

Performance Testing and Certification using Item-Response Theory

Stéphanie van den Berg, Phd
University of Twente, the Netherlands

Topics

- Central ideas of classical and modern test theory
- Applications: Measuring canine aggression and fearfulness
- Optimizing performance testing and certification
- IRT, genetics, and breeding

Classical test theory

- ObservedScore = TrueScore + Error
- Sum scores: $ObservedScore = \sum_{i=1}^k ItemScore_i$
- No validity without reliability
- Lower bound estimates of reliability

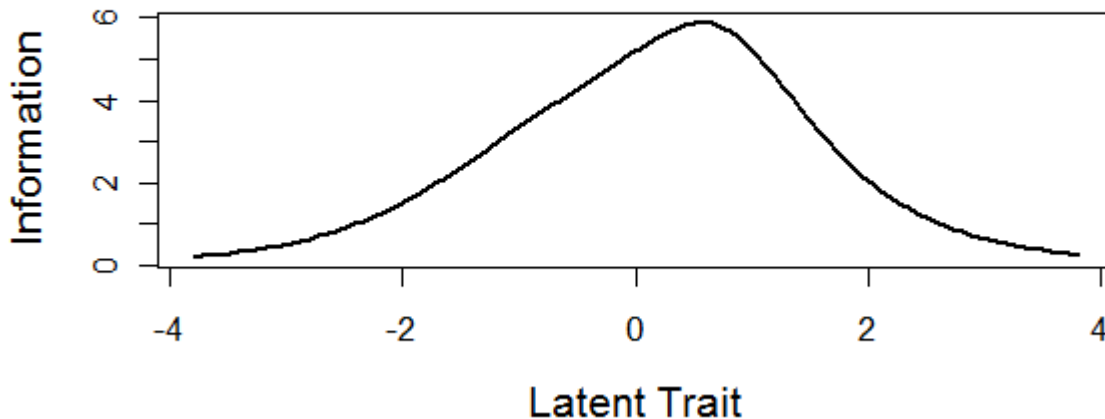
(Sijtsma, Psychometrika, 2009, 74(1): 107–120)

- Cronbach's alpha
- Lambda-2

Modern test theory

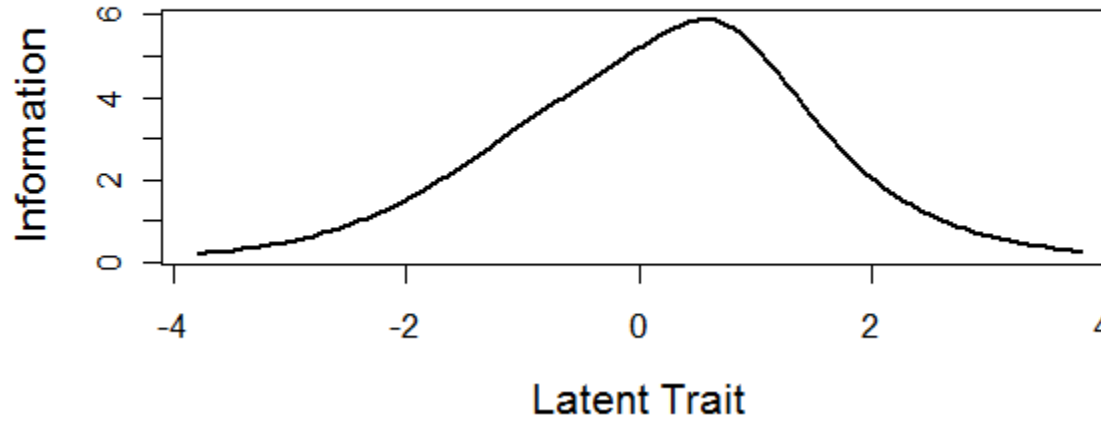
- $\text{ObservedScore}_i = \text{TrueScore}_i + \text{Error}_i$
- Estimated scores: $\hat{\theta} = \theta + \epsilon$
- Measurement error variance depends on θ

