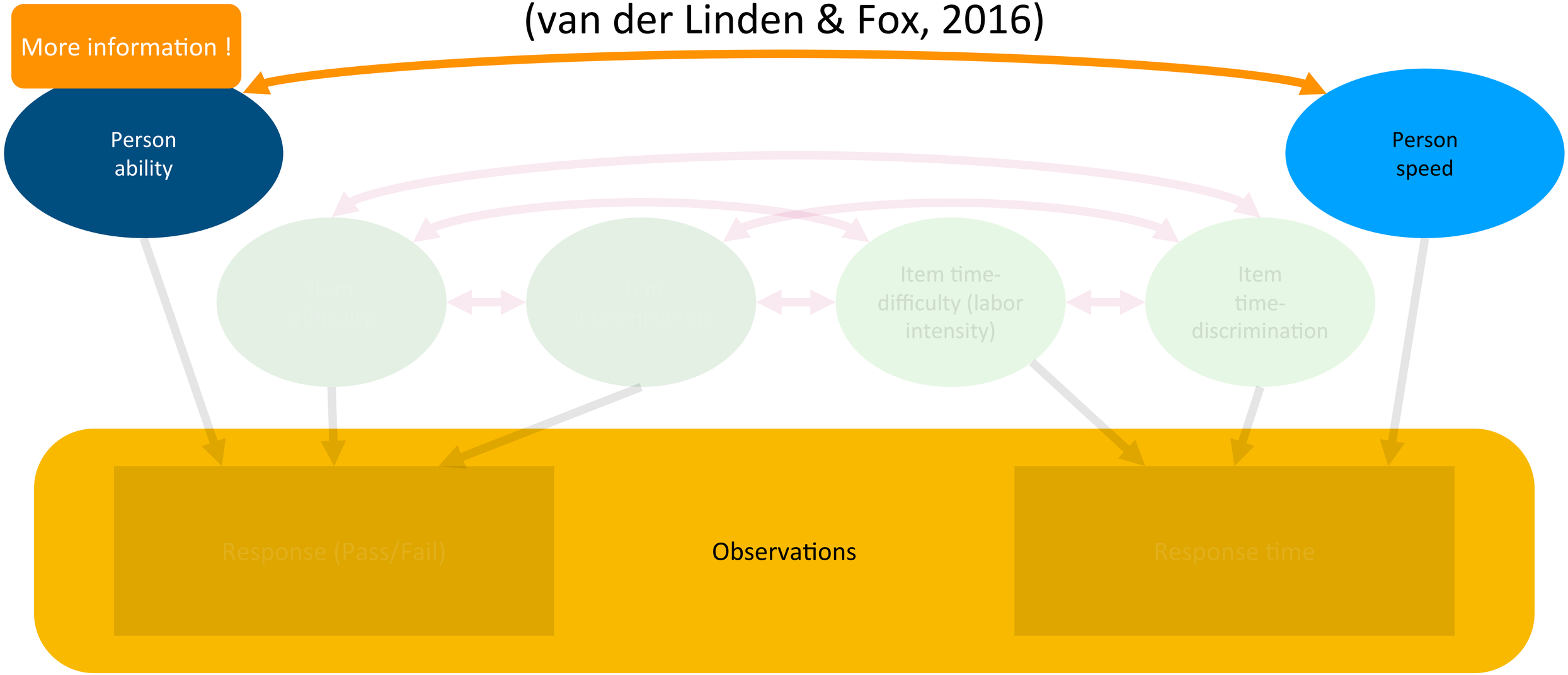
Joint modeling approach
(van der Linden & Fox, 2016)



Joint modeling approach
(van der Linden & Fox, 2016)



Joint modeling approach
(van der Linden & Fox, 2016)



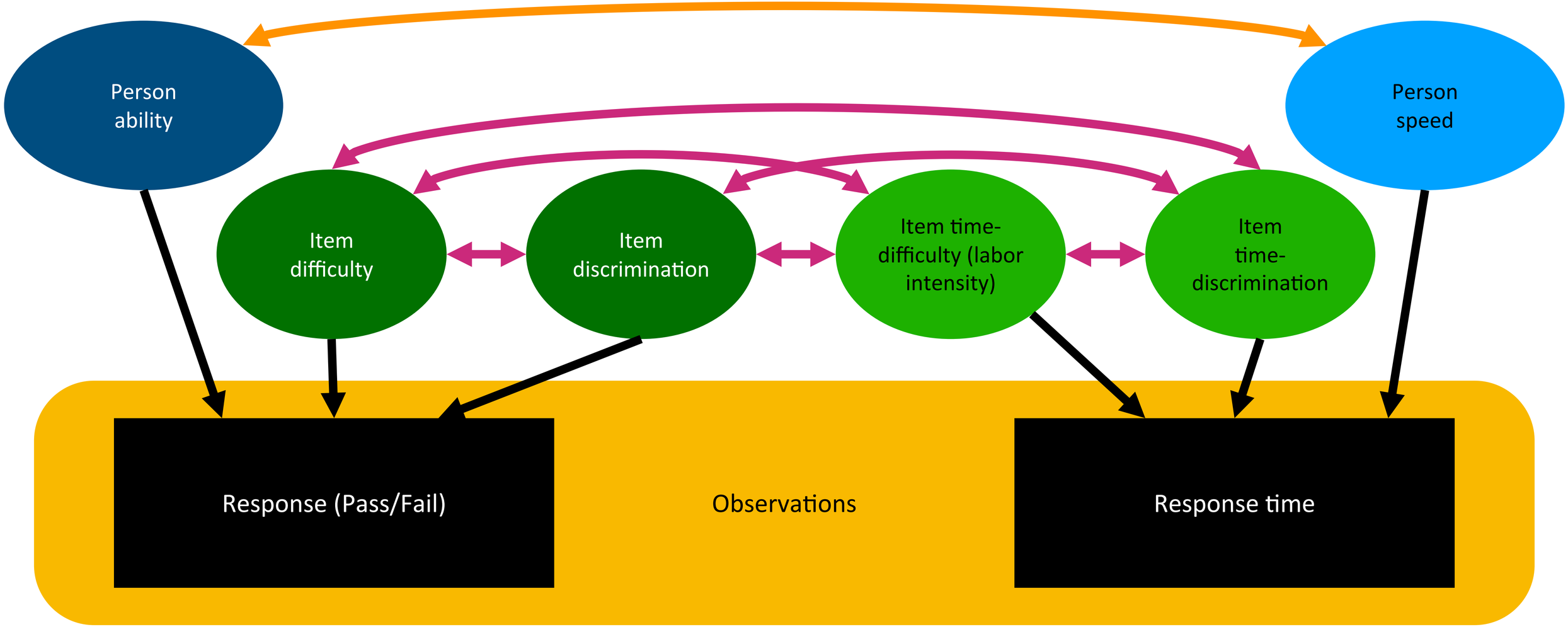
Example 1 : Speed-accuracy trade-offs
(Myszkowski, 2019)

- Responses: 2/3-parameter normal ogive model
- Response times: 2-parameter lognormal model
- Bayesian MCMC estimation using the R package LNIRT (Fox et al., 2007)
- r = -.47 between speed and accuracy (95% HPD [−0.61, −0.32])

