$\theta \geq 0$

Han van der Maas / Dylan Molenaar
University of Amsterdam
What is Psychometrics, 2015

Psychometric interests

- Cognitive psychometrics
- Response times in psychometrics
- Networks as alternatives for latent variable
  - Mutualism model of intelligence
- Rekentuin/Math Garden project
  - Scoring rules
  - learning
  - Self-organizing computer adaptive tests
- Latent classes, developmental transitions
- Psychometrics of Chess

plan

- Theory
- Break
- R session
- Break
- More theory and discussion
- Lunch

Cognitive psychometrics

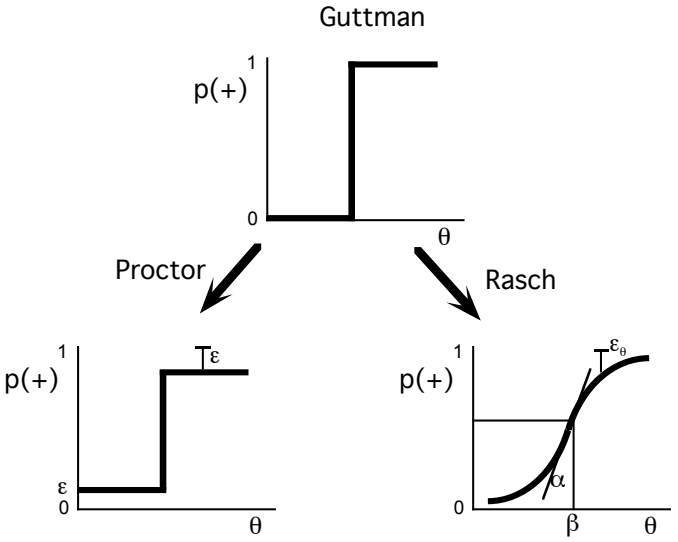
- Riefer et. Al (2002): cognitive psychometrics
- Rouder et al. (2014): lognormal race model
- Vandekerckhove (2011): Hierarchical diffusion models
- De la Torre (2009): Cognitive diagnosis models
- Fischer (1972): Linear Logistic Test Model
- Van der Linden (2009): Hierarchical model
- …

My long term issues with psychometrics

- No quantitative measurement in IRT
  - Rasch paradox
- IRT function (logistic) is not based on theory
  - No process model
- No good model for guessing
  - 3 PL sucks ;-)
- No substantial definition of one-dimensionality
  - Statistical definition (one factor model or 2PL fits) depends on sample characteristics and size
- …

Rasch paradox

- What is the scale level of IRT ability estimates?
- Are ability estimates any better than sum scores?
  - What is the scale level of a Guttman scale?
    - Ordinal (we can rank order subjects, see Birnbaum, 1968)
  - What is the scale level of a Rasch scale?
    - Interval (we can scale subjects, Sijtsma & Molenaar, 2002)
    - Perhaps ordered metric (Bezembinder)
    - At least better than ordinal
  - This is peculiar!



Why is this peculiar?

- All we did was adding error to the deterministic Guttman model.
- How can error bring us to a higher measurement scale?
- This comes across as a Baron von Münchhausen adventure.
- We pull ourselves out of the measurement problems of social science by adding random error.

In other words

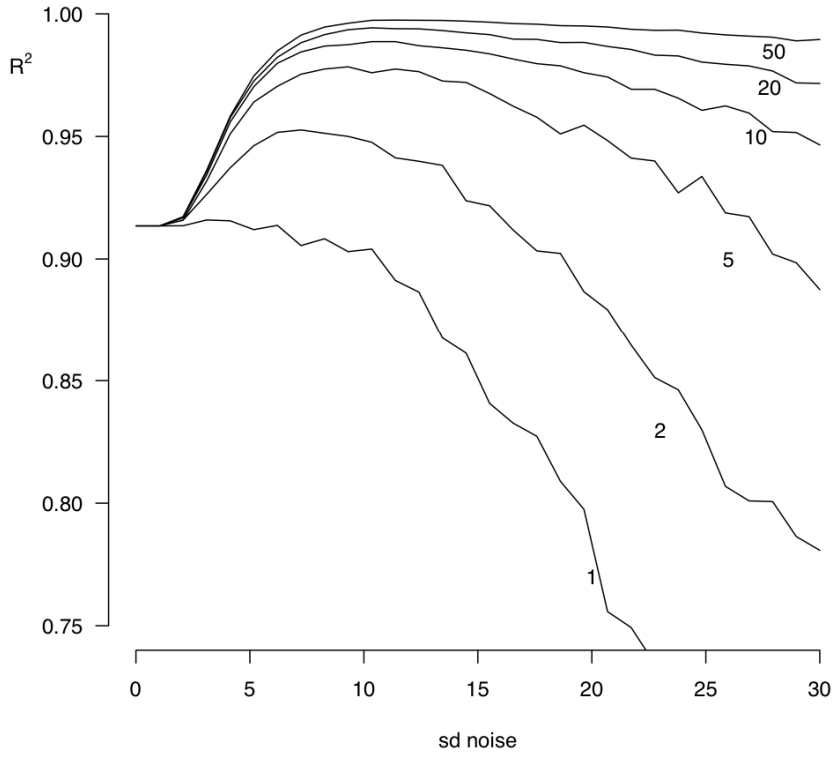
- A good test consists of items with high discrimination, making up a scale with interval properties. If we now improve the items to make them even better discriminating, a scale paradox emerges.
- At a certain point the slopes are so steep that Rasch items turn into Guttman items, and we are left with an ordinal scale. So by improving our items we loose measurement properties!
- So
  - Either Guttman scales are also interval
  - Or Rasch scales are ordinal
- Or is this scale paradox a real effect?

Scale paradox

- I have 4 persons (A, B, C, D) with lengths of 1.60, 1.65, 1.90, 1.95
- I compare each pair n times. If A>B, A gets one point
- I do this without error 10 times, I always find A>B>C>D
  - Only ordinal information (scores 0,10,20,30)
- Now, I ask my 5 year old daughter to do the same. She gets scores 3, 6, 23, 29
  - Her correlation with the real lengths is much better: She finds interval data!

Simulation

- Compare A+noise with B
- Vary sd of noise
- Vary number of comparisons n (1, 2, 5, 10, 20, 50)



Conclusion: error helps! There even seems to be an 'optimal level of sloppiness' (depending on n).

It turns out that:

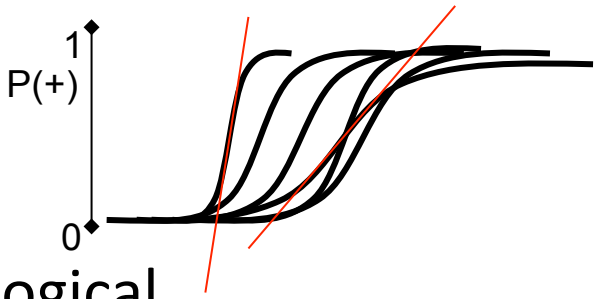
- Scale paradox occurs for bimodally distributed β
- Scale paradox occurs when $\sigma_\theta >> \sigma_\beta$
  - This is a rather trivial case: we should select items with different difficulties
    - But 'classic' test often consists of items of equal difficulty
- But no scale paradox when difficulties are well spaced over the scale (which is the normal case)

My long term issues with psychometrics

- No quantitative measurement in IRT
  - Rasch paradox
- IRT function (logistic) is not based on theory
  - No process model
- No good model for guessing
  - 3 PL sucks ;-)
- No substantial definition of one-dimensionality
  - Statistical definition (one factor model or 2PL fits) depends on sample characteristics and size
- …

2PL model in item response theory

$$P_{jk}(+ \mid \theta_k) = \frac{e^{\alpha_j(\theta_k - \beta_j)}}{1 + e^{\alpha_j(\theta_k - \beta_j)}}$$



- Main model for educational and psychological measurement in Psychometrics
- Justifications
  - Logistic equation does the job (Toolbox)
  - Derivation from 'desirable' statistical properties (Fischer)
  - Derivation from 'desirable' measurement properties (Roskam)
  - *Derivation from psychological process model (Tuerlinckx & De Boeck)*

