A Bayesian and Frequentist Multiverse Pipeline for MPT models

Applications to Recognition Memory

Henrik Singmann
(University of Warwick)

Daniel W. Heck
(University of Mannheim)

Marius Barth
(Universität zu Köln)

Julia Groß
(University of Mannheim)

Beatrice G. Kuhlmann
(University of Mannheim)

Multiverse Approach

- Statistical analysis usually requires several (more or less) *arbitrary* decisions between reasonable alternatives:
    - Data processing and preparation (e.g., exclusion criteria, aggregation levels)
    - Analysis framework (e.g., statistical vs. cognitive model, frequentist vs. Bayes)
    - Statistical analysis (e.g., testing vs. estimation, fixed vs. random effects, pooling)
- Combination of decisions spans **multiverse** of data and results (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016):
    - Usually one path through multiverse (or '*garden of forking paths*', Gelman & Loken, 2013) is reported
    - *Valid conclusions cannot be contingent on arbitrary decisions*
- **Multiverse analysis** attempts exploration of possible results space:
    - Reduces problem of selective reporting by making fragility or robustness of results transparent.
    - Conclusions arising from many paths are more credible than conclusions arising from few paths.
    - Helps identification of most consequential choices.
- **Limits of multiverse approach:**
    - Implementation not trivial.
    - Results must be commensurable across multiverse (e.g., estimation versus hypothesis testing).

Current Project

- DFG *Scientific Network* grant to Julia Groß and Beatrice Kuhlmann
  - "Hierarchical MPT Modeling – Methodological Comparisons and Application Guidelines"
  - 6 meetings over 3 years with 15 people plus external experts
  - Multinomial processing tree (MPT) models: class of discrete-state cognitive models for multinomial data (Riefer & Batchelder, 1988)
  - MPT models traditionally analysed with frequentist methods (i.e., χ^2/G^2) and aggregated data
  - *Several hierarchical-Bayesian approaches exist. Do we need those?*
- *Today*: First results from model for recognition memory

Our Multiverse

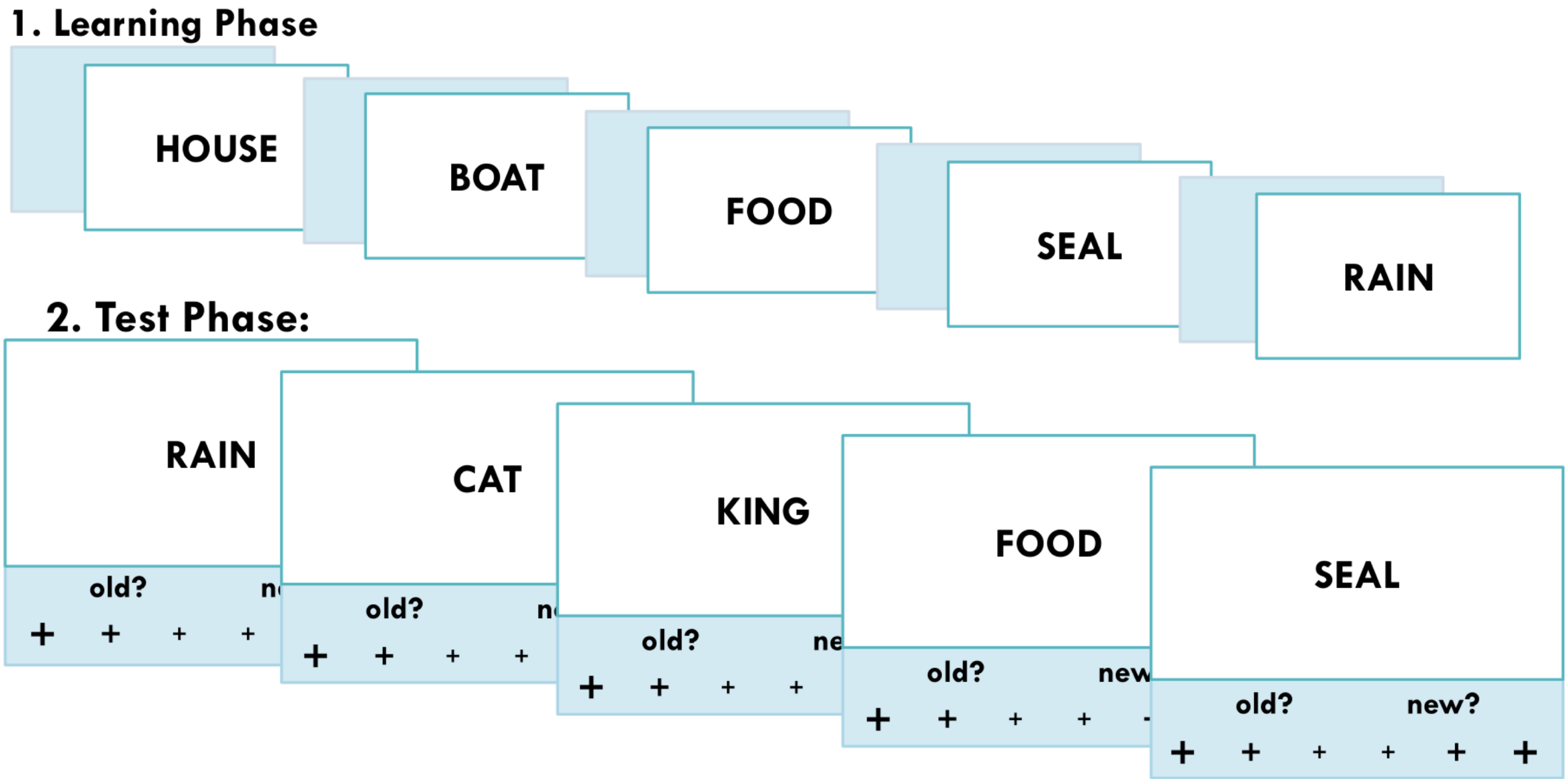
- Statistical framework:
  - Frequentist (i.e., maximum-likelihood)
  - Bayesian (i.e., MCMC)
- Pooling:
  - Complete pooling (aggregated data)
  - No pooling (individual-level data)
  - Partial pooling (hierarchical-modelling): Individual-level parameters with group-level distribution
- Results:
  1. ***Parameter point estimates*: MLE and posterior mean**
  2. *Parameter uncertainty*: ML-SE and MCMC-SE
  3. *Model adequacy*: G^2 *p*-value and posterior predictive *p*-value (Klauer, 2010)