CAPE negative

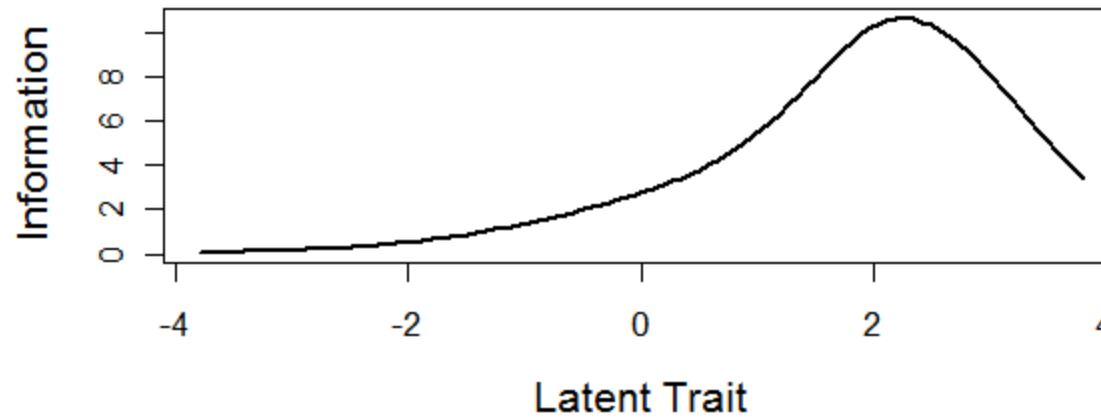


$$\text{Var}(\epsilon) = \frac{1}{\text{Information}}$$

CAPE negative



CAPE positive



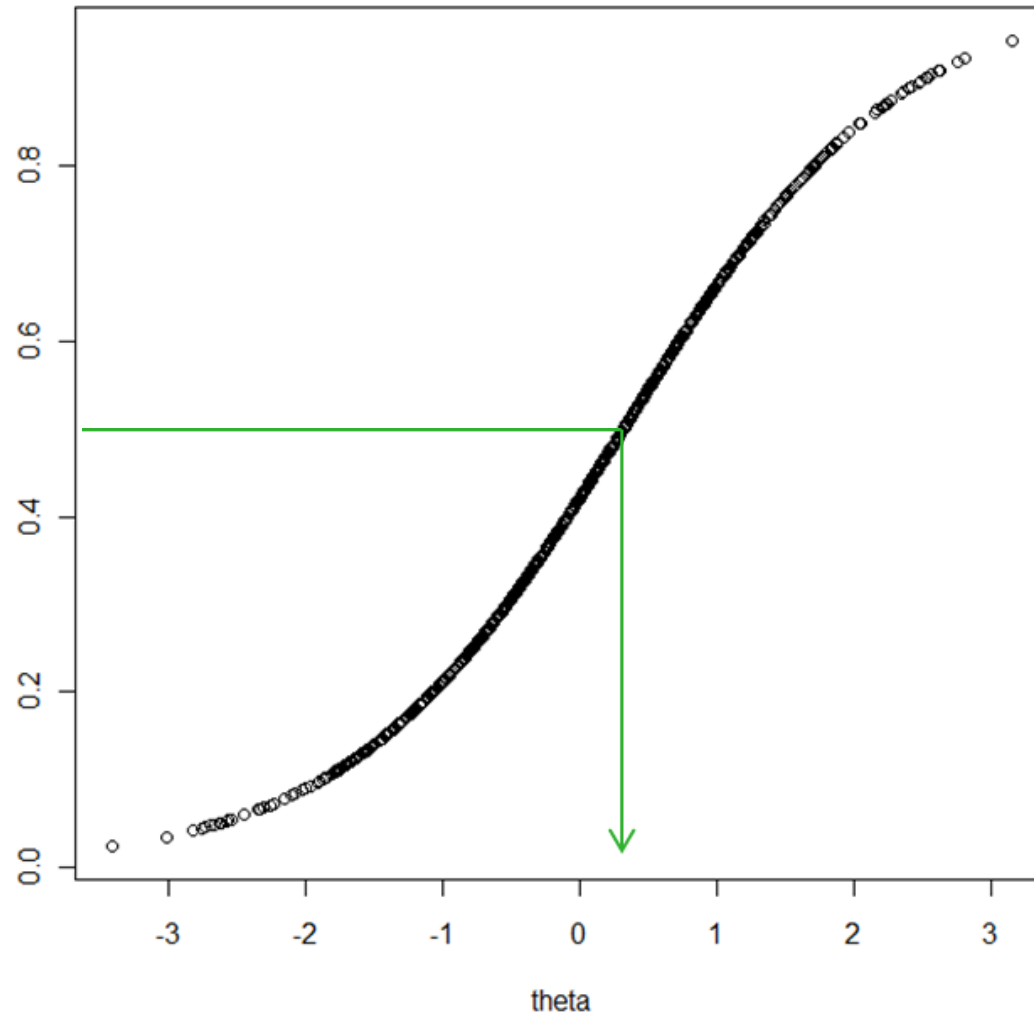
Item Response theory models

- Modelling probabilities of certain responses on test items (subtests)
 - For example: pass/fail, present/absent,
- Probabilities are a function of dog characteristics and item characteristics.