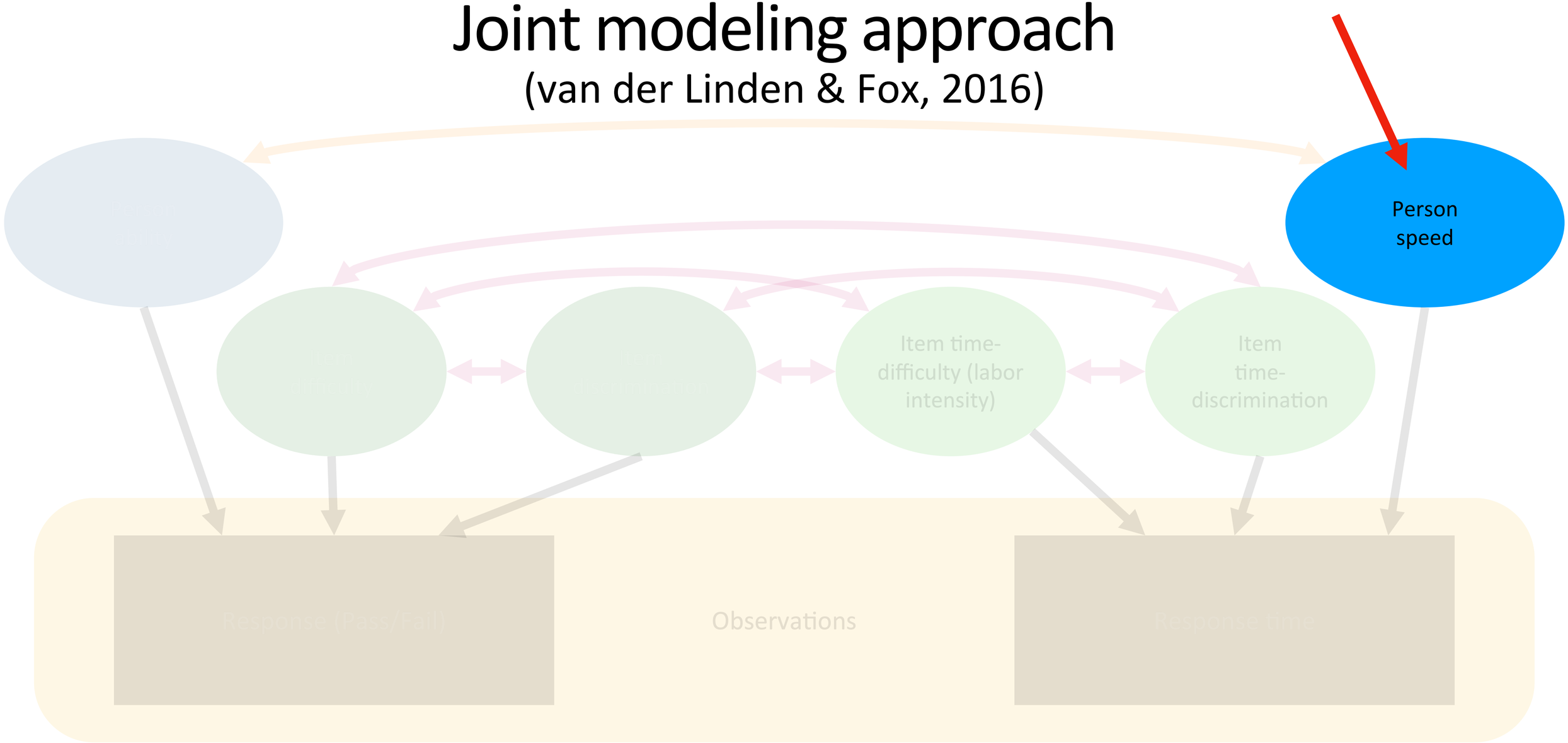
Example 2 : Speed trajectories
(Myszkowski, Storme, Kubiak & Baron, 2022)

- Spatial reasoning test : Progressive matrices generated using IMak (Blum & Holling, 2018)
- Big Five : Synthetic Aperture Personality Assessment (SAPA) (Revelle et al., 2010)
- Analysis :
  - 1/ Variable speed joint IRT model using LNIRT (Fox et al., 2007) for R
  - 2/ Latent profile analysis of latent speed parameters using tidyLPA (Rosenberg et al., 2018) for R

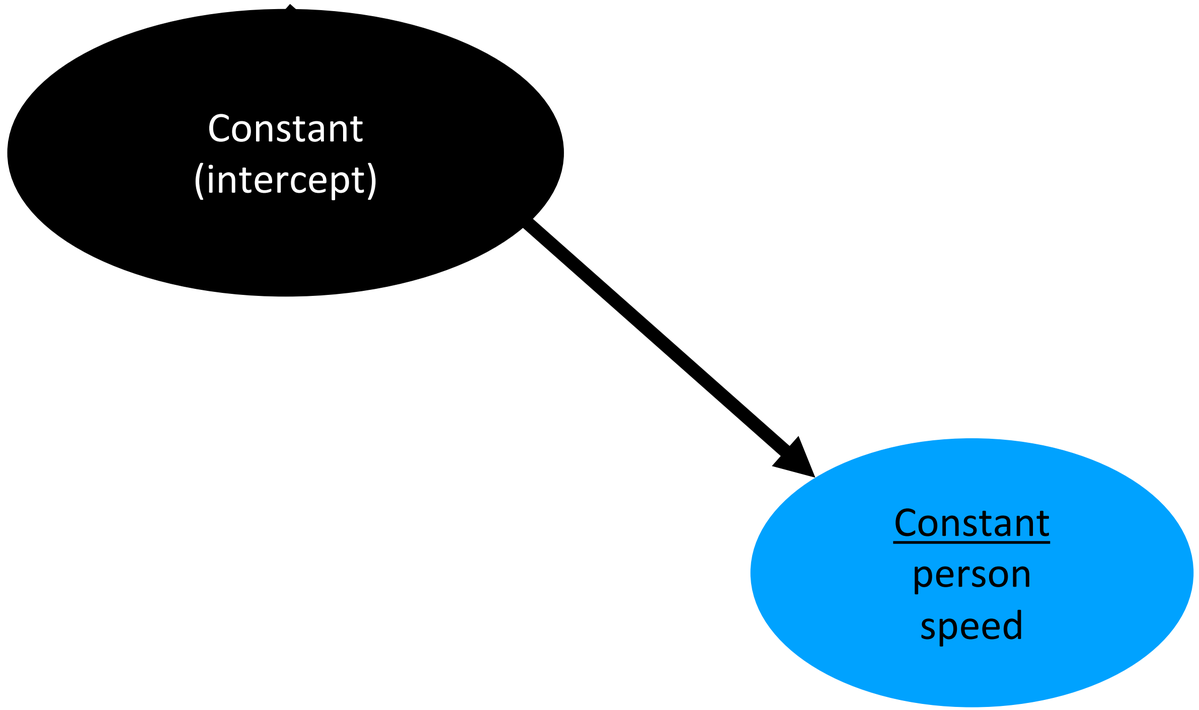
Joint modeling approach
(van der Linden & Fox, 2016)



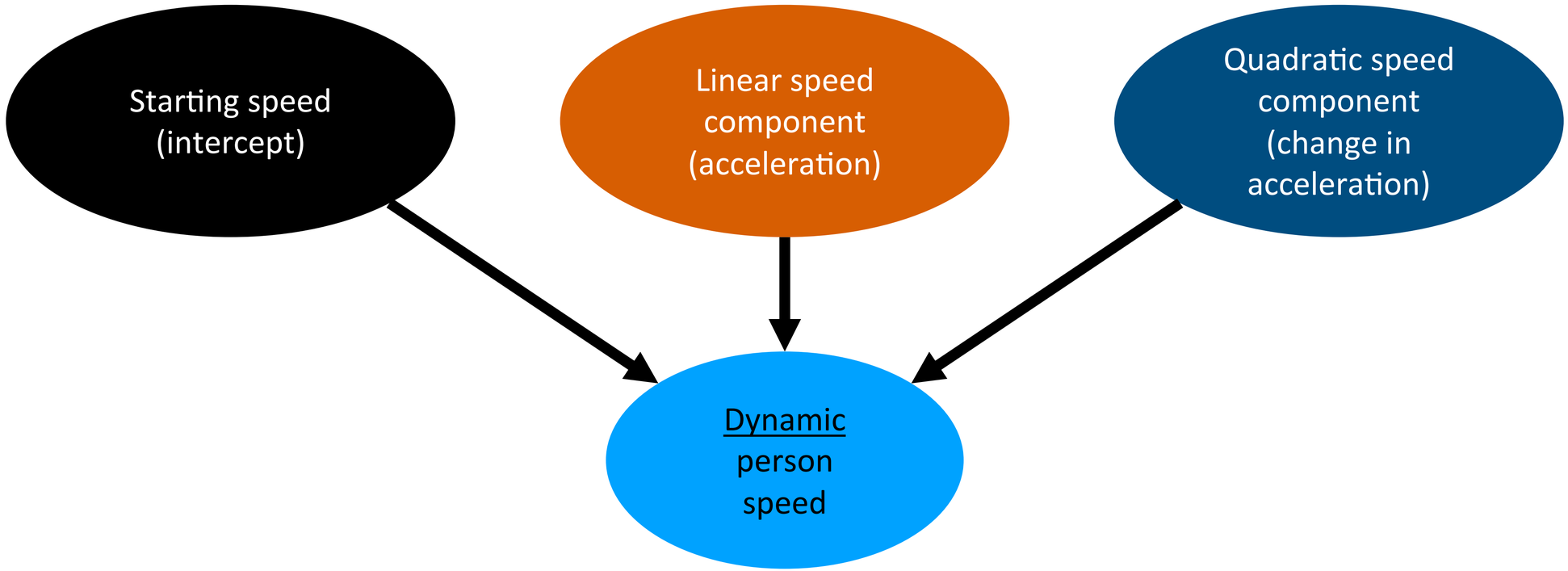
Joint modeling approach
(van der Linden & Fox, 2016)