New model: Q-diffusion model

- A bridge between (mathematical) psychology and psychometrics
- Justification for logistic function in IRT
- Allows for quantitative measurement
- New solution for guessing
- New definition of one-dimensionality

Diffusion model of decision making

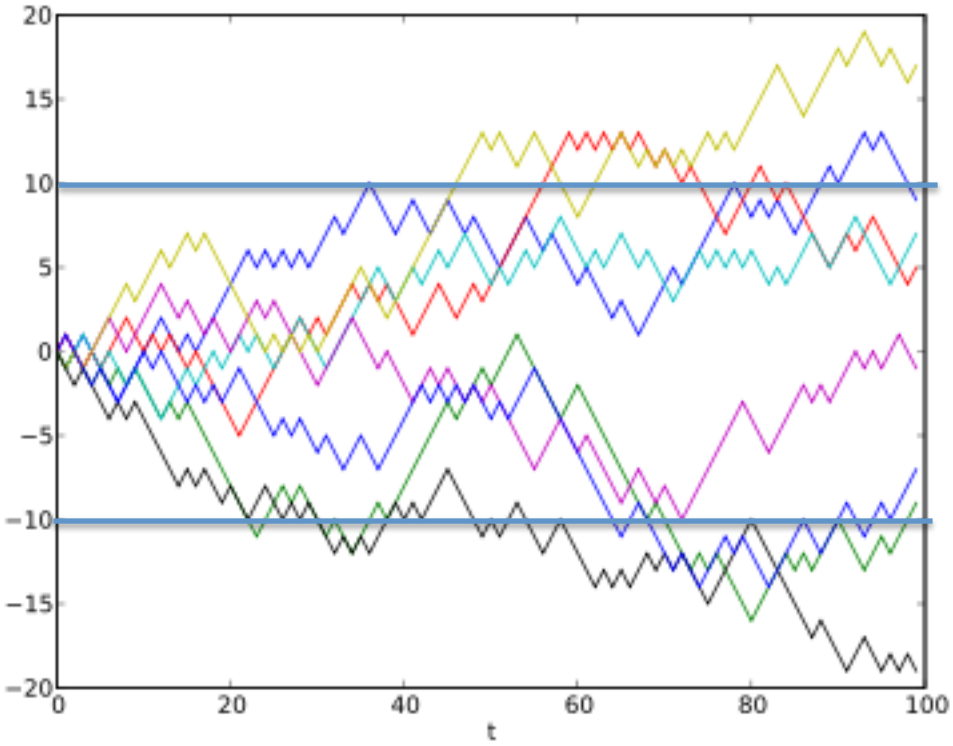
- Model of simple *two choice* decisions

- Stochastic accumulation of evidence for one of the alternatives
- Stops when decision threshold is reached
- Explains the speed accuracy trade-off
- Implements SPRT (optimal accuracy given RT)

- Very influential and very well studied model
- More complex biological realistic models reduce to this model

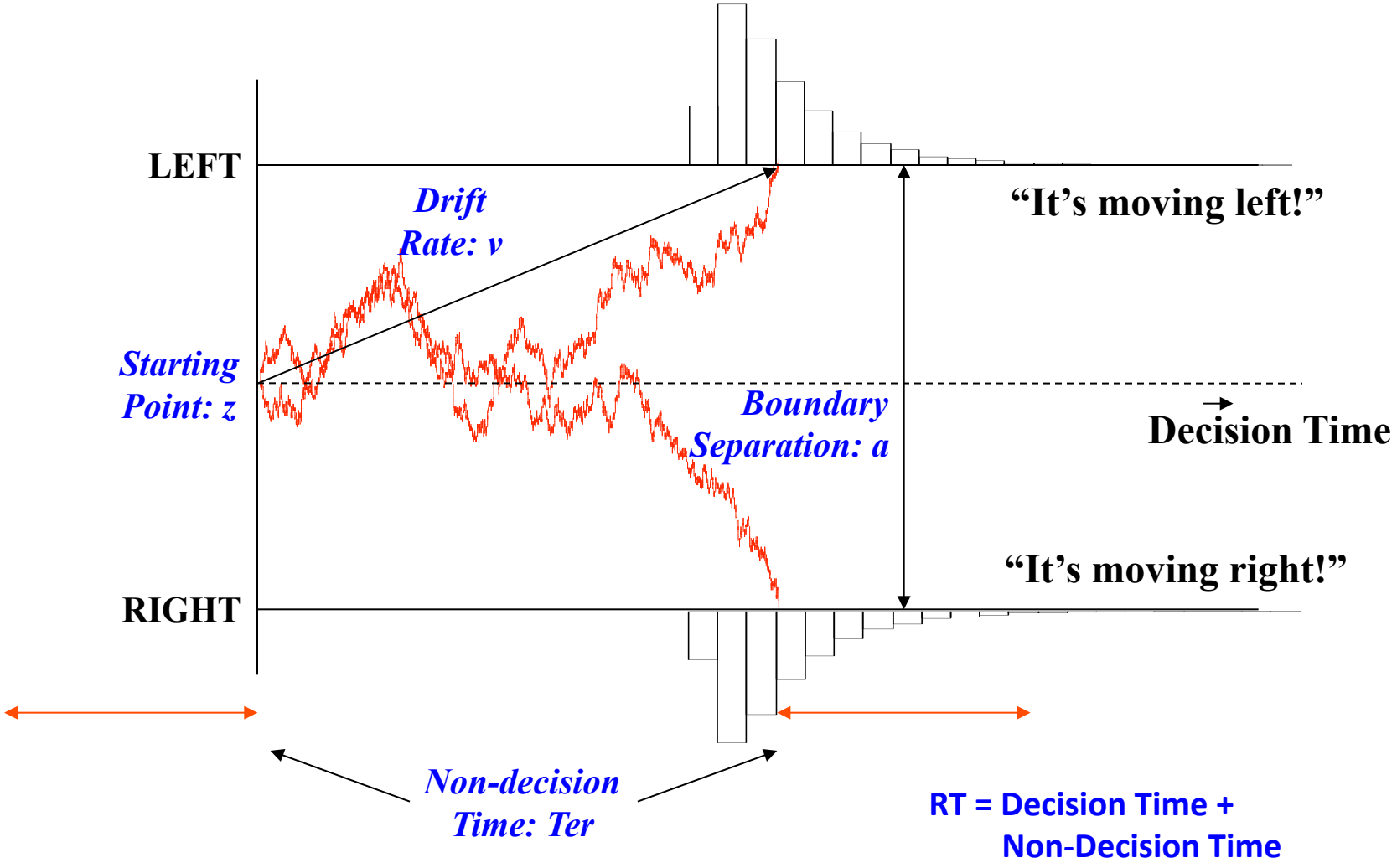
Random walk model

- 2 boundaries
- P(step up) = .6
- P(step down) = .4
- Accuracy = P(hit upper boundary)
- RT = mean t to hit a bound
- Speed accurracy trade-off: lower bounds -> faster but less accurate
- RT(+) = RT(-) ?