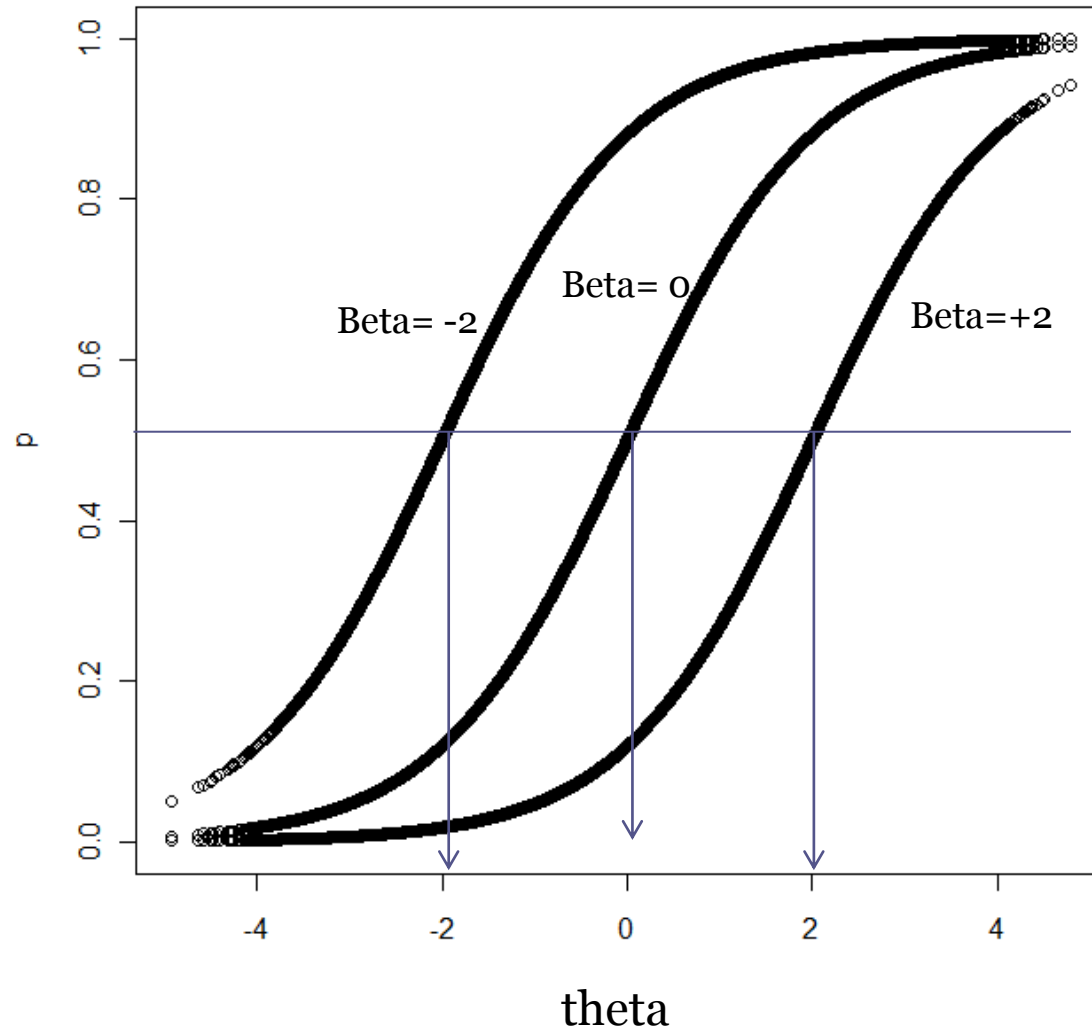
- $$p(\text{pass}|\theta_i, \beta_j) = \frac{1}{1 + \exp(\beta_j - \theta_i)}$$

- $$\text{logit}[p(\text{pass}|\theta_i, \beta_j)] = \log \left[\frac{p(\text{pass}|\theta_i, \beta_j)}{p(\text{fail}|\theta_i, \beta_j)} \right] = \theta_i - \beta_j$$