Our Multiverse Pipeline (in **R**)

- Frequentist (uses `MPTinR`; Singmann & Kellen, 2013):
  - (1) Traditional approach: Frequentist asymptotic complete pooling
  - (2) Frequentist asymptotic no pooling
  - (3) Frequentist no-pooling with parametric bootstrap
  - (4) Frequentist no-pooling with non-parametric bootstrap
- Bayesian (uses `TreeBUGS`; Heck, Arnold, & Arnold, 2018):
  - (5) Bayesian complete pooling (custom C++ sampler)
  - (6) Bayesian no pooling (unique method, custom C++ sampler)
  - (7) Bayesian partial pooling I a, Jags: Beta-MPT Jags (Smith & Batchelder, 2010)
  - (8) Bayesian partial pooling I b, C++: Beta-MPT C++
  - (9) Bayesian partial pooling II: Latent trait MPT (Klauer, 2010; Jags)
  - (10) Bayesian partial pooling III: Latent trait MPT w/o correlation parameters (Jags)
- (11) latent-class approach (Klauer, 2006)

- All implemented in R packages
  - `MPTmultiverse`: (1)-(10): https://cran.r-project.org/package=MPTmultiverse
  - `hmpt` (for latent-class only): https://github.com/mpt-network/hmpt

Example Application: Recognition Memory

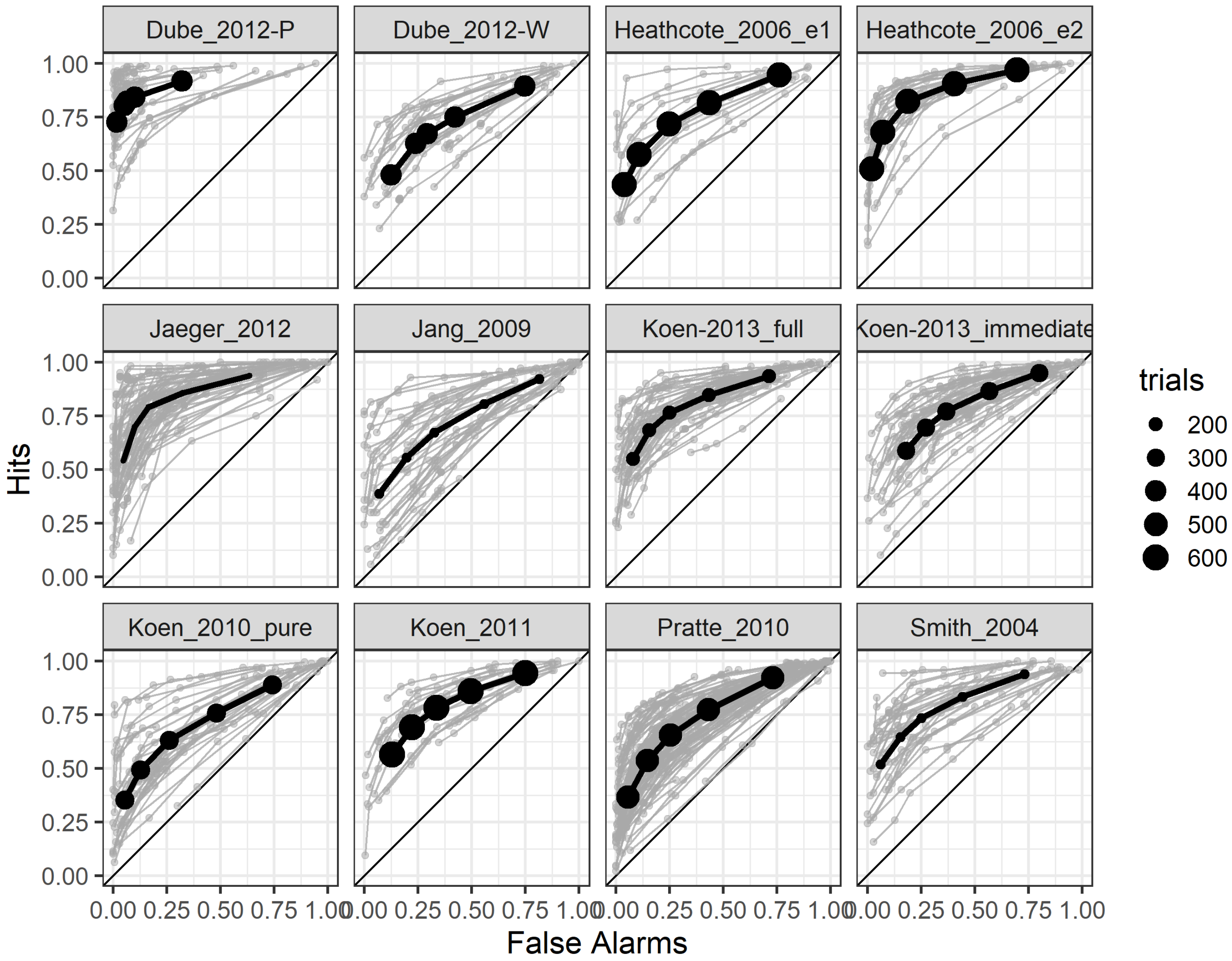


6-point ROCs

Data set	Sample *N*	Mean No. of Trials
Dube & Rotello (2012, E1, Pictures (P))	27	400
Dube & Rotello (2012, E1, Words (W))	22	400
Heathcote et al. (2006, Exp. 1)	16	560
Heathcote et al. (2006, Exp. 2)	23	560
Jaeger et al. (2012, Exp. 1, no cue)	63	120
Jang et al. (2009)	33	140
Koen & Yonelinas (2010, pure study)	32	320
Koen & Yonelinas (2011)	20	600
Koen et al. (2013, Exp. 2, full attention)	48	200
Koen et al. (2013, Exp. 4, immediate test)	48	300
Pratte et al. (2010)	97	480
Smith & Duncan (2004, Exp. 2)	30	140

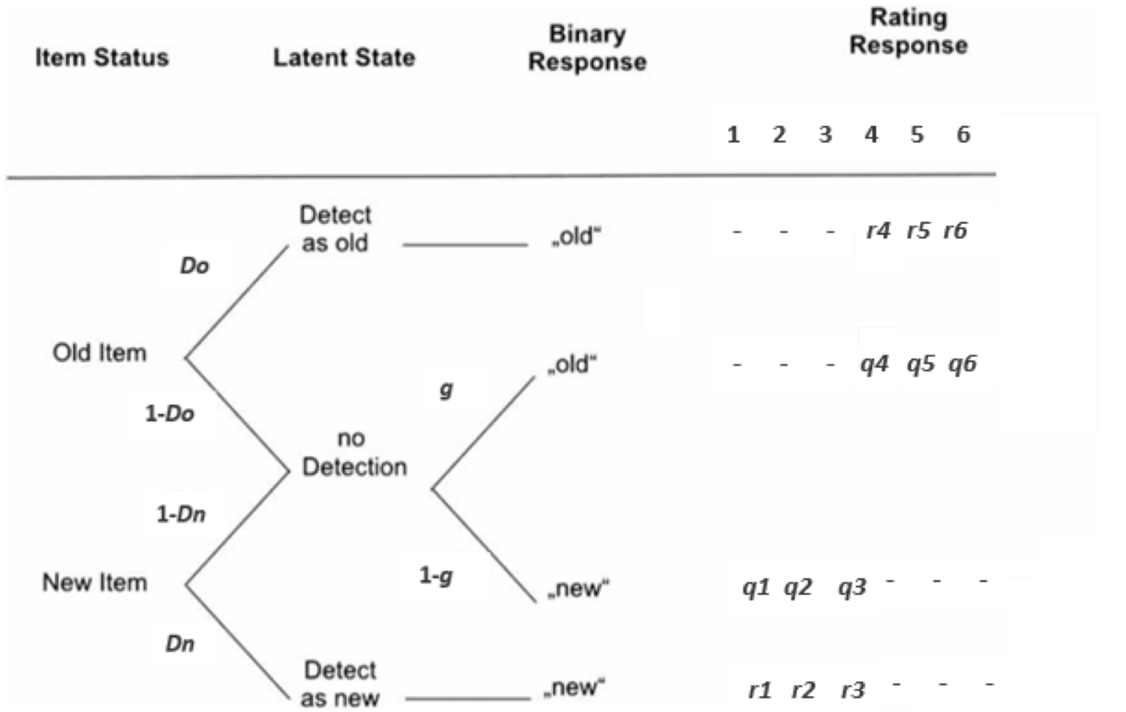
- Total *N* = 459
- Mean trials = 350

- Still missing:
  - 8-point ROCs
  - 4 studies
  - Total *N*: 459
  - reasons: more model variants possible, takes longer