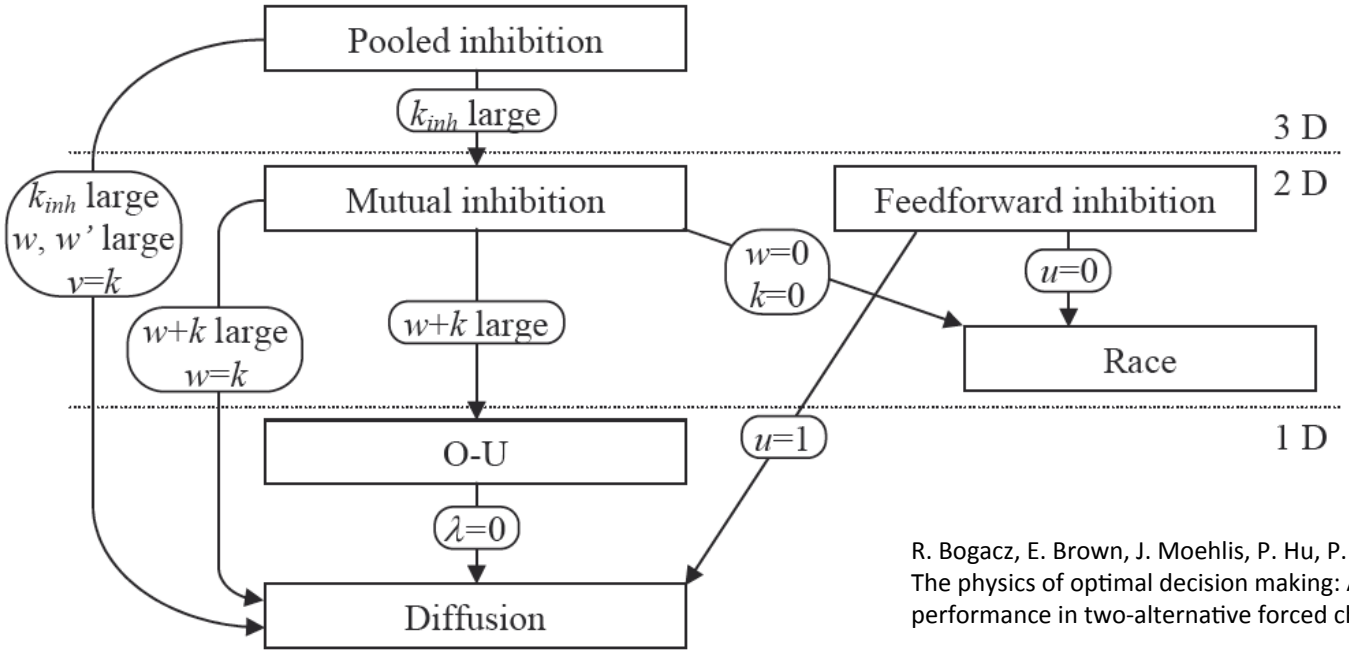


The Diffusion Model

(continuous time and steps)



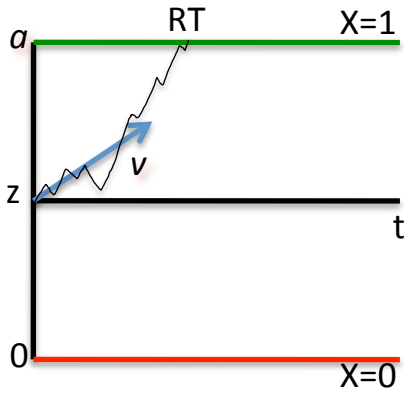
Many, many models



R. Bogacz, E. Brown, J. Moehlis, P. Hu, P. Holmes and J.D. Cohen (2005) The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks.

Relation to IRT

- Two main parameters:
  - Boundary separation (decision criterion): a
  - Drift rate (rate of evidence accumulation): v



$$P(X=1) = \frac{e^{-av}-1}{e^{-2av}-1} = \frac{e^{av}}{1+e^{av}}$$ if z=a/2

$$\mathrm{E}(DT) = \frac{a}{2v}\frac{1-e^{-av}}{1+e^{-av}} \approx \frac{a}{2v}$$ if P(X=1) is large

Tuerlinckx & De Boeck (2005):
Drift = ability – difficulty ($v = \theta - \beta$)
Discrimination = boundary separation ($a = \alpha$)

$$P(X=1) = \frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}} \qquad \mathrm{E}(DT) \approx \frac{\alpha}{2(\theta-\beta)}$$

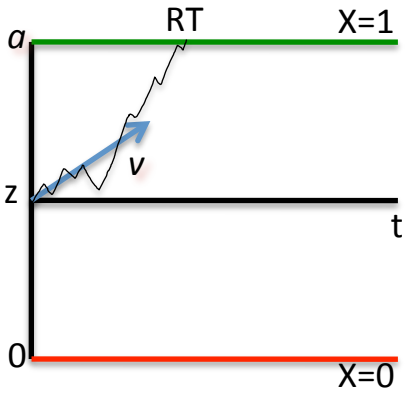
This seems great

- 2PL derived from a popular and important process model
- A model of RT as bonus
- New interpretation of the discrimination parameter

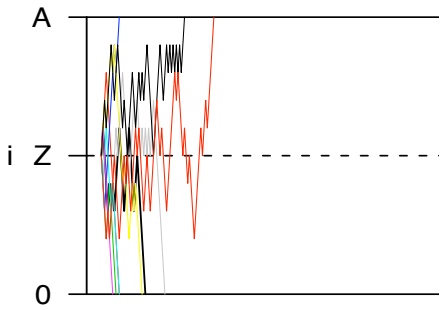
- But…

Three qualitative predictions

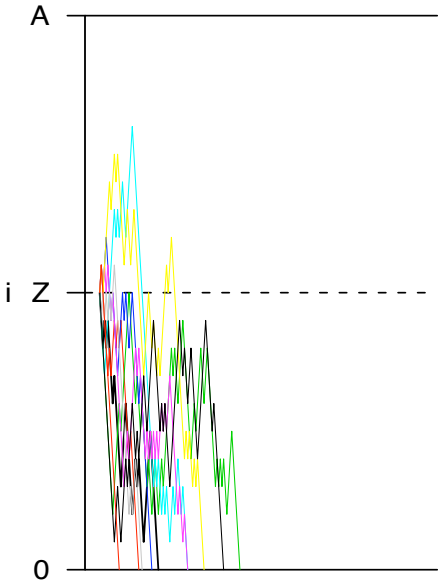
1. Subjects are slowest when v=0, that is when $\theta=\beta$
   - When $\theta<<\beta$, responding is very fast
2. For negative v, i.e. $\theta<\beta$, allowing more time to think reduces the probability correct.
3. By reducing time to think P(+) -> .5, irrespective of θ



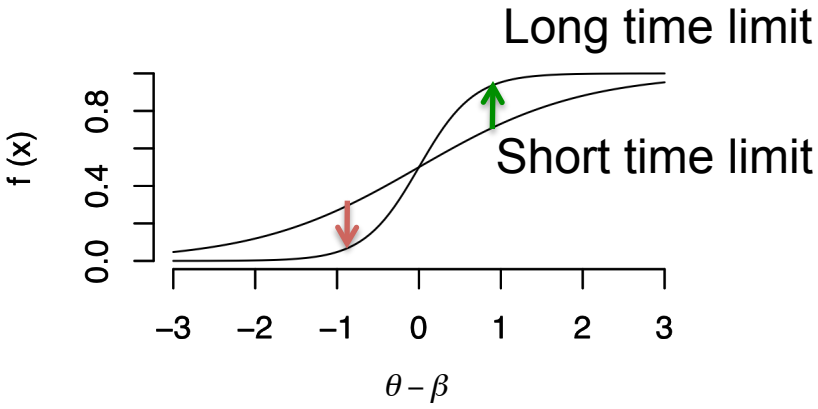
For negative v (θ<β), longer time limits reduce the probability correct.



Short time limit



Long time limit



Very short time limits reduce P+ to .5

Attitudes: "I stick to my decisions" (agree/not agree)

1. RT decreases with $|\theta-\beta|$
   - OK: extreme answers are fast
2. if $\theta<\beta$ and time limit increases P(agree)←0
   - OK: the longer I think about: I don't stick to my decisions
3. If time to think (a) →0, P(agree) →1/2
   - OK: I just say something

Ability: “24+79” (93/114/103/130)

1. RT decreases with $|\theta-\beta|$
   - NO, best subjects are fastest
2. if $\theta<\beta$ and time limit increases P(agree)←0
   - NO, they guess because they know that they don’t know
3. If time to think (a) →0, P(+) →1/2
   - NO: P(+) →0 for open questions and to 1/M for M multiple choice questions

Thus

For two choice attitude/personality tests the diffusion IRT model works well!

- But this case is relatively rare

But not for ability tests:

Problem of Guessing

- In ability testing subjects often guess and don't score below chance level. They will not score worse when given more time to think.