- $p(\text{pass}|\theta_i, \beta_j)$



- $p(\text{pass}|\theta_i, \beta_j)$

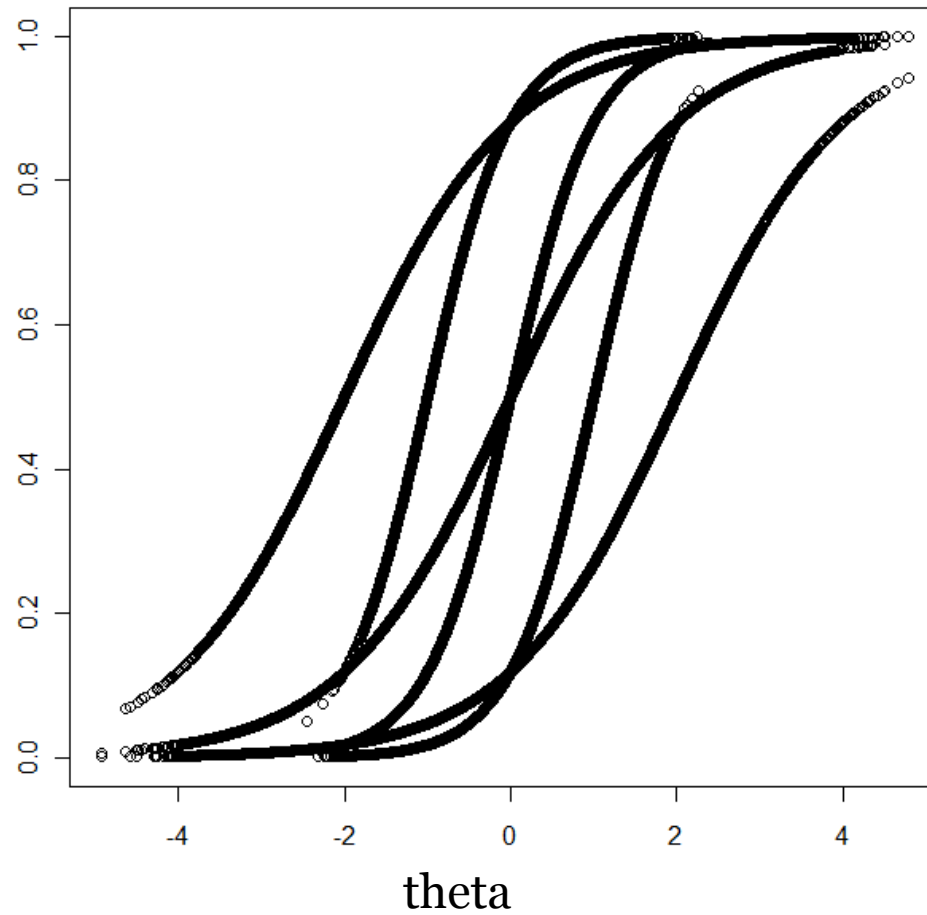


$$\text{Logit}(p) = \alpha\theta - \beta$$



Discrimination
parameter

$$p(\text{pass}|\theta_i, \beta_j, \alpha_j)_\alpha$$



Application to canine aggression



Contents lists available at ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim



Evaluation of the C-BARQ as a measure of stranger-directed aggression in three common dog breeds

Stéphanie M. van den Berg^{a,*}, Henri C.M. Heuven^{b,1}, Linda van den Berg^{c,2}, Deborah L. Duffy^{d,3}, James A. Serpell^{d,3}

Table 4

Parameter estimates (SE) and Lagrange multiplier tests for DIF across breeds based on the 1-parameter model.

Item	β parameter	LM	df	p	Abs. dif
1	1.955 (0.168)	0.64	2	0.73	0.00
2	3.023 (0.180)	9.22	2	0.01*	0.02
3	0.483 (0.145)	6.66	2	0.04*	0.02
4	0.540 (0.143)	3.01	2	0.22	0.01
5	1.476 (0.161)	5.07	2	0.08	0.02
6	-0.945 (0.145)	0.99	2	0.61	0.01
7	-1.225 (0.155)	10.58	2	0.01*	0.03
8	2.373 (0.163)	1.28	2	0.53	0.01
9	-0.510 (0.148)	2.95	2	0.23	0.02
10	1.244 (0.154)	3.74	2	0.15	0.02

* Statistically significant at an alpha of 0.05.

Table 5

Means and standard deviations of individual θ parameters based on the 1-parameter model.

	Mean (SE)	Standard deviation
German Shepherds	0 (fixed)	2.658 (0.147)
Golden Retrievers	-2.206 (0.253)	2.666 (0.217)
Labrador Retrievers	-1.402 (0.211)	3.329 (0.185)

Application to fearfulness

Comparing German Shepherds, Golden Retrievers and Irish soft-coated wheaten terriers on DMA fearfulness subscale