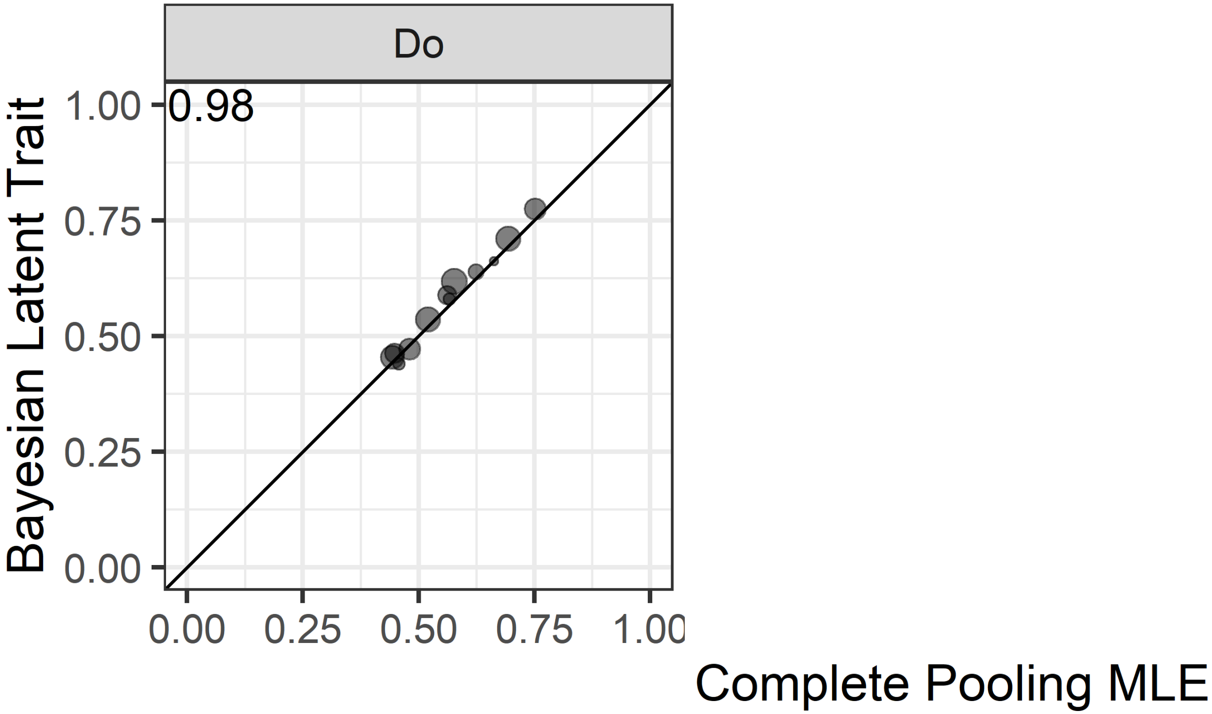


Dube_2012-P
Dube_2012-W
Heathcote_2006_e1
Heathcote_2006_e2
Jaeger_2012
Jang_2009
Koen-2013_full
Koen-2013_immediate
Koen_2010_pure
Koen_2011
Pratte_2010
Smith_2004
Hits
False Alarms
0.00
0.25
0.50
0.75
1.00
trials
200
300
400
500
600

2-high threshold model (2HTM) for 6-point confidence-rating data
(e.g., Bröder, et al., 2013)