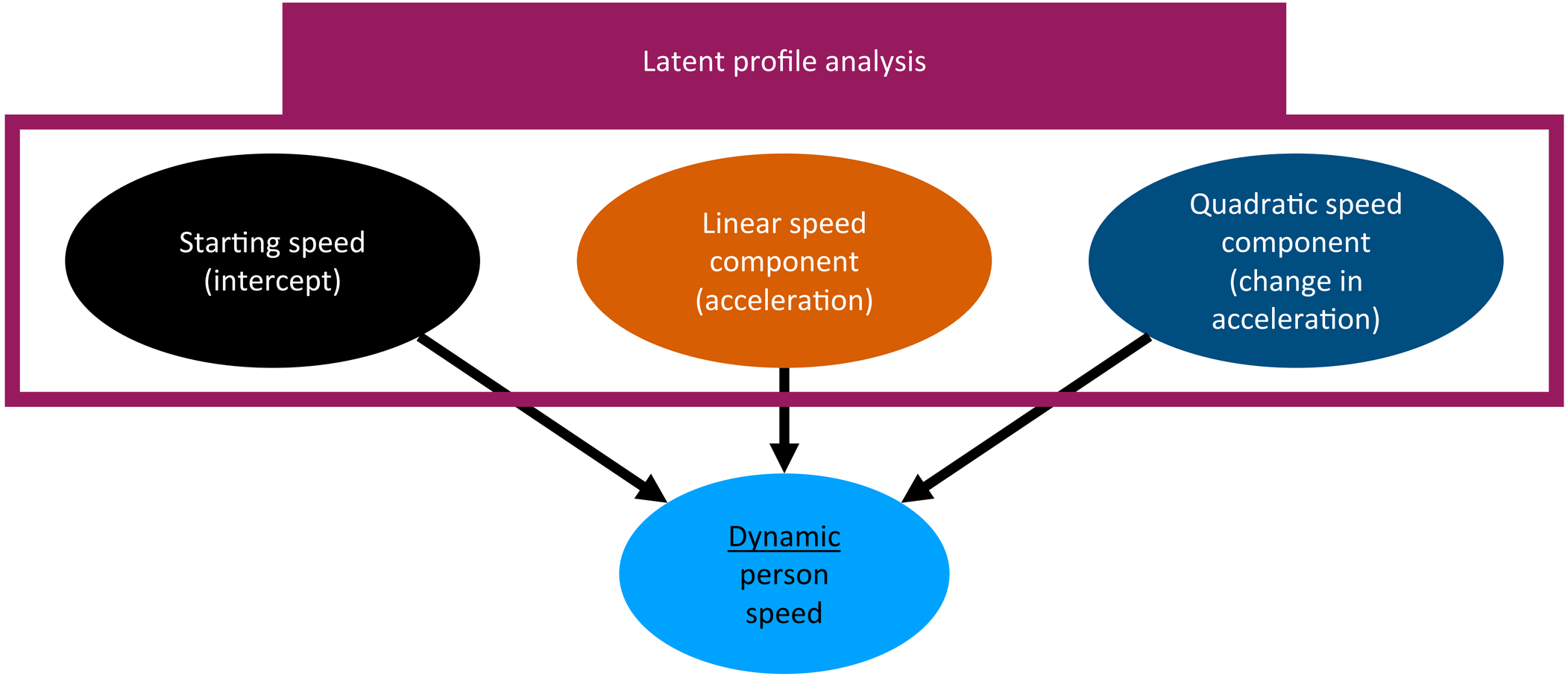
Constant speed model



Dynamic speed model



Dynamic speed model



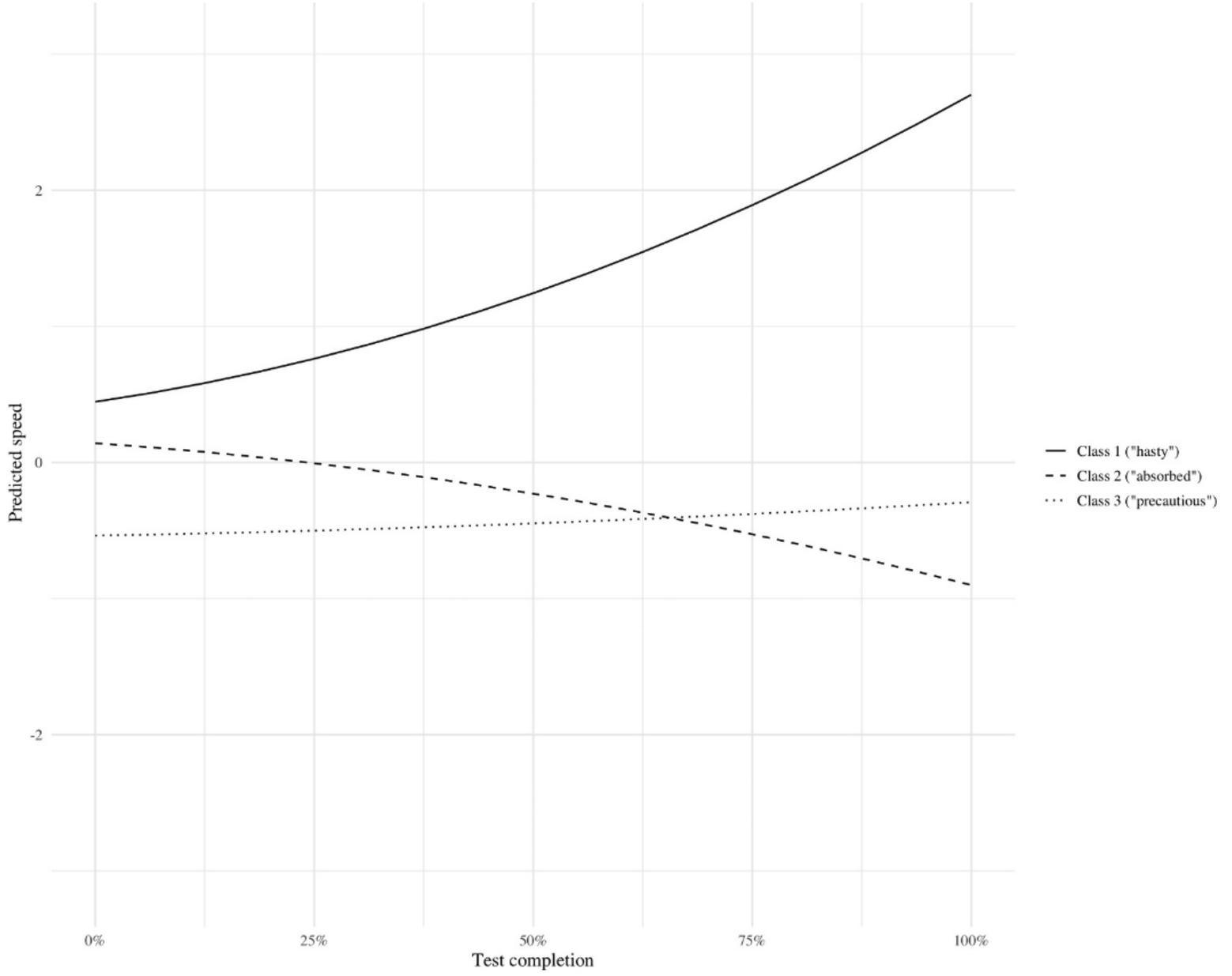


Fig. 4. Predicted speed trajectories of the classes identified with Latent Profile Analysis.