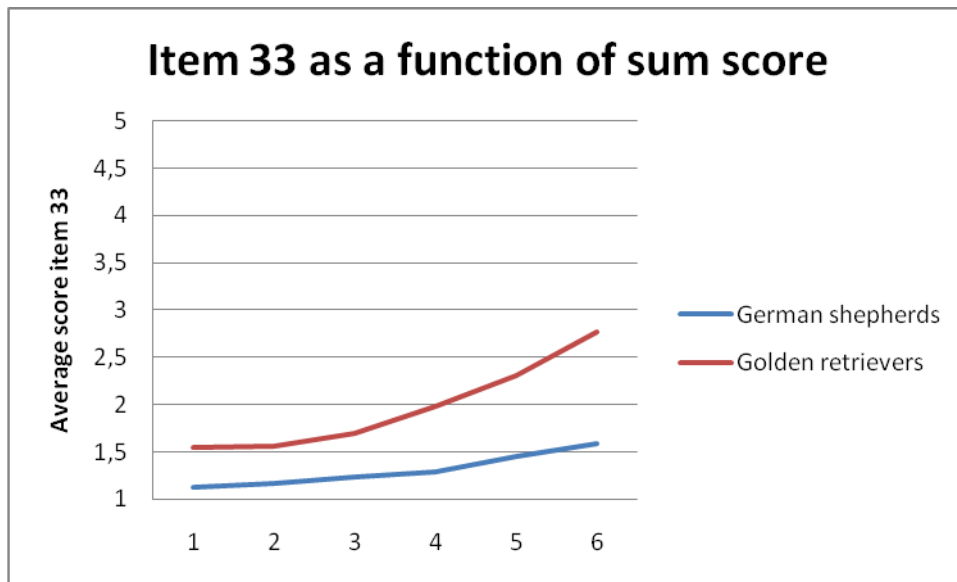
Table 2.2 Overall DIF across breeds

	LM	Abs dif
17 Överr rädsla	359.10	0.19
19 Överr nyfiken*	33.94	0.04
20 Överr kvars nyfiken	92.47	0.08
22 Ljudkänsl rädsla	285.48	0.12
23 Ljudkänsl nyfiken*	168.77	0.19
24 Ljudkänsl kvars rädsla	16.84	0.03
33 Skott	1103.19	0.43

No measurement invariance across breeds, mainly due to wheaten terriers, and particularly the reaction to gun shots

Application to fearfulness

Comparing German Shepherds and Golden Retrievers



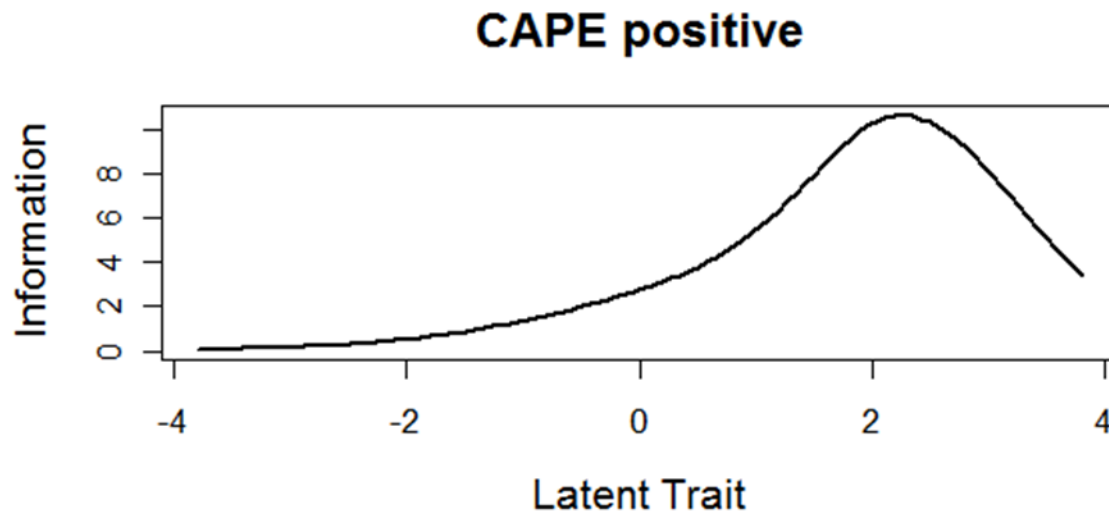
Corrected for overall fearfulness differences, German shepherds are less responsive to gun shots than Golden Retrievers

Results

- Good fit for a 1-parameter IRT model for stranger-directed aggression and 2-parameter model for fearfulness
- Breeds differ on *overall* level of stranger-directed aggression.
- Clear breed differences for expression of fear
- IRT-models seem promising tools for canine behaviour

Optimizing performance testing and certification

- Where do you want to have high levels of information (small measurement error)?
 - Across the entire scale?
 - Around the cut-off point?



Most items have beta parameters values around 2.

Conclusion

- IRT able to quantify the information provided by a specific item in a certain population.
- IRT helps you determine what type of items you need to optimize your test, given its goal.
- General principle: for high cut-off points, use items with low endorsement levels (positive beta parameters), and vice versa

IRT, genetics, and breeding

- Determining heritability, corrected for measurement error
- Selecting the best dogs in a breeding program
- Finding “genes for aggression”: use items with widely different response rates, gives best information across the entire scale.