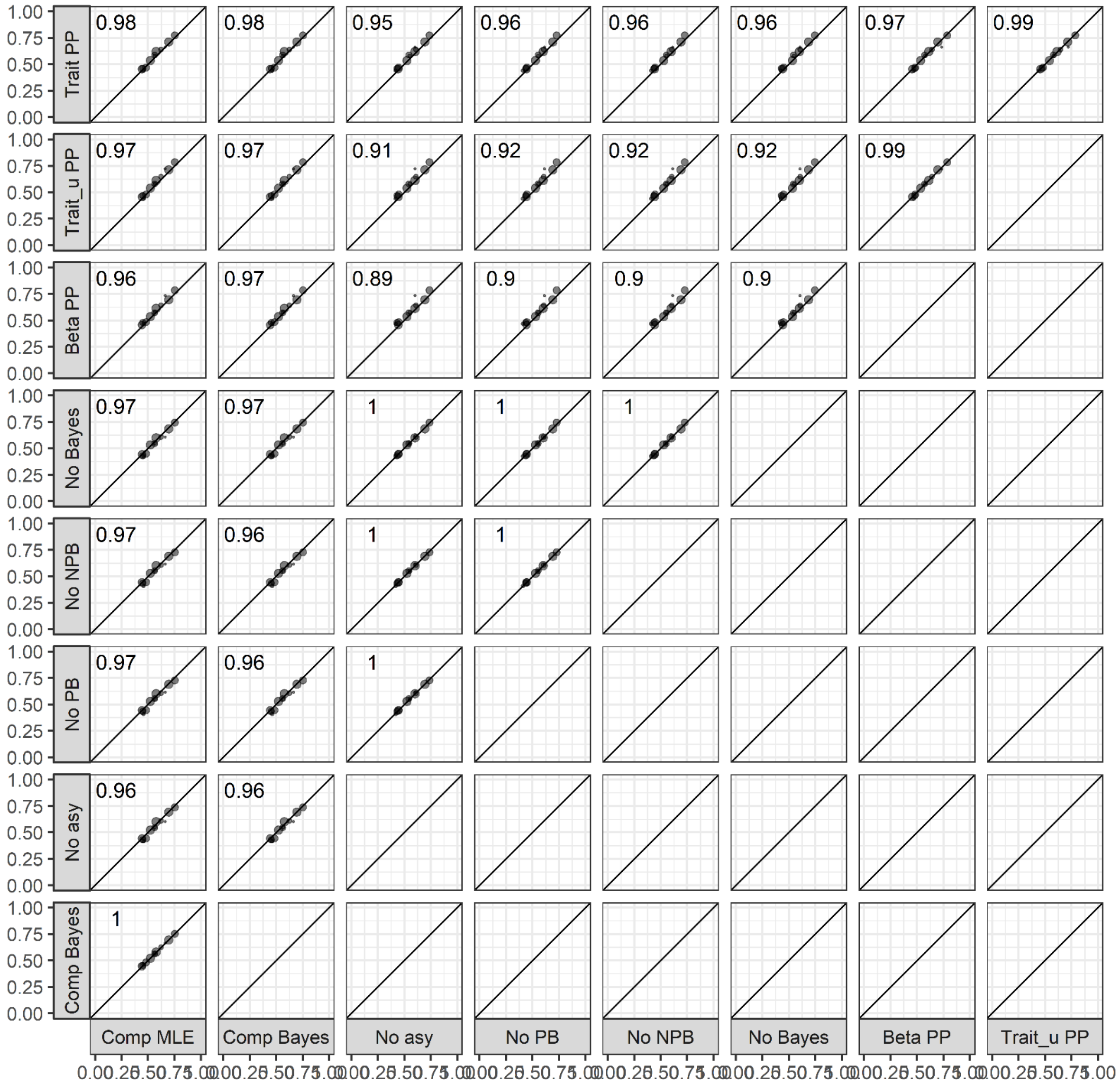


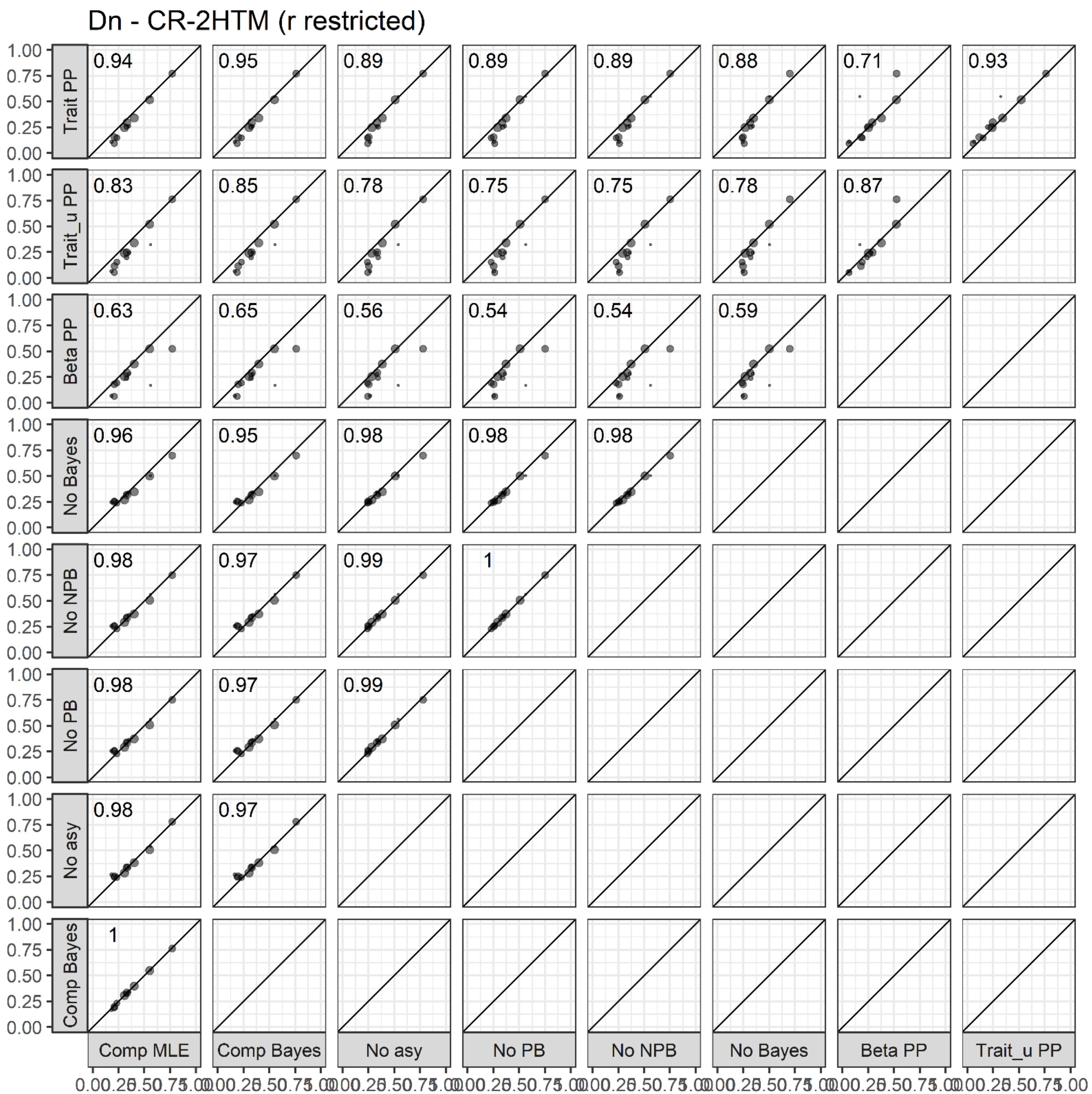
- 3 core parameters:
  - Do: Probability to detect old item as old
  - Dn: Probability to detect new item as new
  - g: Probability to guess item is old (conditional on non-detection)
- 8 response mapping parameters (at least one needs to be equated for identifiability)