(Myszkowski, Storme, Kubiak & Baron, 2022)

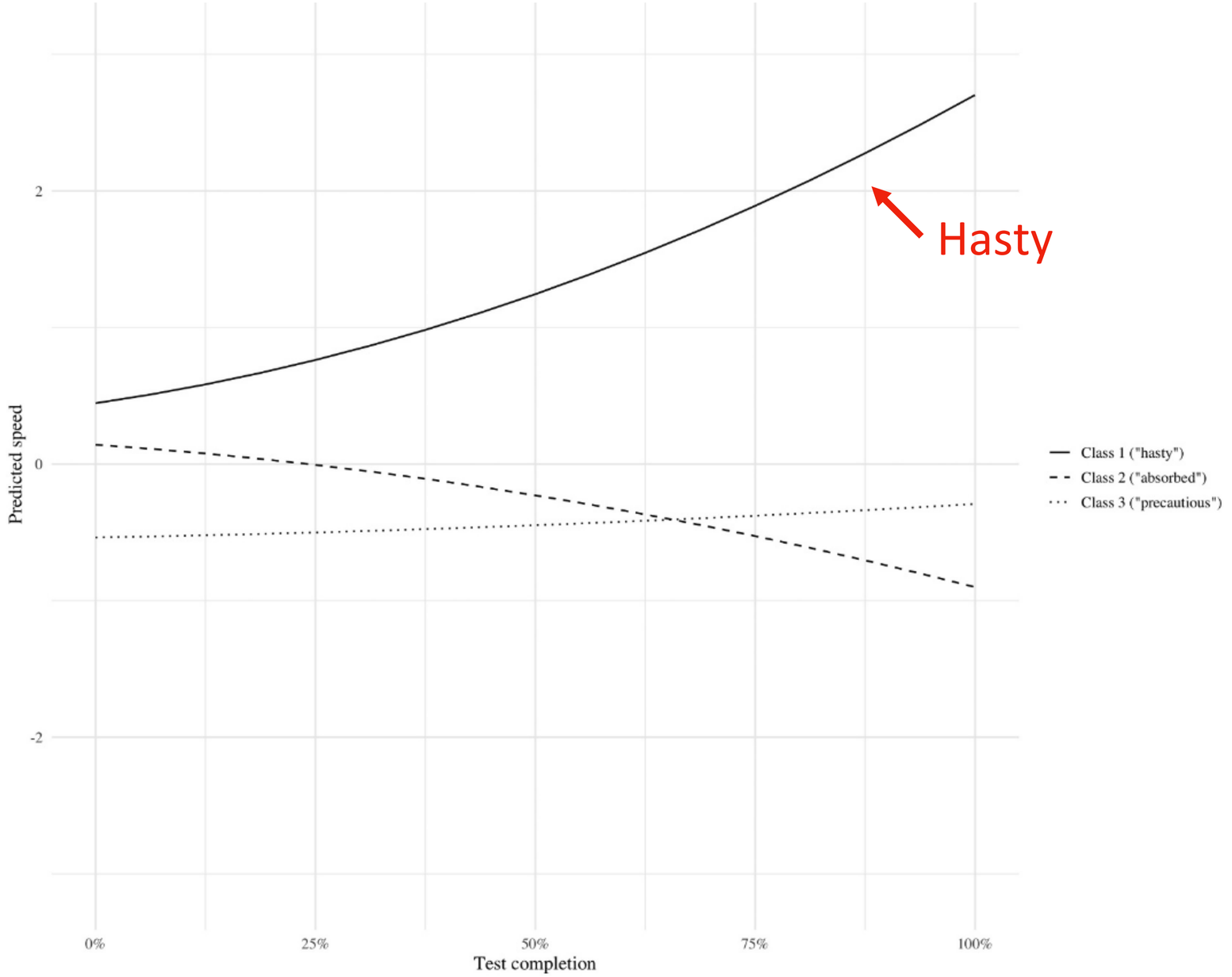


Fig. 4. Predicted speed trajectories of the classes identified with Latent Profile Analysis.

(Myszkowski, Storme, Kubiak & Baron, 2022)

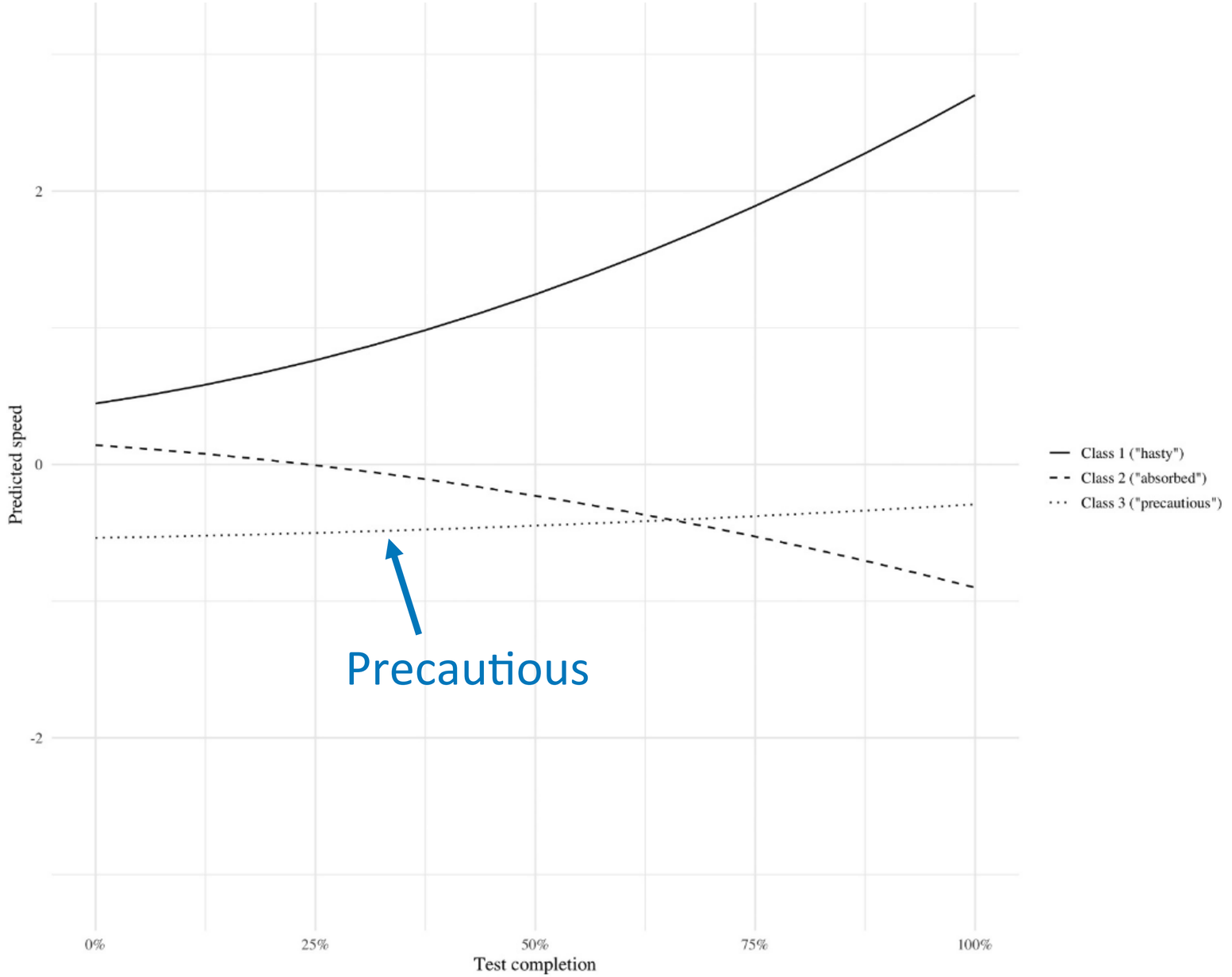


Fig. 4. Predicted speed trajectories of the classes identified with Latent Profile Analysis.