Heritability of fearfulness in Irish soft-coated wheaten terriers

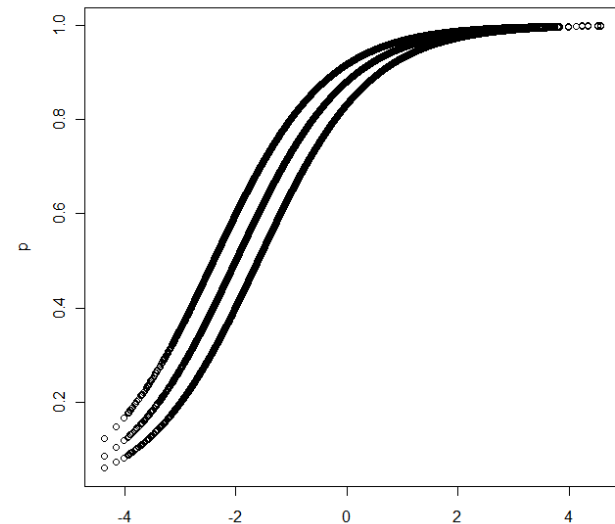
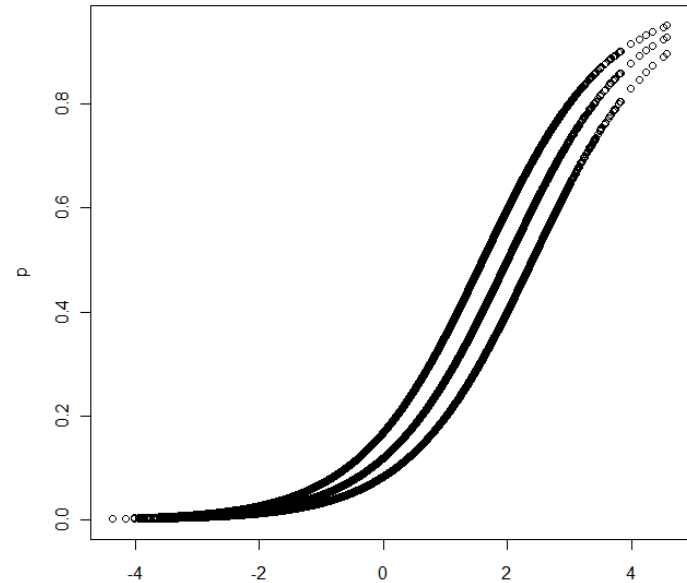
Table 1: Posterior means and standard deviations of genetic parameters with and without an IRT model.

Parameter	Standard analysis	Analysis with IRT model
σ_E^2	0.055±0.004	0.855±0.137
σ_A^2	0.024±0.004	0.553±0.137
h^2	0.30±0.04	0.39±0.09

Van den Berg, S.M., Fikse, F., Arvelius, P., Glas, C.A.W., & Strandberg E. (2010). [Integrating phenotypic measurement models with animal models](#). *Proceedings of the 9th World Congress on Genetics applied to Livestock Production*, Leipzig, Germany, August 1-6, 2010.