Guessing

- Subjects usually don't score below chance level
  - Typical solution: 3PL

- Our solution is based on a conceptual analysis of abilities.
  - What are abilities?
- In standard IRT ability (θ) can be anything: a trait, attitude, capacity, competence, ability etc.
  - Ability (θ) is an individual difference variable
- But if we analyze ability at the level of the individual, ability is very special case

Note: Ability ≥ 0

- Think of walking (or chess ability, language comprehension, etc.)
  - No negative ability (no negative difficulty)
  - Real zero point (no ability)
  - With sufficient time and ability>0 any item can be solved
  - This in contrast to many attitudes and traits (which can have negative values)
- So we need a diffusion IRT model with positive v (or θ)

Idea: *a* and *v* both have an item and a person part

drift *v*

ability person v^p (>0)

difficulty item v^i (>0)

requirement

$v=f(v^p, v^i)>0$

one sensible solution

$v=v^p/v^i$

boundary separation *a*

response caution a^p (>0)

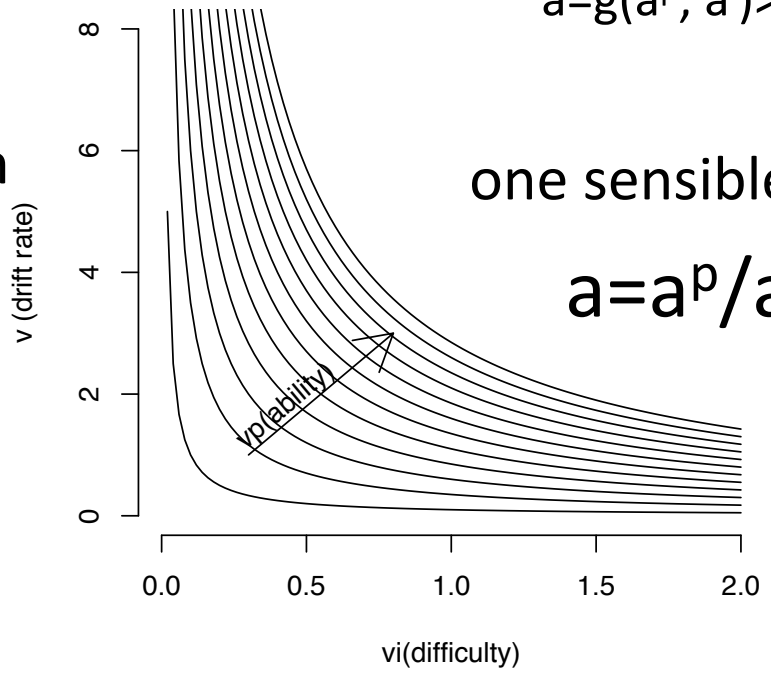
time pressure a^i (>0)

requirement

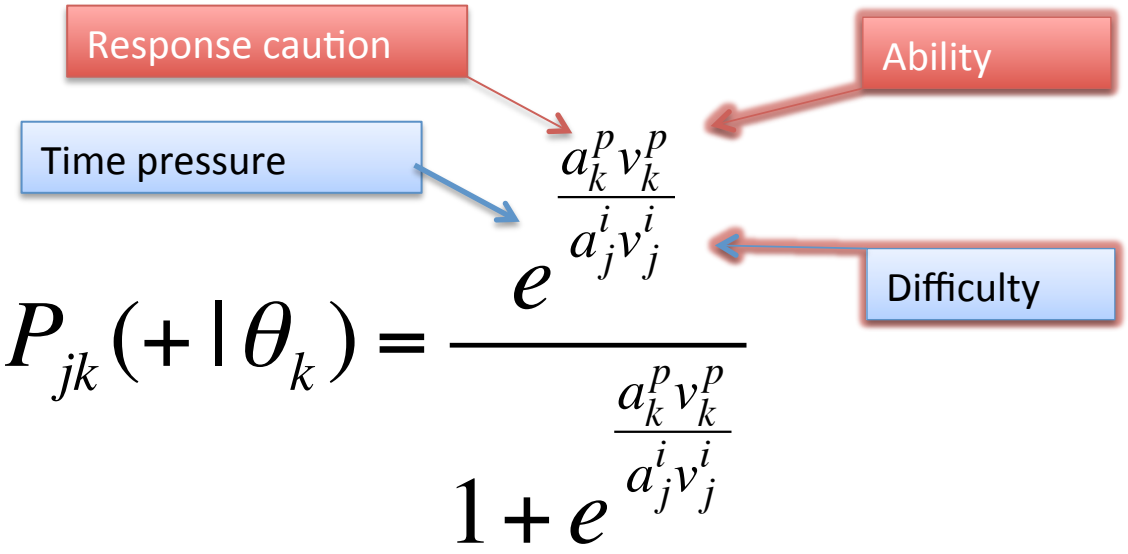
$a=g(a^p, a^i)>0$

one sensible solution

$a=a^p/a^i$



Quotient diffusion 2PLM (Q-diffusion model)



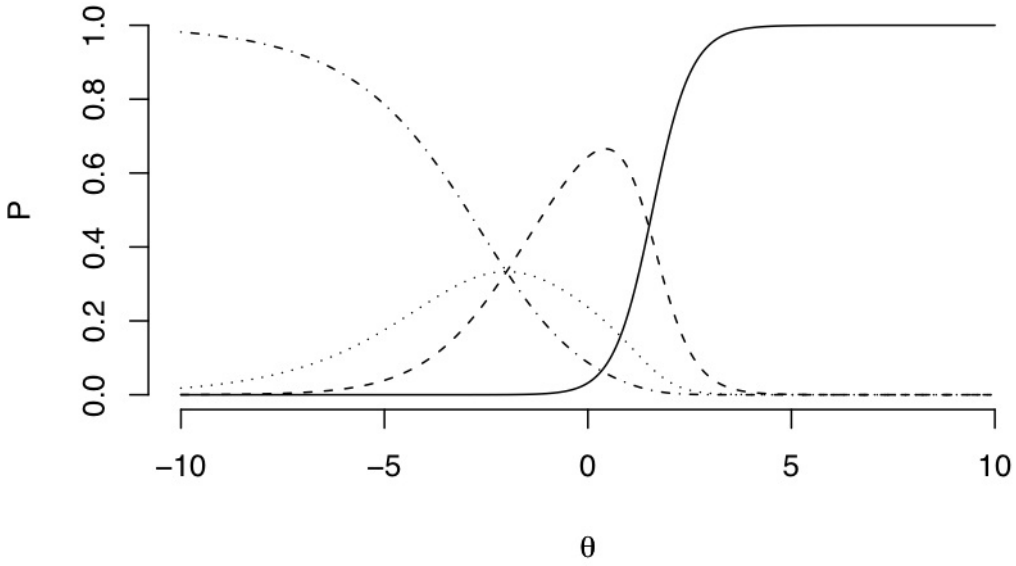
All *a* and *v* are positive

Extension to MC

Polytomous IRT

- Most general polytomous IRT model (non ordered responses) is Bock's nominal response model:

$$P_m(\theta) = \frac{e^{b_m^* + a_m^* \theta}}{\sum_k^M e^{b_k^* + a_k^* \theta}}$$



Identification

- Nominal response model can be identified by setting one a* and one b* to zero (or the sums of a* and b*)
- For two alternatives (M=2) we get:

$$P_2(\theta) = \frac{e^{b_2^* + a_2^*\theta}}{e^{0+0\theta} + e^{b_2^* + a_2^*\theta}} = \frac{e^{b_2^* + a_2^*\theta}}{1 + e^{b_2^* + a_2^*\theta}}$$

- which is the 2PL model!