(Myszkowski, Storme, Kubiak & Baron, 2022)

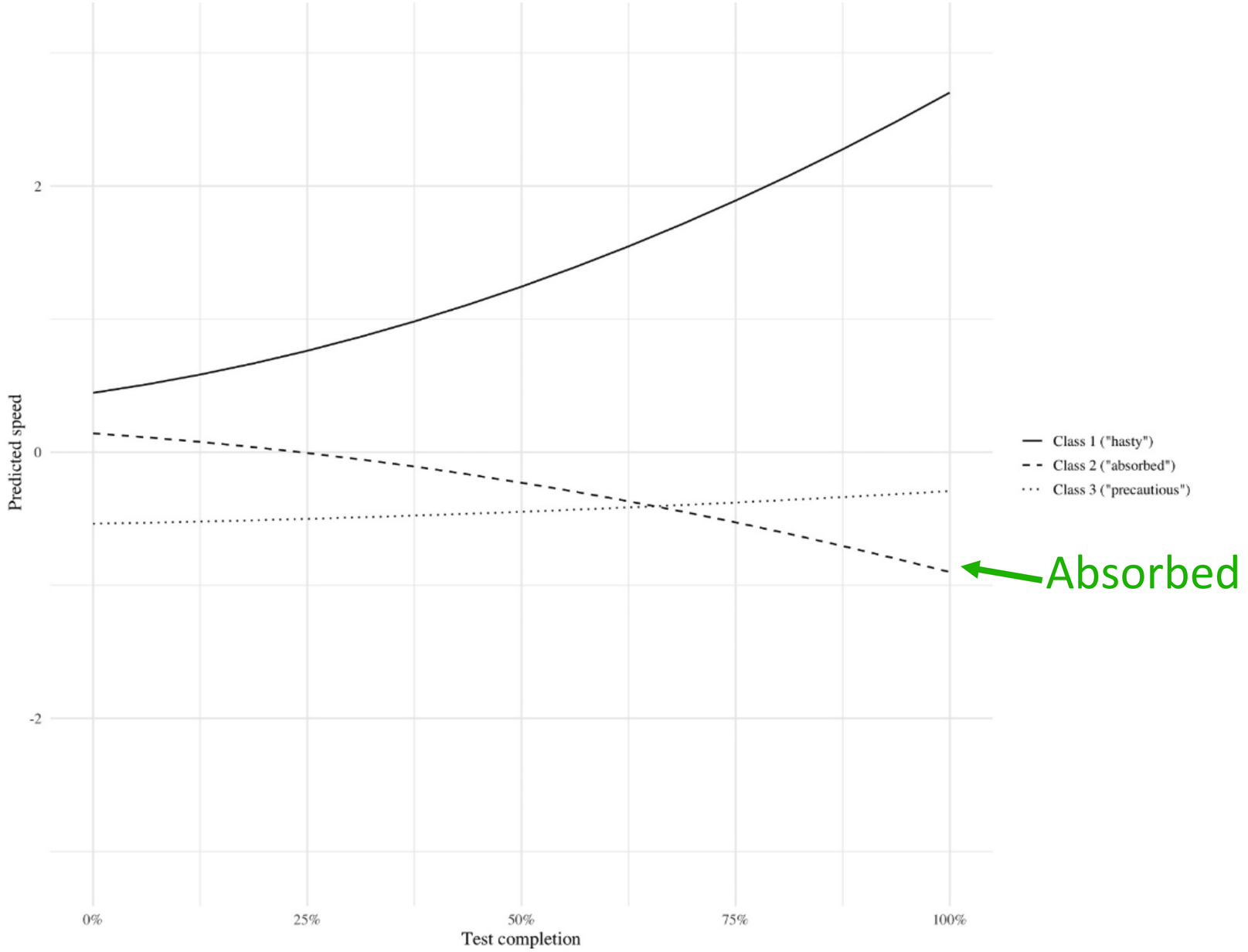


Fig. 4. Predicted speed trajectories of the classes identified with Latent Profile Analysis.

(Myszkowski, Storme, Kubiak & Baron, 2022)

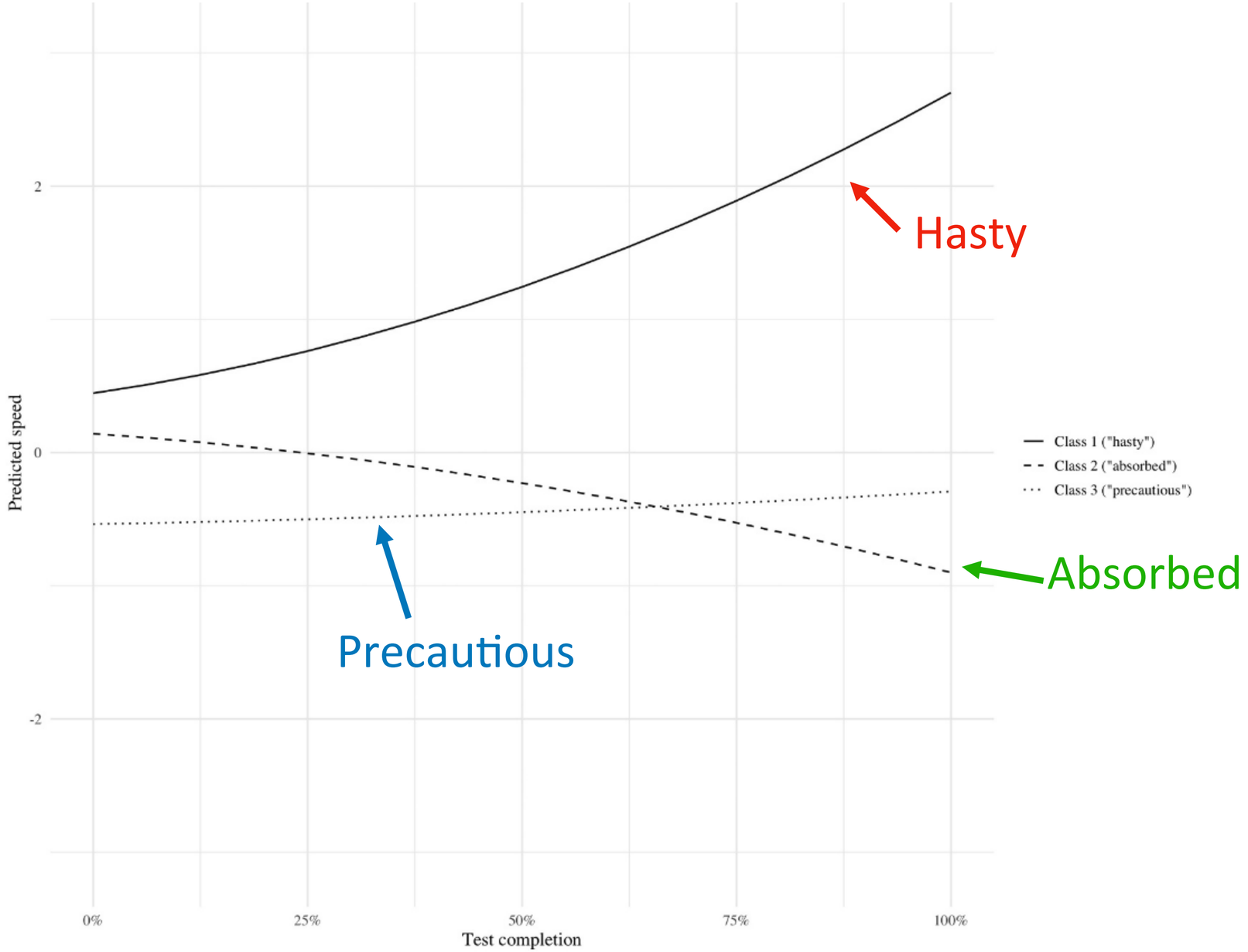


Fig. 4. Predicted speed trajectories of the classes identified with Latent Profile Analysis.