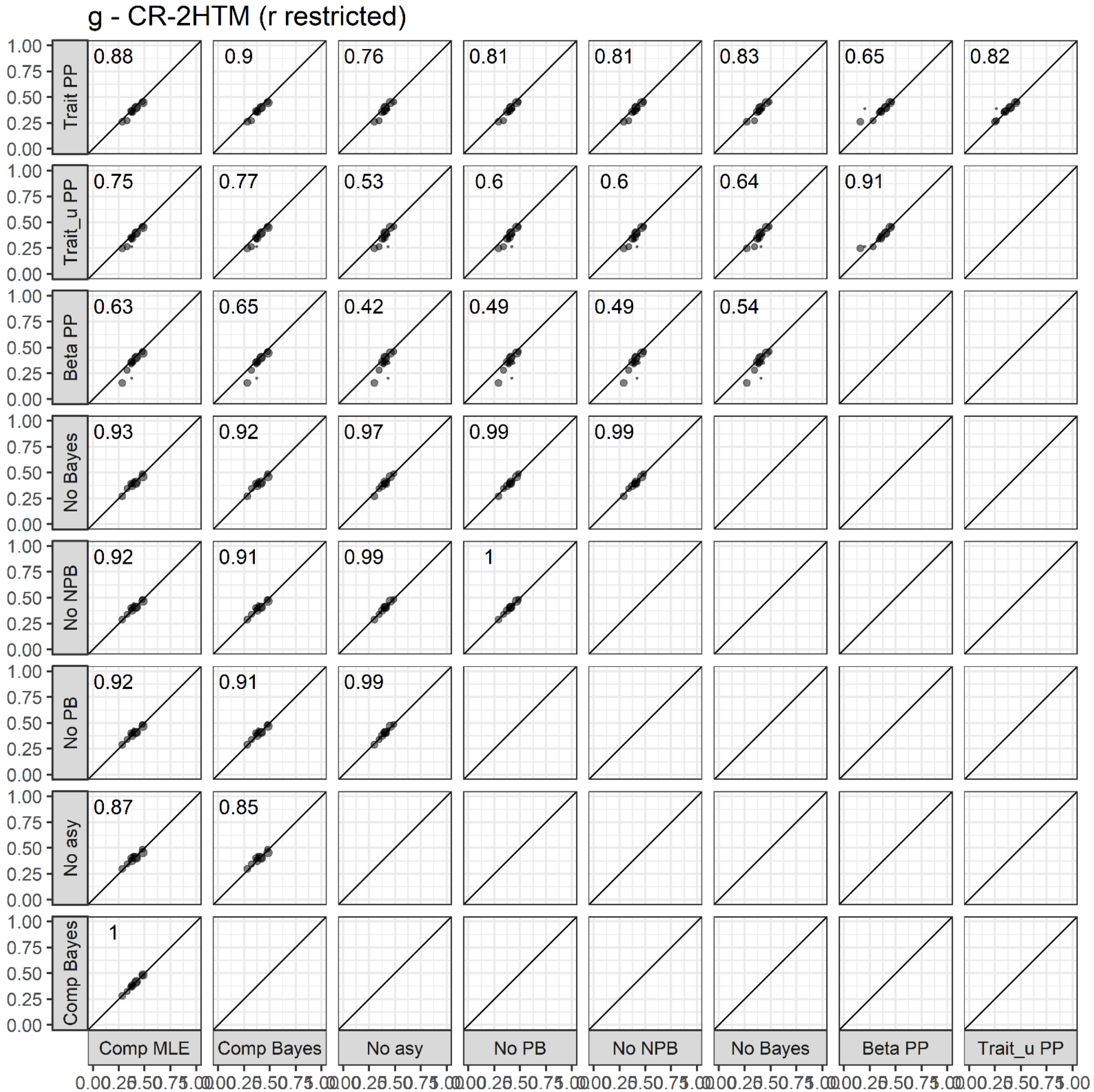


- Each data point is one study
- Value corner: *CCC* = concordance correlation coefficient (measure of absolute agreement)

Do - CR-2HTM (r restricted)