Same trick for M>2

If $a^*_1 \ldots a^*_{M-1}$ and $b^*_1 \ldots b^*_{M-1}$ are set to zero:

$$P_m(\theta) = \frac{e^{b_m^* + a_m^* \theta}}{(M-1)e^{0+0\theta} + e^{b_2^* + a_2^* \theta}} = \frac{e^{b_m^* + a_m^* \theta}}{(M-1) + e^{b_m^* + a_m^* \theta}} =$$

$$\frac{\frac{e^{b_m^* + a_m^* \theta}}{e^{\ln(M-1)}}}{\frac{e^{\ln(M-1)}}{e^{\ln(M-1)}} + \frac{e^{b_m^* + a_m^* \theta}}{e^{\ln(M-1)}}} = \frac{e^{b_m^* + a_m^* \theta - \ln(M-1)}}{1 + e^{b_m^* + a_m^* \theta - \ln(M-1)}}$$

which, given m is the correct answer, a*=a and b*=-ba, gives:

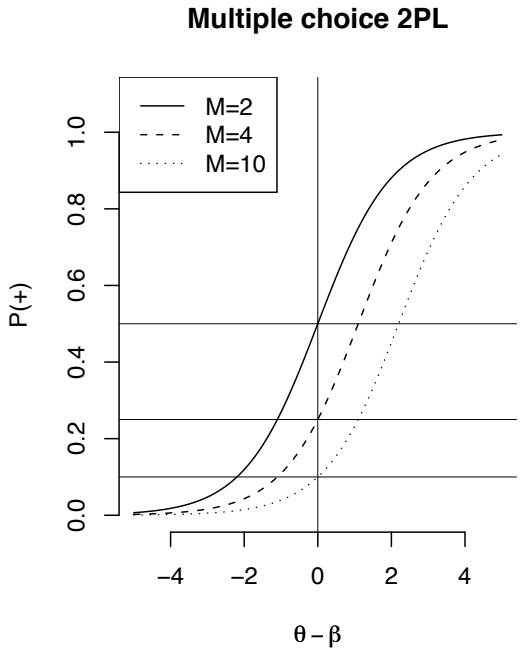
$$P(+ \mid \theta) = \frac{e^{\alpha(\theta - \beta) - \ln(M-1)}}{1 + e^{\alpha(\theta - \beta) - \ln(M-1)}}$$

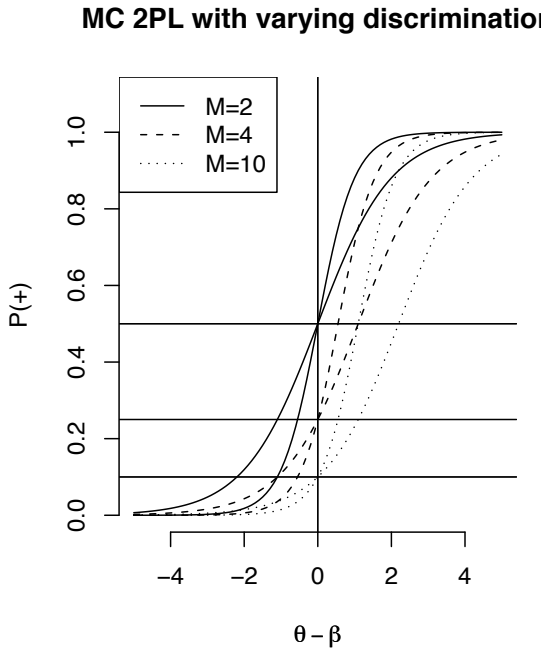
MC2PL

$$P(+\mid\theta) = \frac{e^{\alpha(\theta-\beta)-\ln(M-1)}}{1+e^{\alpha(\theta-\beta)-\ln(M-1)}}$$

Note: M=2 gives standard 2PL

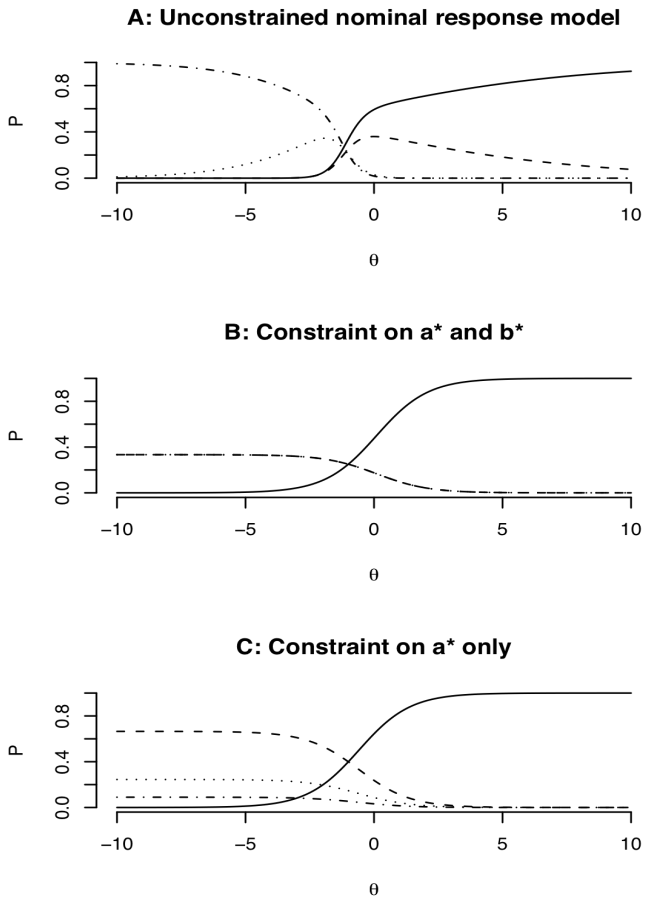
$$P(+\mid\theta=\beta) = \frac{e^{-\ln(M-1)}}{1+e^{-\ln(M-1)}} = \frac{1}{M}$$



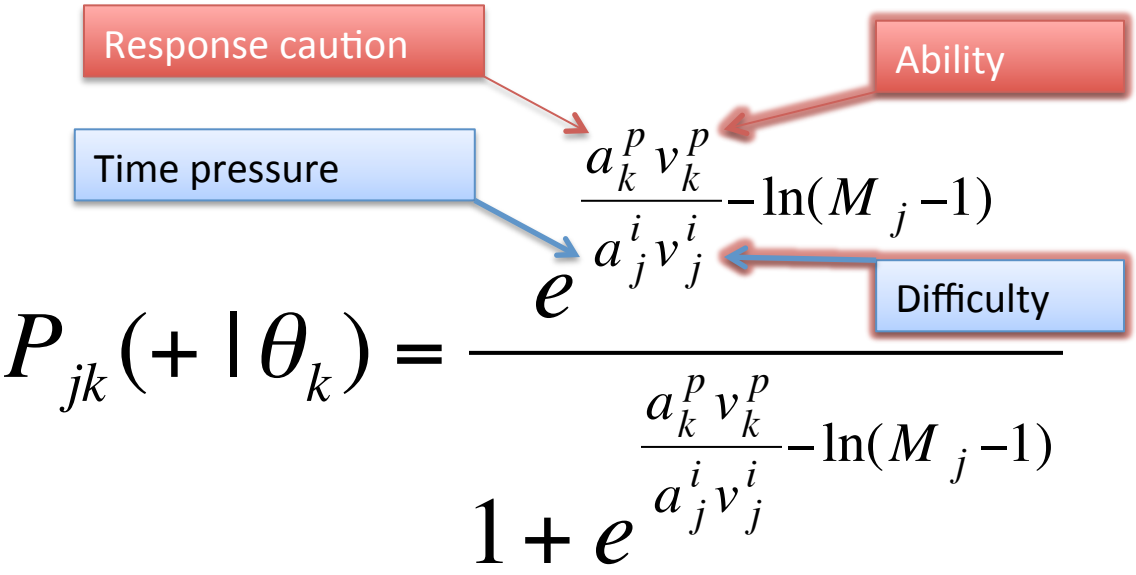


Thus:

- Multiple choice dichotomous model
  - Assuming equal attractiveness of incorrect alternatives
- Guessing at 1/M
  - Note if $\theta=\beta$ subjects guess



Quotient diffusion 2PLM (Q-diffusion model)