(Myszkowski, Storme, Kubiak & Baron, 2022)

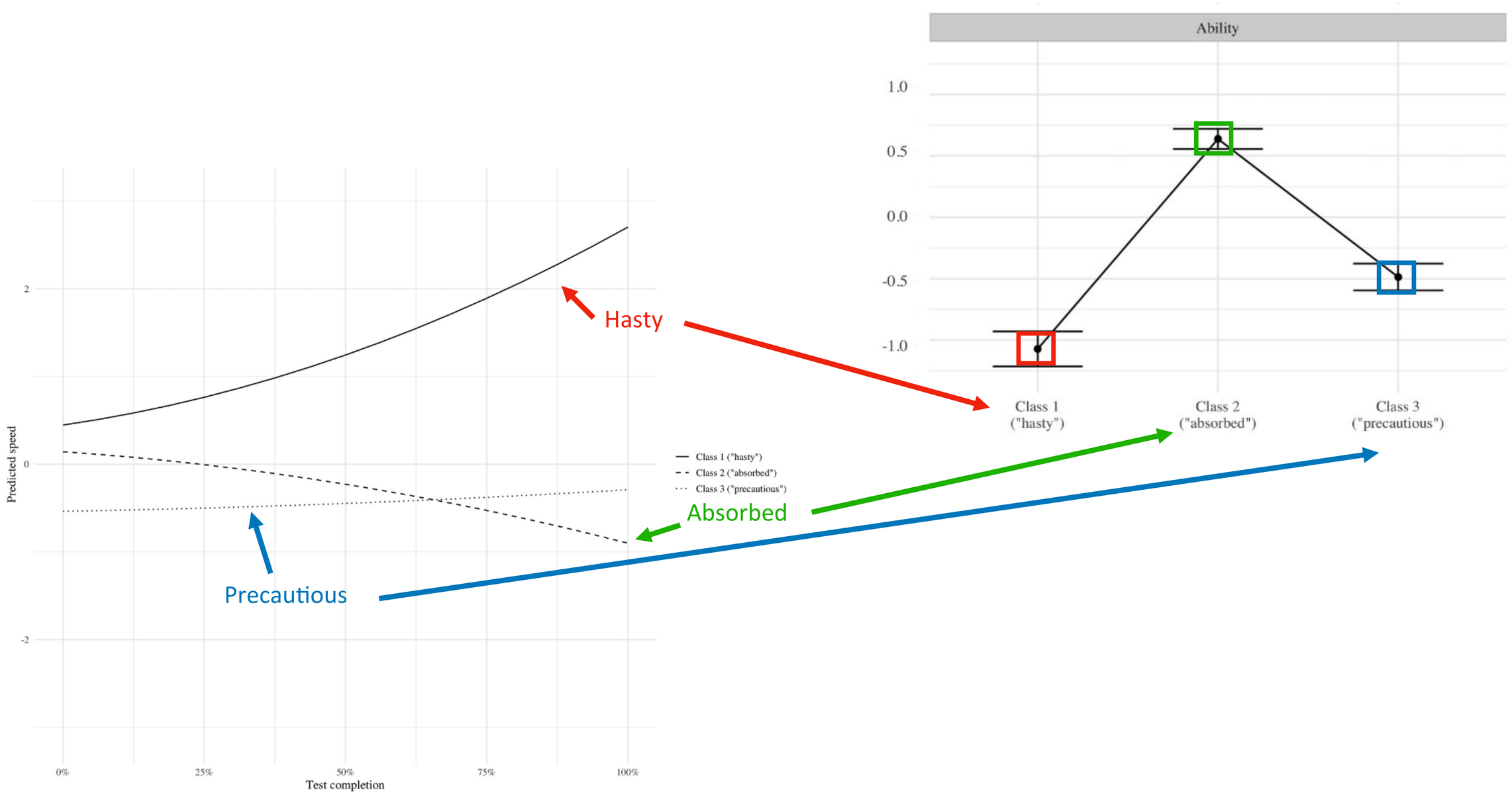
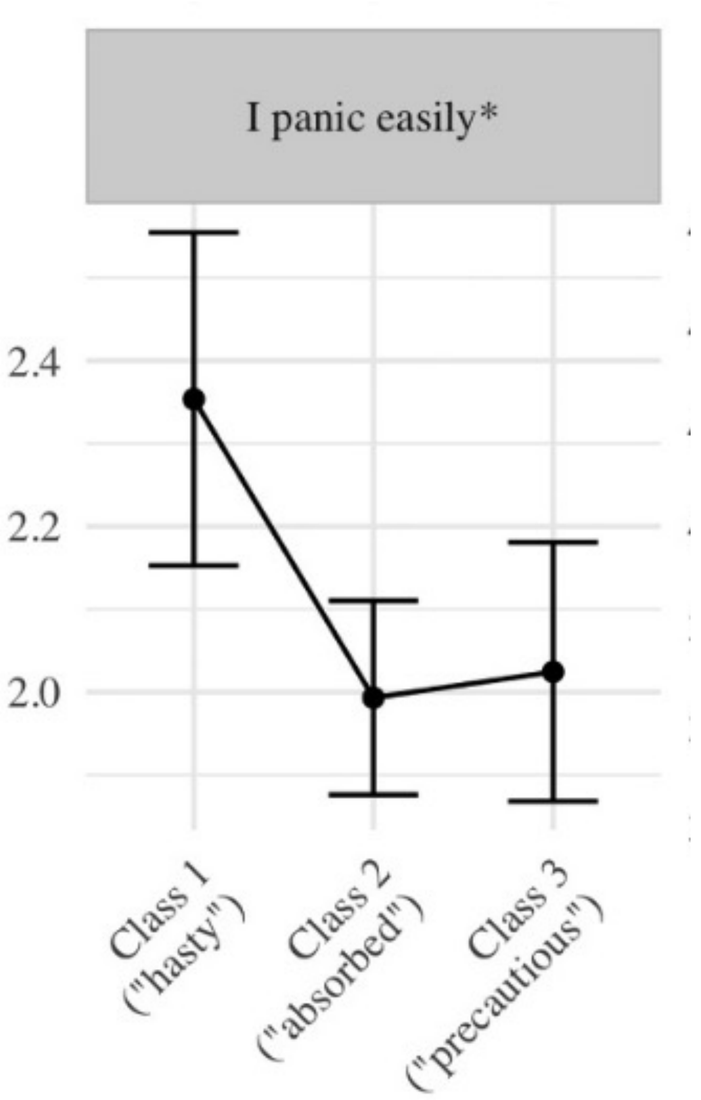
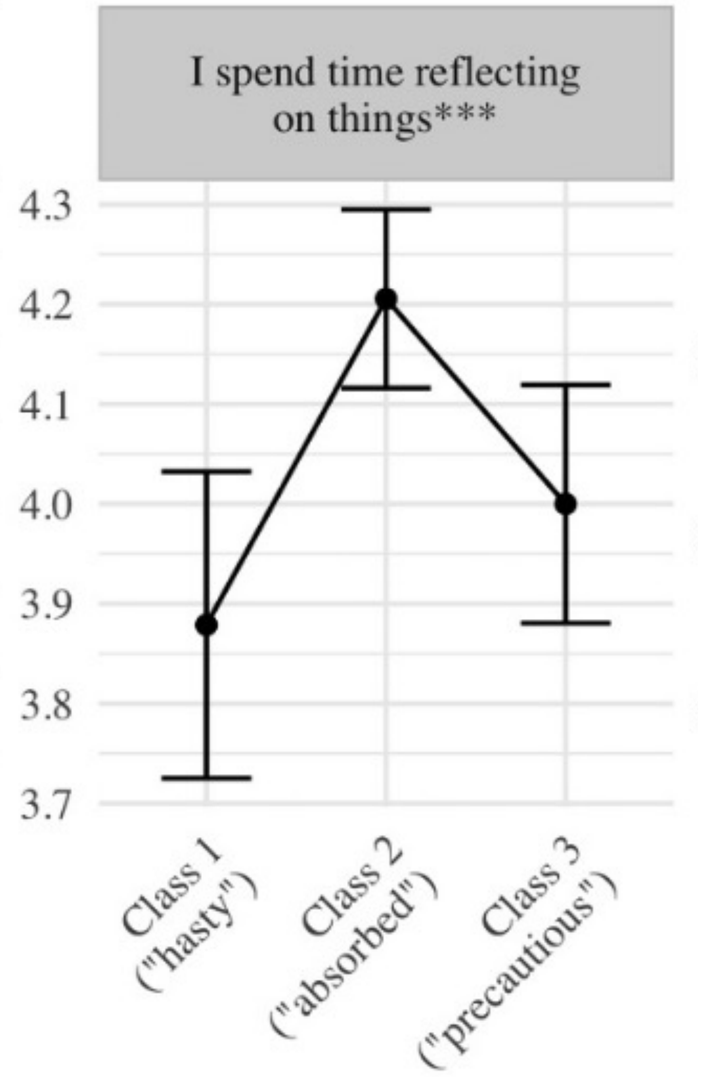