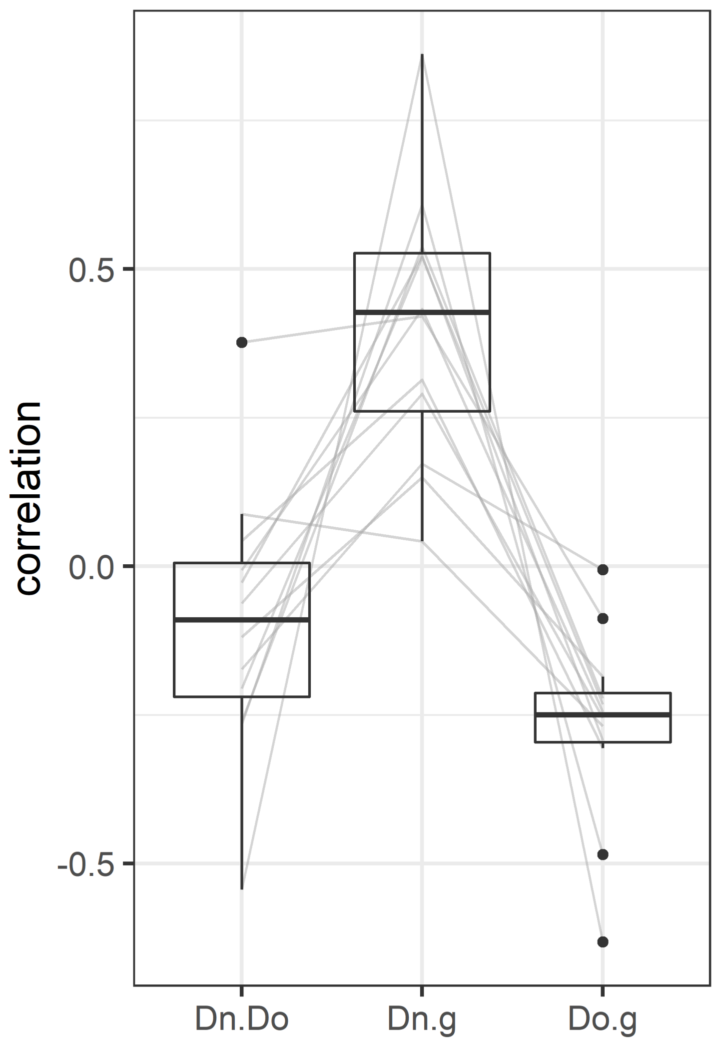



Dn - CR-2HTM (r restricted)
Trait PP
Trait_u PP
Beta PP
No Bayes
No NPB
No PB
No asy
Comp Bayes
Comp MLE
Comp Bayes
No asy
No PB
No NPB
No Bayes
Beta PP
Trait_u PP
0.94 0.95 0.89 0.89 0.89 0.88 0.71 0.93
0.83 0.85 0.78 0.75 0.75 0.78 0.87
0.63 0.65 0.56 0.54 0.54 0.59
0.96 0.95 0.98 0.98 0.98
0.98 0.97 0.99 1
0.98 0.97 0.99
0.98 0.97
1



g - CR-2HTM (r restricted)
Trait PP
Trait_u PP
Beta PP
No Bayes
No NPB
No PB
No asy
Comp Bayes
Comp MLE
Comp Bayes
No asy
No PB
No NPB
No Bayes
Beta PP
Trait_u PP
0.88
0.9
0.76
0.81
0.81
0.83
0.65
0.82
0.75
0.77
0.53
0.6
0.6
0.64
0.91
0.63
0.65
0.42
0.49
0.49
0.54
0.93
0.92
0.97
0.99
0.99
0.92
0.91
0.99
1
0.92
0.91
0.99
0.87
0.85
1

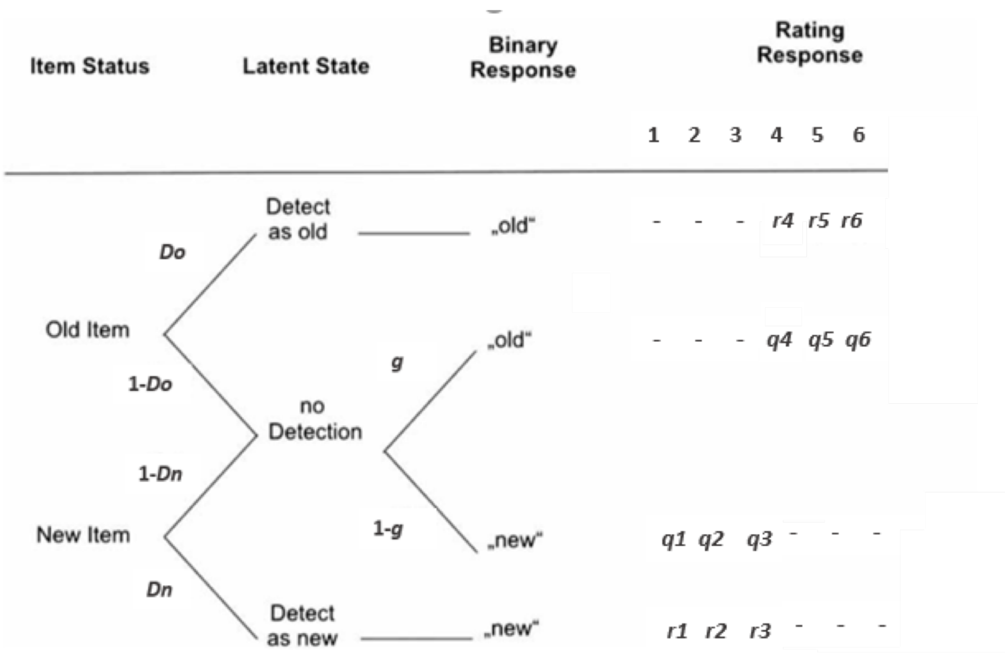
Parameter trade-off/fungibility (based on Latent-trait group-level posteriors)



Conclusions

- Overall, methods appear to agree with each other (maximal difference ≈ .25)
- Agreement between estimation method, depends on parameter
  - Even for structurally very similar core parameters (*Do* versus *Dn*) we see differences
  - Compared to latent trait methods, *Dn* and *g* are slightly **underestimated** by other methods
  - Effect of over/under-estimation does not appear to be related to sample size
- Parameters that show imprecision in estimation, seem to show larger parameter trade-offs
- *Recommendation*: Use multiverse approach to take uncertainty of modeling framework into account

2-high threshold model (2HTM) for 6-point confidence-rating data
(e.g., Bröder, et al., 2013)



$$r1 = r_1$$
$$r2 = (1 - r_1)r_2$$
$$r3 = (1 - r_1)(1 - r_2)$$
same for q

$$r6 = r_6$$
$$r5 = (1 - r_6)r_5$$
$$r4 = (1 - r_6)(1 - r_5)$$
same for q

- For r, r_1 and r_6 expected to have most data
- For q, q_5 and q_2 expected to have most data