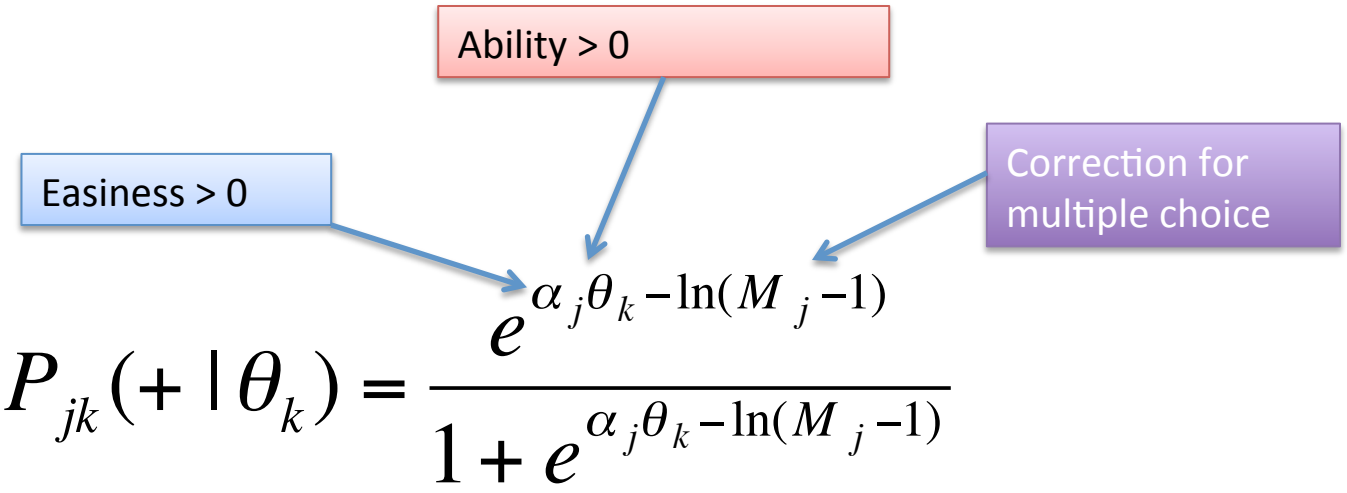


All *a* and *v* are positive

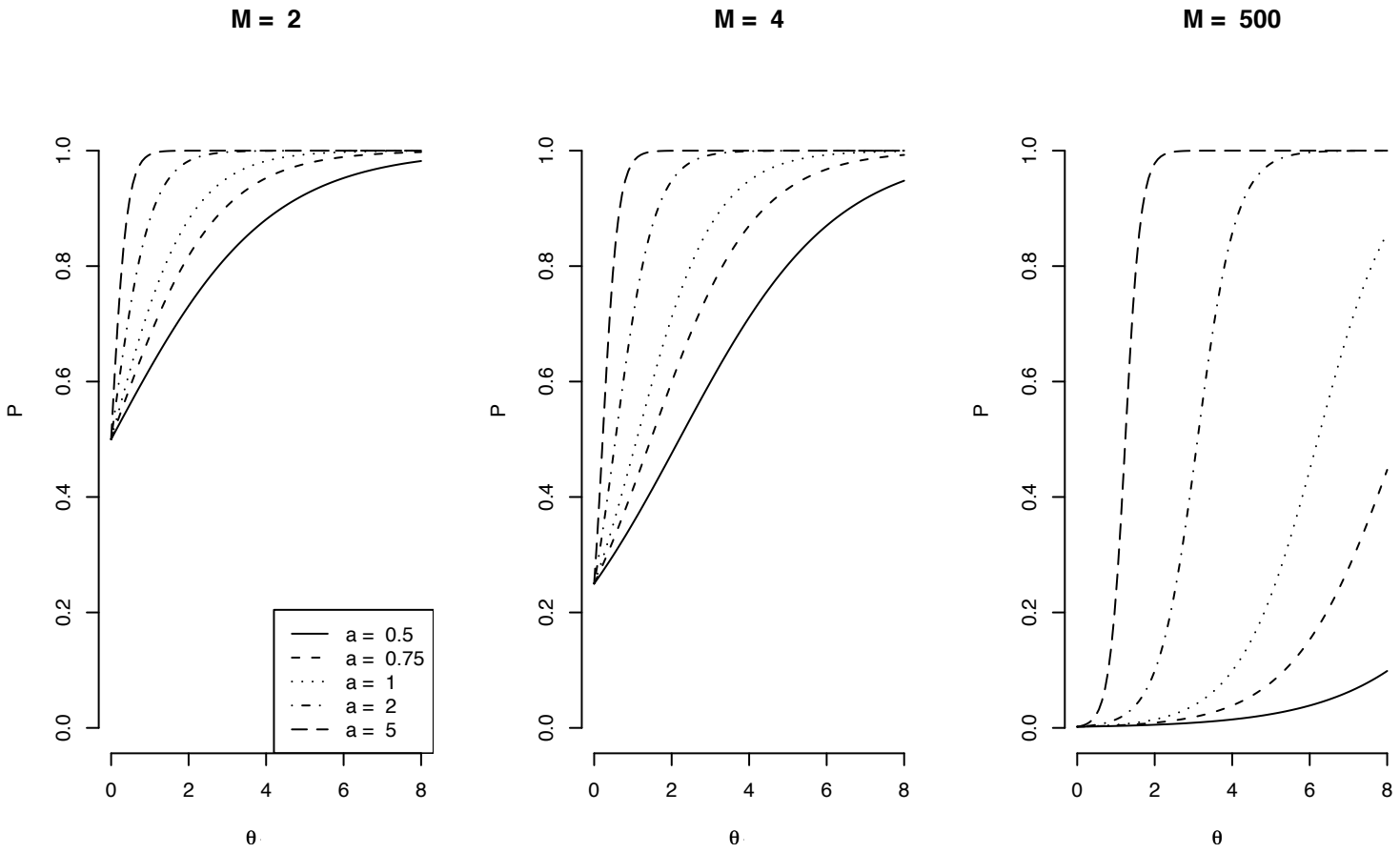
Simple case for data without RT's

$\theta_k = a_k^p v_k^p$ ⇒ Ignore a,v distinction (no RT's)

$\alpha_j = 1/a^i v^i_j$ ⇒ equal time limits for all item in test



ICC's of new guessing model



Advantages

- Process model for attitude and ability traits
- 'Mechanistic' interpretations of IRT parameters
  - θ is product of drift rate and boundary separation
    - Meaningful zero point (ratio properties)
    - Speed accuracy trade-off incorporated
  - α is easiness parameter, reciprocal of product of item drift rate and time pressure
  - β is intercept parameter, guessing probability
- Model of accuracy, but also model of RT
- Guessing explained by restricting the 2PL
- Simple extension to MC

D-diffusion model

Q-diffusion for ability

$$P_{jk}(+\,|\,\theta_k) = \frac{e^{\frac{a_k^p v_k^p}{a_j^i v_j^i} - \ln(M_j - 1)}}{1 + e^{\frac{a_k^p v_k^p}{a_j^i v_j^i} - \ln(M_j - 1)}}$$

D-diffusion for attitude

$$P_{jk}(+\,|\,\theta_k) = \frac{e^{\frac{a_k^p}{a_j^i}(v_k^p - v_j^i) - \ln(M_j - 1)}}{1 + e^{\frac{a_k^p}{a_j^i}(v_k^p - v_j^i) - \ln(M_j - 1)}}$$

a_j^i is item discrimination (time pressure)

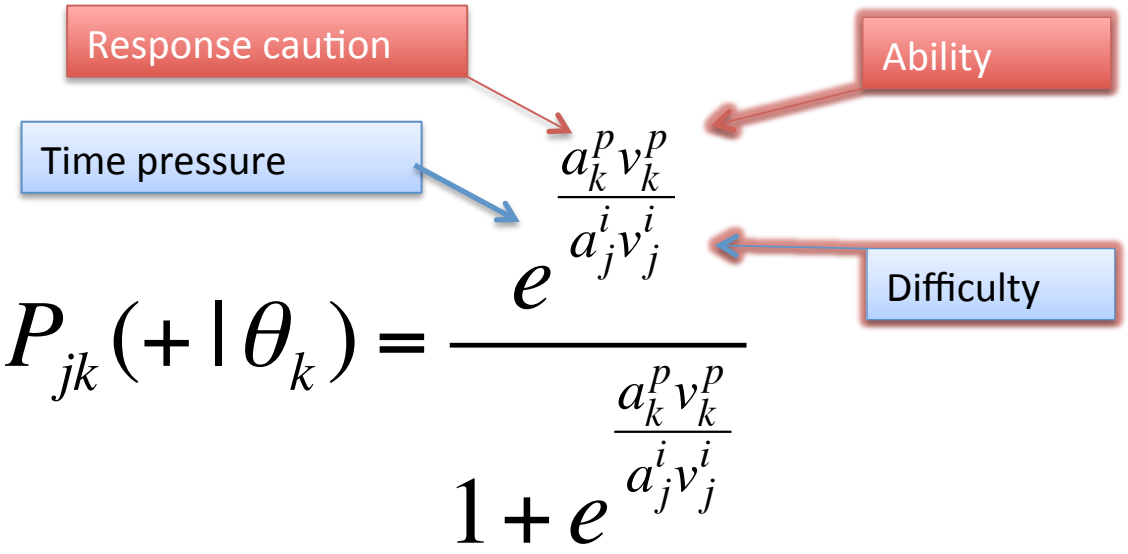
a_k^p is person discrimination (response caution)