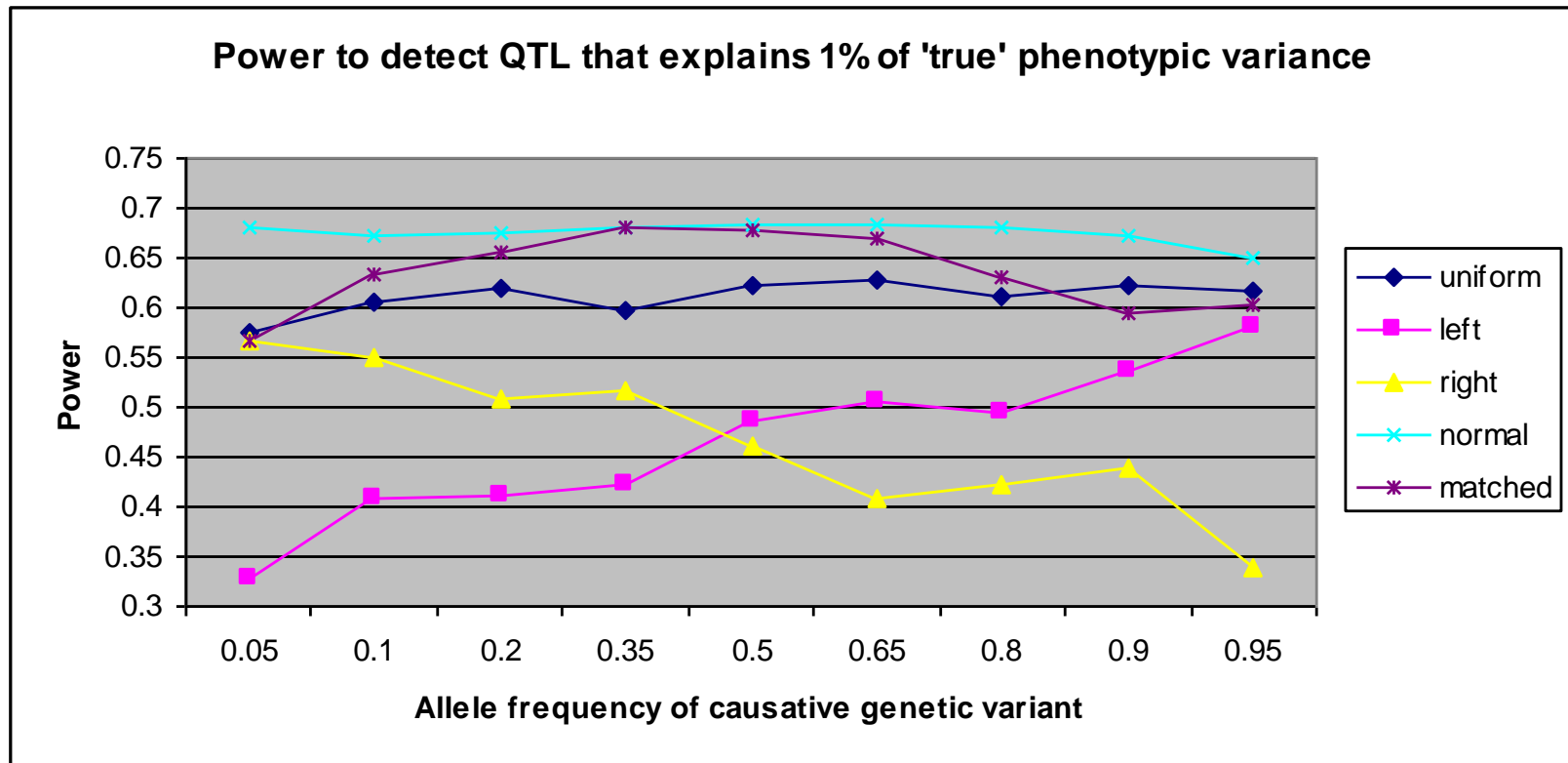
IRT and breeding

- Selecting “top dogs” for breeding
 - Use test items that are only passed by very few dogs (positive beta parameter values), you can omit non-informative items
- Selecting low scoring dogs (e.g., docile dogs): use items that many dogs pass