- Data provides 10 independent data points, full model has 11 free parameters:
  - 3 core parameters: Dn, Do, and g.
  - 8 response mapping parameters
- For parameter identifiability: at least one parameter needs to be equated

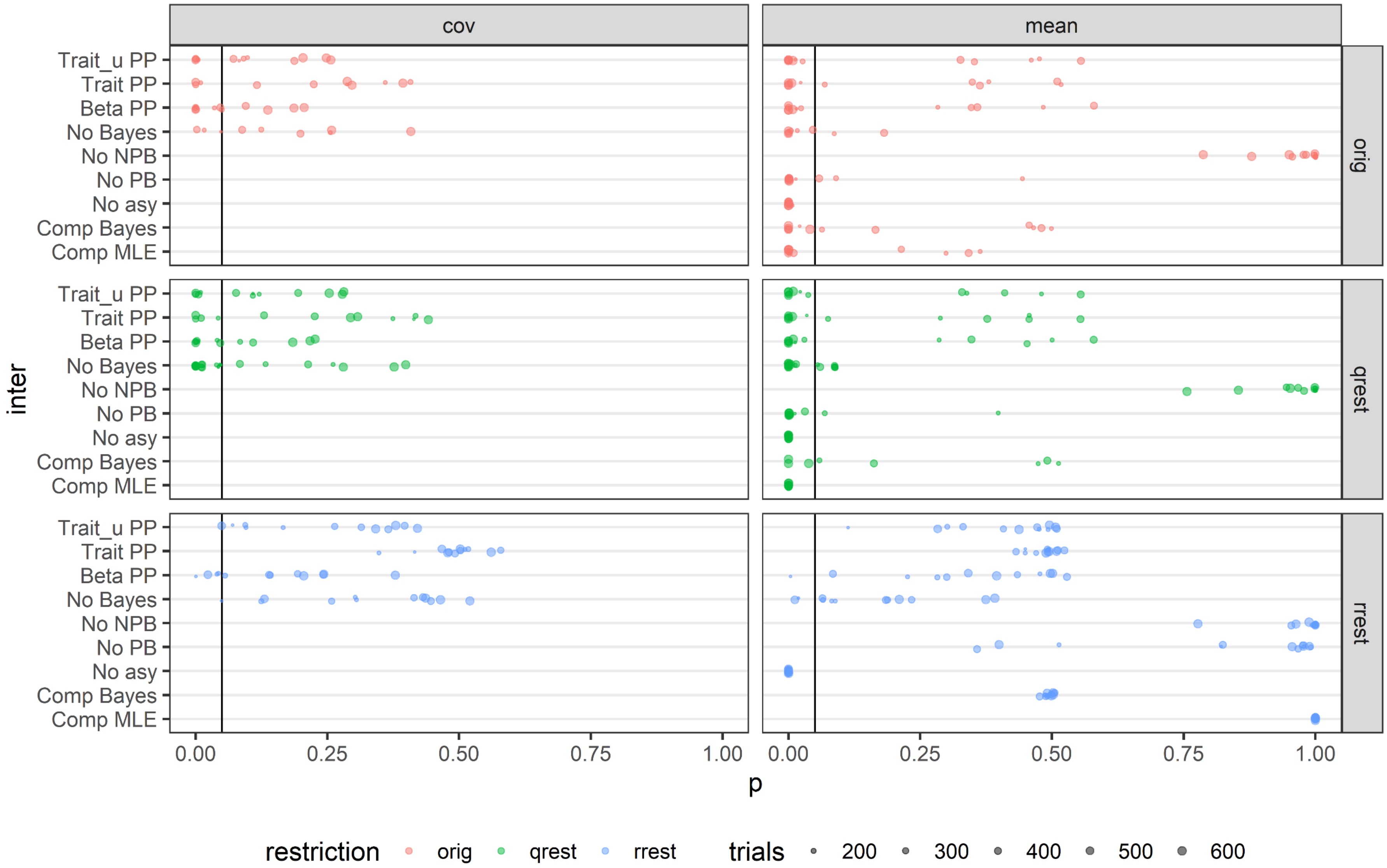
Identifiability Restrictions

$r1 = r_1$
$r2 = (1 - r_1)r_2$
$r3 = (1 - r_1)(1 - r_2)$
same for q

$r6 = r_6$
$r5 = (1 - r_6)r_5$
$r4 = (1 - r_6)(1 - r_5)$
same for q

- original Bröder et al. (2013) variant:
  - $q_5 = q_2$
  - $r_5 = r_2$
- only *q*-restricted:
  - $q_5 = q_2$
- only *r*-restricted:
  - $r_5 = r_2$

- original Bröder et al. (2013) variant:
  - $q_5 = q_2$
  - $r_5 = r_2$
- only *q*-restricted:
  - $q_5 = q_2$
- only *r*-restricted:
  - $r_5 = r_2$


cov
mean
orig
qrest
rrest
Trait_u PP
Trait PP
Beta PP
No Bayes
No NPB
No PB
No asy
Comp Bayes
Comp MLE
inter
0.00
0.25
0.50
0.75
1.00
p
restriction
orig
qrest
rrest
trials
200
300
400
500
600