Person fit

- Discrimination parameter in 2PL is an item parameter. Can be used to detect bad items.
- But also person can fit badly (e.g., cheaters)
- Requires to discrimination parameters a^p and a^i, but this model (Strandmark and Linn, 1987) is unidentified with accuracy data only
- *Note that the D-diffusion model with RT allows for person discrimination parameters* a^p/a^i

pauze

Quotient diffusion 2PLM (Q-diffusion model)



All *a* and *v* are positive

One-dimensionality

- The Q-diffusion model is a restricted 2PL models
  - β = guessing probabilty
  - α = difficulty
- In the Q-diffusion model:
  - *In the end all items will be passed, for all items, $P(+|\theta>0,a^i \rightarrow \infty)=1$, e.g. one Guttman item*

- If not (because the item also requires a jump) it measures another additional ability. The test is not unidimensional!
  - So any ability test with some items that I can solve and others that I will *never* solve (P_+ does not increase with longer time limit), tests for more than one ability
- *This implies a substantial (and not statistical definition) of one-dimensionality*
  - *Which is almost always too strict!*

Quantitative measurement

- Key is additive concatenation
  - 1cm + 1cm = 2 cm; 130kg-110kg=80kg-70kg
- Is this the case for the Rasch probability P+ ?
- Or for θ?

- But what about RT and v (drift rate)?
  - Time is quantitative!
  - Number of solved items is quantitative (is it?)

Interesting and unsolved puzzle!

Criticism

- Diffusion model is simplistic model of simple perceptual discriminations. It is too simple for complex processes involved in typical ability items in IRT
  - But 2PL is simplistic too
  - Better a simple model than no model at all
- We don't need a process model
  - Then IRT is curve fitting without explanatory value
- 3PL will fit better
  - 3PL is difficult to fit and uses many more parameters then the Q-diffusion IRT model
  - The point of the Q-diffusion IRT is not goodness of fit but theoretical clarity
- The MC extension does work well for the accuracy part of the model but not for response times
  - Use race models
- Open questions

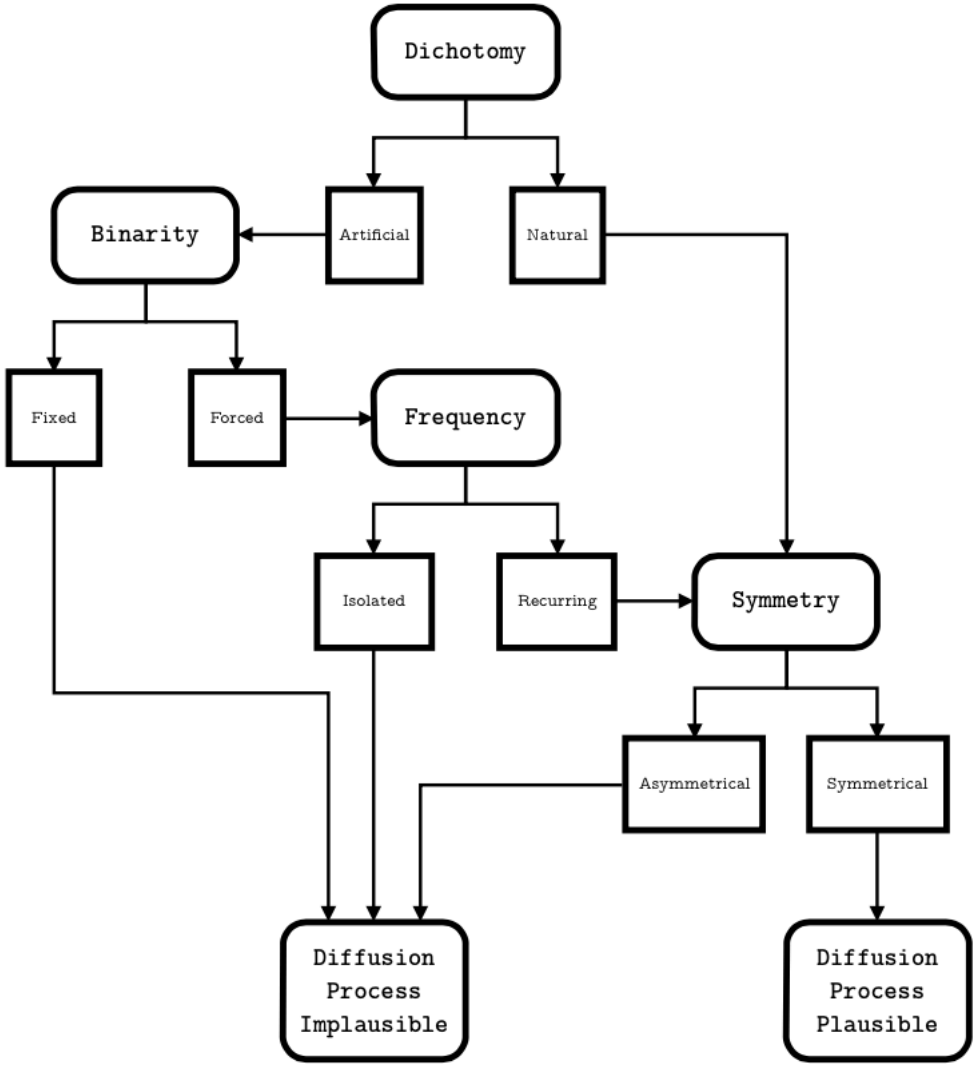


Figure 4. Decision Tree for Determining the Plausibility of a Diffusion Explanation based on Dichotomy, Binarity, Frequency and Symmetry

- Natural Dichotomy: Gender
- Artificial Dichotomy: 3+6 = 8 or 9
- Fixed binarity: binary scored MC
- Frequency recurring: answer option always 8 and 9
- Asymmetry: word/non word

Thus:

- Natural symmetrical case
- Artificial fixed recurring symmetrical case

are OK!

Joost Kruis, internship 2014

Relation to other IRT models

- Ramsay's Q model
- Guessing models
- van der Linden's IRT RT model

Ramsay's quotient model

- Ramsay (1989) investigates simple models for S_{kj}
  - Difference model ($\theta-\beta$): Rasch model
  - Quotient model (θ/β)

$$P_{jk}(+\mid\theta_k)=\frac{e^{S_{kj}}}{1+e^{S_{kj}}}$$

$$P_{jk}(+\mid\theta_k)=\frac{e^{\theta_k/\beta_j}}{K+e^{\theta_k/\beta_j}},\theta_k\geq 0,\beta_j>0$$