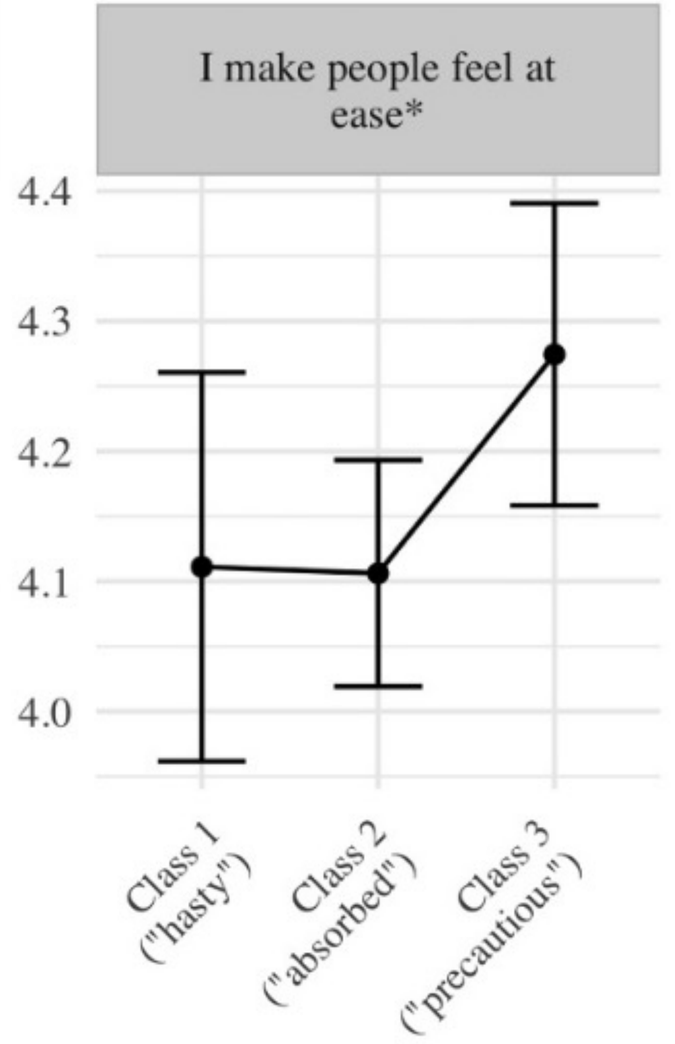
Fig. 4. Predicted speed trajectories of the classes identified with Latent Profile Analysis.



(Myszkowski, Storme, Kubiak & Baron, 2022)



(Myszkowski, Storme, Kubiak & Baron, 2022)



(Myszkowski, Storme, Kubiak & Baron, 2022)

Example 2 : Personality and test taking
(Myszkowski, Storme, Kubiak & Baron, 2022)

- Big five trait differences across trajectory types (One-Way MANOVA; p<.001)
  - "**Hasty**" (18% cases): **low** openness, **low** emotional stability
  - "**Absorbed**" (53% cases): **low** agreeableness, **high** openness
  - "**Precautious**" (29% cases): **high** agreeableness, **high** conscientiousness