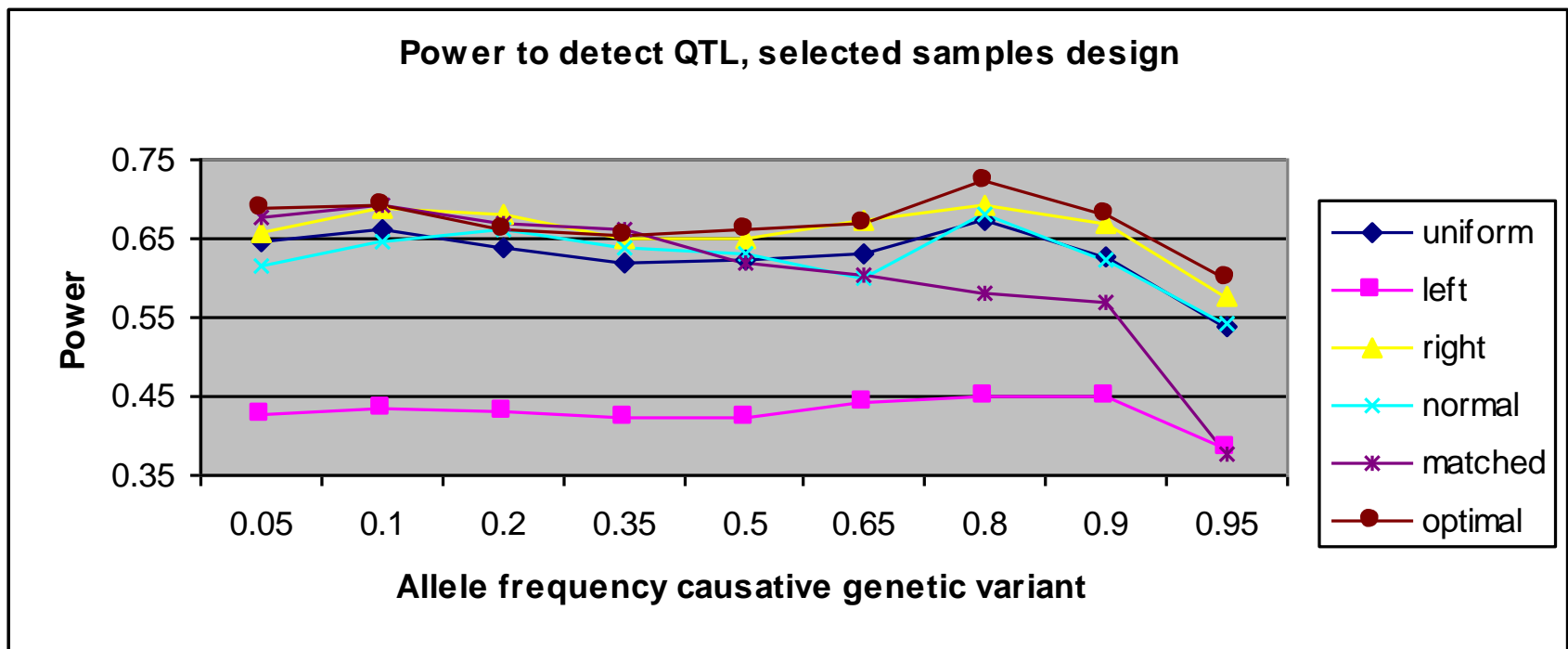
IRT and genetics

Van den Berg & Service, in preparation.



IRT and genetics

Van den Berg & Service, in preparation.



Increasing sample size with IRT

Table 1. Samples, number of subjects, files, Tests and links between the different tests of the phenotype harmonization project De Moor, Van den Berg, et al. (2011)

#	Sample name	N	File	NEO	TC I	EPQ	JEP O	IPIP	ABV	Nyra	MPQ
1	Munich	476	Munich-Germany.NEO-PI-R.TCI.MMPI2.07042011.sav	NEO-PI-R	TCI						
2	Italy	800	CILENTO.NEO-PI-R.08042011.sav	NEO-PI-R							
3	Young Finns	2058	YoungFinns.NEO.04042011.sav	NEO-FFI							
4	LBC1936 (GB)	1052	LBC1936.IPIP.NEO.08042011.sav	NEO-FFI				IPIP			
5	NTR	30598	NTR.ABV.NEO-FFI.08042011.sav	NEO-FFI					ABV		
6	Croatia VIS	918	CROATIA_VIS.EPQ-R.06042011.sav			EPQ-R					
7	Croatia KOR	810	CROATIA_KORCULA.EPQ-R.13042011.sav			EPQ-R					
8	Estonia	1731	EST_NEOPI3_data_180411.sav	NEO-PI-3							
9	NESDA (NL)	2981	NESDA.NEO-FFI.19042011.sav	NEO-FFI							
10	Finnish twins	30654	fintwin.epq.neo.27.04.2011.sav	NEO-PI-R		EPQ (1 item link!)					
11	BLSA (VS)	1917	BLSA.NEO-PI-R.6MAY11.sav	NEO-PI-R							
12	NBS (NL)	1832	NBS.EPQ-S.03052011.sav			EPQ-R					
13	QIMR	27065	QIMR_all.sav	NEO-FFI/PI-R	TCI	EPQ-R	JEPQ				
14	Swedish twins	36535	personality_from_q73_20110519.dta							Nyra	
15	ORCADES (GB)	602	ORCADES.EPQ-R.05052011.sav			EPQ-R					
16	USA Cogend	2712	COGEND.NEO-FFI.12052011.sav	NEO-FFI							
17	LBC1921 (GB)	478	LBC1921.IPIP.08042011.sav					IPIP			
18	Minnesota	2232	sibs.recode.all.sav								MPQ
19	Finland	1698	HBCS_TPQNEOPI.sav	NEO-PI-R	TCI						
20	ERF (NL)	2657	ERF_NEOFFI_IRT.sav	NEO-FFI							
21	SAGE (VS)	649	SAGE-COGA.TCI-TPQ.18052011.sav		TCI						

IRT, genetics and breeding

- **Conclusions:**
 - Separate measurement error from “environmental variance”
 - Optimize breeding programs and gene-finding studies
 - Increase sample size by including dogs that were tested in alternative ways

Further reading

- Van den Berg, S.M., Glas, C.A.W., Boomsma, D.I. (2007). [Variance decomposition using an IRT measurement model](#). *Behavior Genetics*, 37, 604-616.
- Van den Berg, S.M., Heuven, H.C.M., van den Berg, L., Duffy, D.L. & Serpell, J.A. (2010). [Evaluation of the C-BARQ as a measure of stranger-directed aggression in three common dog breeds](#). *Applied Animal Behaviour Science*, 124, 136-141.
- Van den Berg, S.M., Fikse, F., Arvelius, P., Glas, C.A.W., & Strandberg E. (2010). [Integrating phenotypic measurement models with animal models](#). *Proceedings of the 9th World Congress on Genetics applied to Livestock Production*, Leipzig, Germany, August 1-6, 2010.

Thanks

- University of Pennsylvania
 - James Serpell
 - Debby Duffy
- SLU:
 - Freddy Fikse
 - Erling Strandberg
- University of Utrecht:
 - Henri Heuven
- University of Twente:
 - Jasper Wouda