- Fits better than Rasch model and Rasch model with guessing in 4 examples

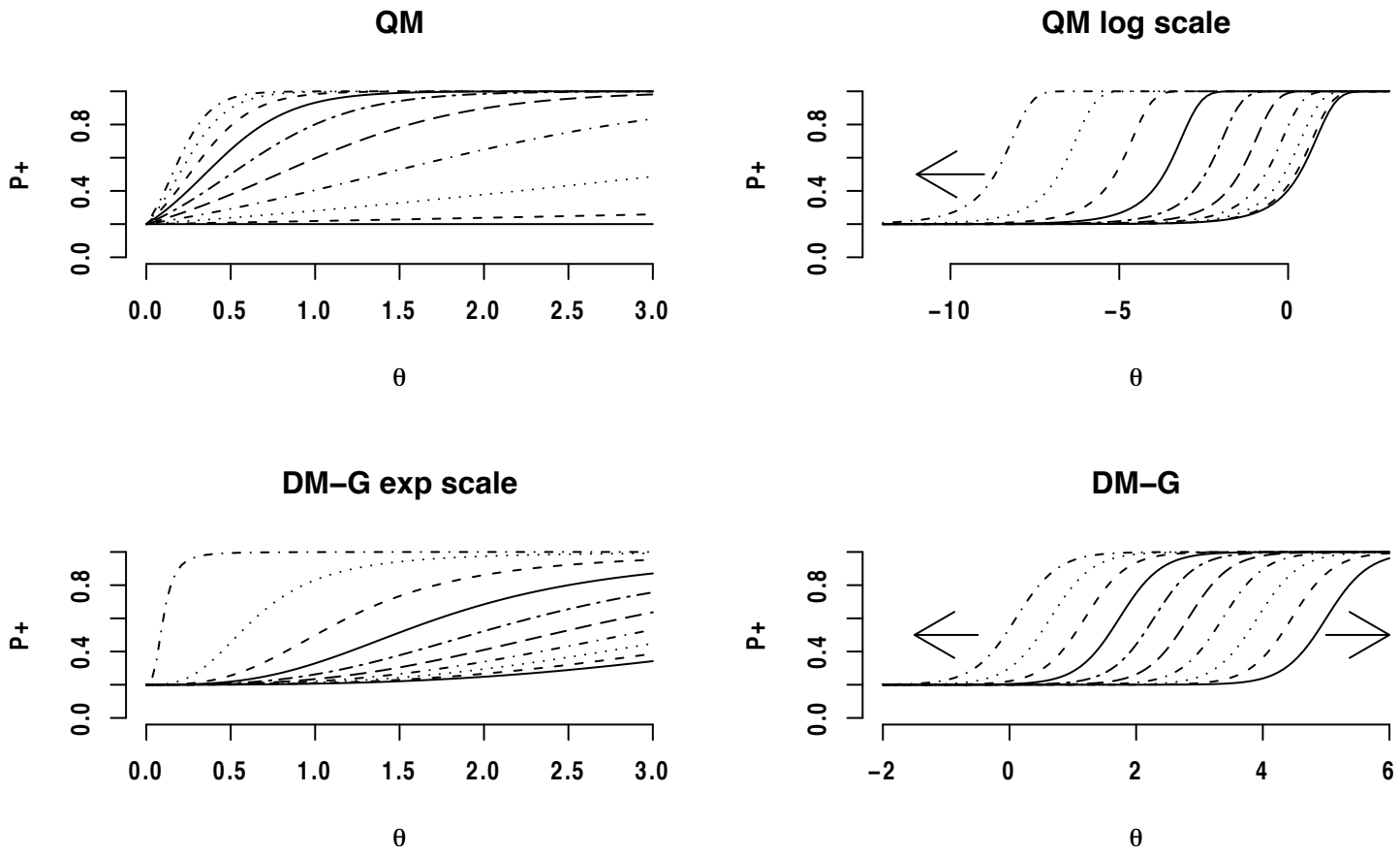
- if $\theta = a^p v^p$, $\beta = a^i v^i$ and K=M-1 the Q-model and the Q-diffusion model are equivalent
- Note: model in Cressie & Holland ('83) also equivalent:

$$P_{jk}(+\mid\theta_k)=\frac{ce^{e^{\theta_k^*-\beta_j^*}}}{(1-c)+ce^{e^{\theta_k^*-\beta_j^*}}}$$

Guessing

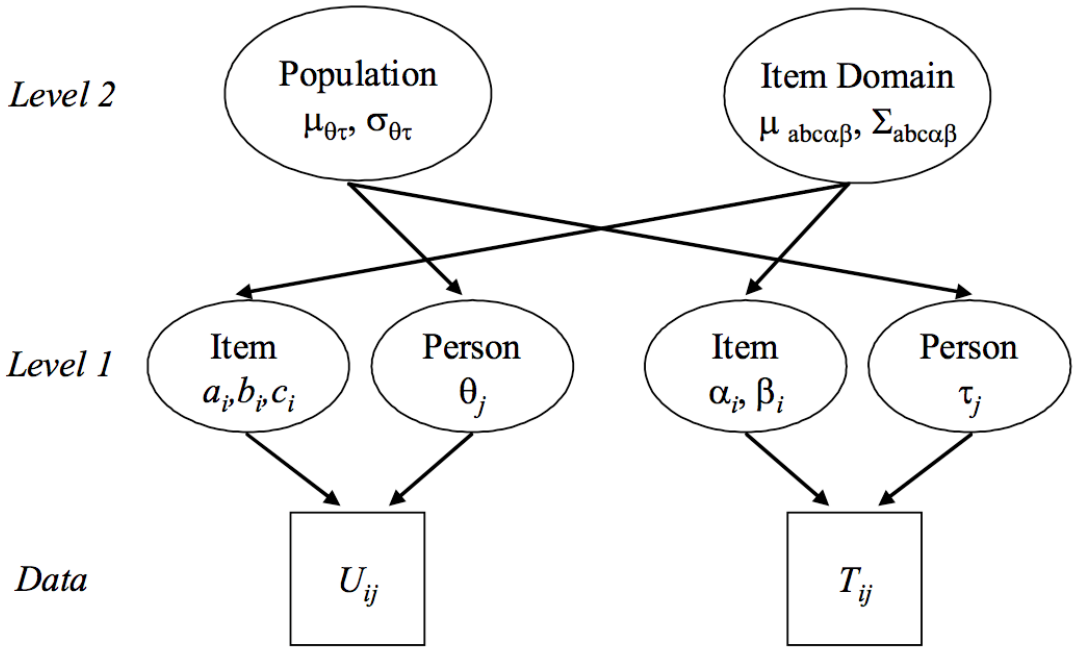
- Two process models
  - 3PL (p and g process)
  - 1PL-AG (p and ability dependent g process)
    - San Martin, del Pino, and De Boeck
  - DINA
- One state models
  - Difficulty dependent guessing
    - Hessen 2005, δ model
  - *Q-diffusion model: one state 'ability based' guessing*

QM versus DM-G



van der Linden's IRT model for RT

- General model: van der Linden's Hierarchical model
- Fundamental equation for RT modeling: $E(DT) = \frac{\xi_j^*}{\tau_k^*} \rightarrow E(\ln(DT) = \xi_j - \tau_k$



Translation of the item and person parameters of van der Linden's model to diffusion model parameters

$$E(DT) \approx \frac{a}{2v} = \frac{1}{2}\frac{a_k^p \,/\, a_j^i}{v_k^p \,/\, v_j^i} = \frac{1}{2}\frac{v_j^i \,/\, a_j^i}{v_k^p \,/\, a_k^p} = \frac{\xi_j^*}{\tau_k^*}; E(\ln(DT) \approx -\ln(2) + \ln\frac{v_j^i}{a_j^i} - \ln\frac{v_k^p}{a_k^p} = \xi_j - \tau_k$$

$$\xi_j = \ln\frac{v_j^i}{a_j^i}$$

$$\tau_k = \ln\frac{v_k^p}{a_k^p}$$

$$\theta_k = a_k^p v_k^p$$

$$\alpha_j = 1/a_j^i v_j^i$$

$$\beta_j = \ln(M_j - 1)$$

- If speed and ability parameters at the second level in van der Linden's model are positively correlated, then individual differences are primarily due to differences in drift rate (for an example, see van der Linden, 2007).
- If these parameters correlate negatively, the individual differences in drift rate are probably similar across subjects, and differences are mainly due to differences in response caution (see example 2 of Klein Entink, Fox, & van der Linden, 2009).

Summary

- A new IRT model for ability testing derived from a psychological process model
- $\theta=\beta=0$ obtains a new psychological meaning
  - $\theta \geq 0$ (no negative ability)
  - P(+)=1/M (guess)
  - if $\theta>\beta$ and d (time limit)$\rightarrow\infty$ then $P_{+}\rightarrow 1$
  - Gives a ratio aspect to IRT
  - Quantitative measurement
- In ability tests α is easiness parameter (no more β)
- Dichotomous scoring of MC items allowed by a simple extension of the 2PL model
- Guessing is allowed by restricting the 2PL model not by extending it!
- Integrated IRT model for accuracy and RT
- New definition of one-dimensionality
- R-package!

thanks

Note: Capacities versus abilities

- For Lumsden's sticks for measuring length the Guttman model is OK

- This also applies to STM
  - Digit span items