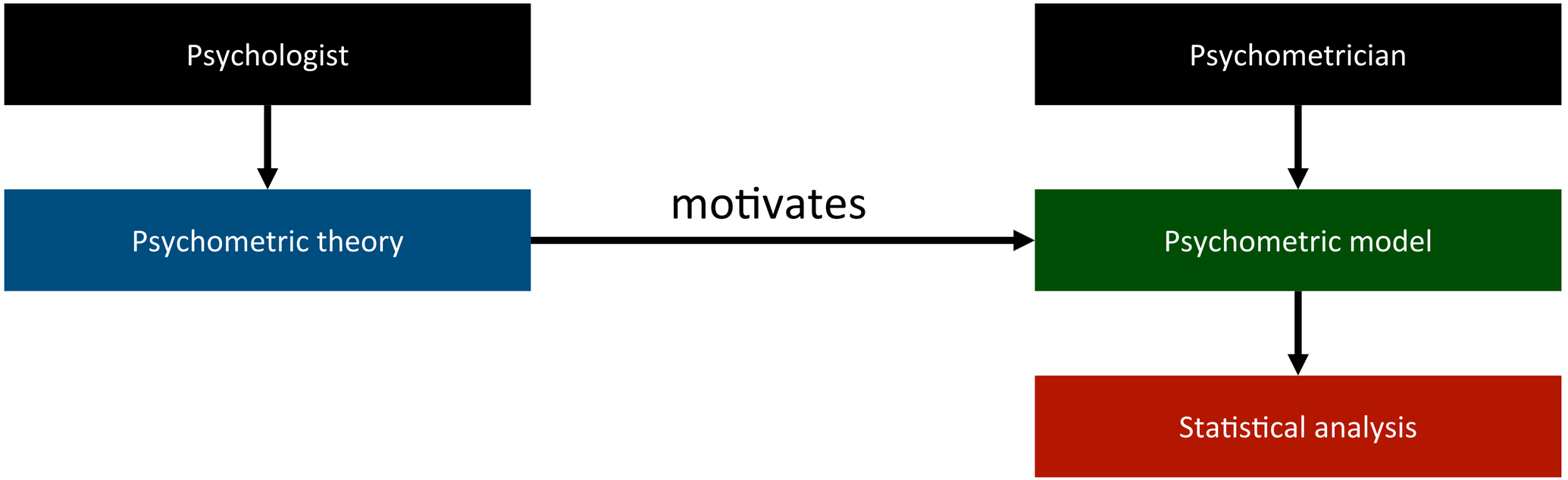
Conclusions

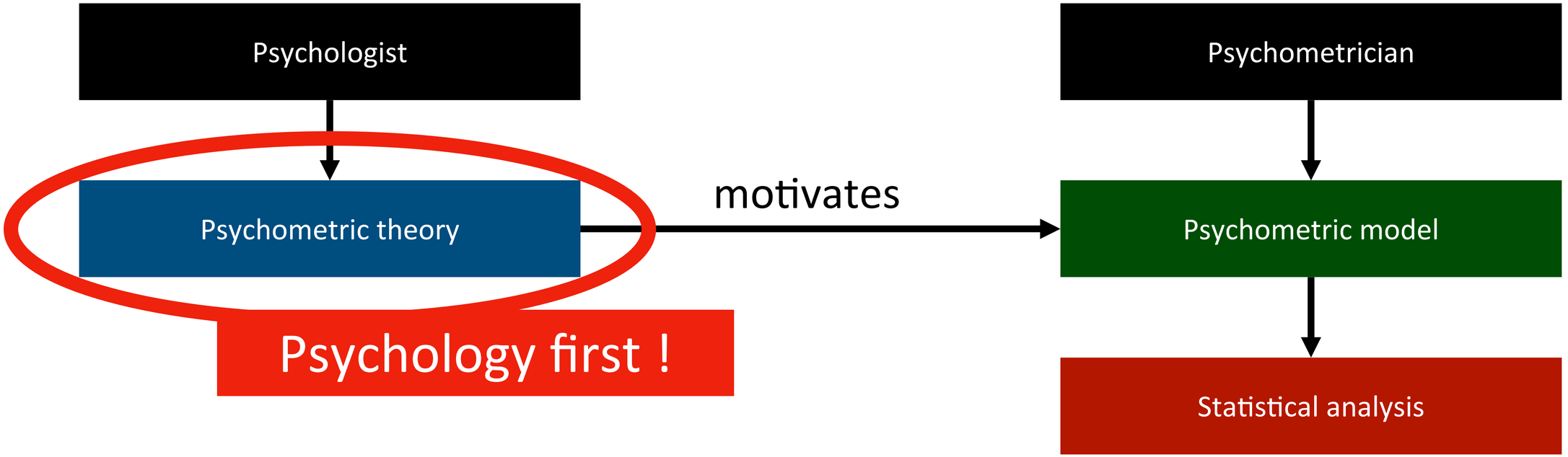
- Collateral information allows…
  - To increase reliability (or to use fewer items)
  - To study and measure various psychological processes (e.g., engagement) going on during the test
  - To control for them
  - To detect unexpected test behavior (e.g., guessing, cheating)

Conclusions
(and unsolicited advice)

Some important considerations for assessment

- All psychological testing situations are unique.
  - Various response formats (discrete, bounded, etc.)
  - Multiple decisions / strategies involved in responses
  - Different traits involved in these decisions
  - Other observable behaviors than the response
- The modern test theory framework allows to accommodate a lot of quirks and messy situations !


Psychologist
Psychometrician
Psychometric theory
motivates
Psychometric model
Statistical analysis


Psychologist
Psychometrician
Psychometric theory
motivates
Psychometric model
Psychology first !
Statistical analysis

Upcoming projects

Manuscripts in revision

- Meta-analysis on critical thinking - creative thinking in education (student 1st author, *Thinking Skills and Creativity*)
- Large-scale algebra test development (co-author, *Journal for Research in Mathematics Education)*

Manuscripts in preparation for 2025

- PISA Creative Thinking
  - Empirical application of IRTree (1st author, *Thinking Skills and Creativity*)
  - General IRT critique (1st author, *Creativity Research Journal*)
  - General psychometric critique (co-author, *Creativity Research Journal*)
- IRT analysis on the impact of early stopping rules in cognitive assessment using large-scale data (student 1st author, journal TBD)
- Dynamic measurement of creativity through the originality-appropriateness trade-off (1st author, journal TBD)

Grant projects

- Co-PI on NSF EDU-Core grant to be concluded in 2025
  - Team currently working on publishing results and possibly reapply for a new grant
- Fall 2025 : Grant for experiment on creative mindsets
  - APA – Division 10 Micro-Grant
- 2026 : Grant for the development of an R library to facilitate PISA data scoring
  - IES – Statistical and Research Methodology in Education | 84.305D
- 2026/2027 : Development of a tool for large-scale measurement of creativity in students in the US
  - NSF – EDU Core Research | 21–588

References

- Beisemann, M. (2022). A flexible approach to modelling over-, under- and equidispersed count data in IRT: The Two-Parameter Conway-Maxwell-Poisson Model. *The British Journal of Mathematical and Statistical Psychology*, *75*(3), 411–443. https://doi.org/10.1111/bmsp.12273
- Birnbaum, A., Lord, F. M., & Novick, M. R. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores* (pp. 397–472). Addison Wesley.
- De Ayala, R. J. (2022). *The theory and practice of item response theory* (Second edition). The Guilford Press.
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-Based Item Response Models of the GLMM Family. *Journal of Statistical Software*, *48*(1), 1–28. https://doi.org/10.18637/jss.v048.c01
- Forthmann, B., Bürkner, P.-C., Szardenings, C., Benedek, M., & Holling, H. (2019). A New Perspective on the Multidimensionality of Divergent Thinking Tasks. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00985
- Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of Responses and Response Times with the Package CIRT. *Journal of statistical software*, *20*(7), 1–14.
- Götz, K. O. (1985). *VAST: Visual Aesthetic Sensitivity Test* (4th ed.). Concept Verlag.
- Lang, J. W. B., Lievens, F., De Fruyt, F., Zettler, I., & Tackett, J. L. (2019). Assessing meaningful within-person variability in Likert-scale rated personality descriptions: An IRT tree approach. *Psychological Assessment*, *31*(4), 474–487. https://doi.org/10.1037/pas0000600
- Lievens, F., Lang, J. W. B., De Fruyt, F., Corstjens, J., Van de Vijver, M., & Bledow, R. (2018). The predictive power of people's intraindividual variability across situations: Implementing whole trait theory in assessment. *Journal of Applied Psychology*, *103*(7), 753–771. https://doi.org/10.1037/apl0000280
- Myszkowski, N. (2019). The first glance is the weakest: "Tasteful" individuals are slower to judge visual art. *Personality and Individual Differences*, *141*, 188–195. https://doi.org/10.1016/j.paid.2019.01.010
- Myszkowski, N. (2021). Development of the R library "jrt": Automated item response theory procedures for judgment data and their application with the consensual assessment technique. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(3), 426–438. https://doi.org/10.1037/aca0000287
- Myszkowski, N., & Storme, M. (2017). Measuring "good taste" with the Visual Aesthetic Sensitivity Test-Revised (VAST-R). *Personality and Individual Differences*, *117*, 91–100. https://doi.org/10.1016/j.paid.2017.05.041
- Myszkowski, N., & Storme, M. (2018). A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence*, *68*, 109–116. https://doi.org/10.1016/j.intell.2018.03.010
- Myszkowski, N., & Storme, M. (2019). Judge response theory? A call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 167–175. https://doi.org/10.1037/aca0000225
- Myszkowski, N., & Storme, M. (2021). Accounting for Variable Task Discrimination in Divergent Thinking Fluency Measurement: An Example of the Benefits of a 2-Parameter Poisson Counts Model and its Bifactor Extension Over the Rasch Poisson Counts Model. *The Journal of Creative Behavior*, *55*(3), 800–818. https://doi.org/10.1002/jocb.490
- Myszkowski, N., Storme, M., Kubiak, E., & Baron, S. (2022). Exploring the associations between personality and response speed trajectories in low-stakes intelligence tests. *Personality and Individual Differences*, *191*(111580), 1–9. https://doi.org/10.1016/j.paid.2022.111580
- Raven, J. C. (1941). Standardization of Progressive Matrices, 1938. *British Journal of Medical Psychology*, *19*(1), 137–150. https://doi.org/10.1111/j.2044-8341.1941.tb00316.x
- Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 68–85. https://doi.org/10.1037/1931-3896.2.2.68
- Storme, M., Celik, P., & Myszkowski, N. (2020). A forgotten antecedent of career adaptability: A study on the predictive role of within-person variability in personality. *Personality and Individual Differences*, *160*, 1–6. https://doi.org/10.1016/j.paid.2020.109936
- Storme, M., Myszkowski, N., Baron, S., & Bernard, D. (2019). Same Test, Better Scores: Boosting the Reliability of Short Online Intelligence Recruitment Tests with Nested Logit Item Response Theory Models. *Journal of Intelligence*, *7*(3), 1–17. https://doi.org/10.3390/jintelligence7030017
- Suh, Y., & Bolt, D. M. (2010). Nested Logit Models for Multiple-Choice Item Response Data. *Psychometrika*, *75*(3), 454–473. https://doi.org/10.1007/s11336-010-9163-7
- van der Linden, W. J., & Fox, J.-P. (2016). Joint Hierarchial Modeling of Responses and Response Times. In *Handbook of Item Response Theory, Volume One: Models* (1st ed., pp. 481–500). Chapman and Hall/CRC. https://doi.org/10.1201/9781315374512

Thank you !