

The Rasch model in a large-scale CAT

The case of personalised assessments in Wales

Ben Smith

November 2018



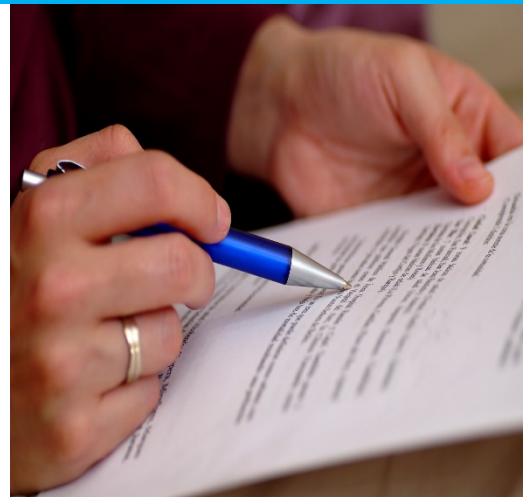
Partnership

We work in partnership with our clients. This is more than a cliché for us: we care about the services we provide and the impact they have on learners. Experience has shown us that the best impact our work can have is when it is undertaken alongside our clients so we make partnership a key feature of our project approach and management method.



Quality

We manage projects effectively and to the highest quality, freeing up experts to concentrate on their specialism, but ensuring that activities are managed to meet expectations. This means only making promises that we know we can keep, and remembering the promises we have made to make sure we deliver.



Expertise

We ensure our teams consist of genuine sector experts with understanding in breadth and depth of both the theory and the practical complex everyday challenges faced by education providers.



Development

We are committed to the improvement of our staff, both to promote the long-term development of our business and as an end in itself: we believe in the value of education for all.



Educationalists

We are educationalists with a strong commitment to improving teaching, learning and assessment, based on intellectual integrity, sound evidence and innovative approaches.

Welsh National Tests (WNTs)



Map data: GeoBasis-DE/BKG, Google

















- Years 2-9 (ages 6-14)
- Each year, tested in:
 - Reading
 - English
 - Welsh
 - Numeracy
 - Procedural
 - Reasoning
- England = national testing very summative
 - National Curriculum Tests at end of primary school (Year 6) and in Year 2
- Wales = more formative - hence annual testing. Until now, traditional paper tests.


Personalised assessments

- Over the next few academic years, WNTs are to be replaced by “personalised assessments” – computer adaptive tests (CATs)
- AlphaPlus is leading a consortium to deliver these new assessments
- The first personalised assessment, procedural numeracy, went live in December 2018
- Reading (English and Welsh) due to go live this year, numerical reasoning in 2020

Question: 1

Shoe sizes of all children in Year 3

Size 2	   
Size 2½	     
Size 3	  
Other	  

Key
 = 2 children

In Year 3, how many children are there altogether?

Question: 5

How many grams are in $\frac{1}{2}$ kilogram?

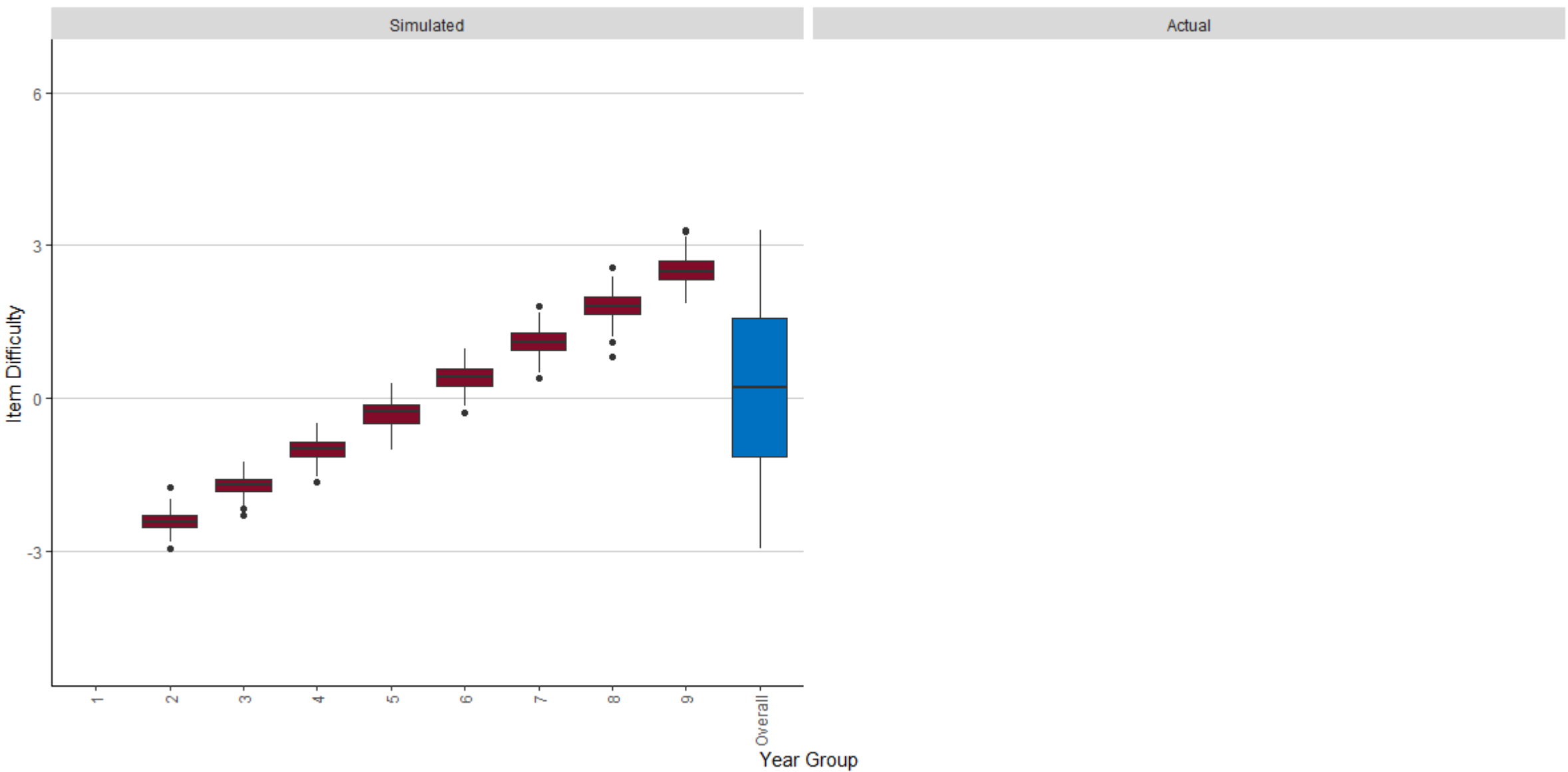
How is the Rasch model involved?

- In a CAT, we need to know how difficult each item is to be able to tailor each learner's test
- All items in the item bank pre-tested and calibrated using the Rasch model (JML estimation in Winsteps)
- During each CAT adaptive algorithm (developed by Angela Verschoor from Cito) fits the Rasch model to each learner's data **after each item they respond to**
 - i.e. learner ability is updated after every item, permitting adaptive item allocation
 - The item which provides most Fisher information is selected for delivery (within item exposure and content constraints)

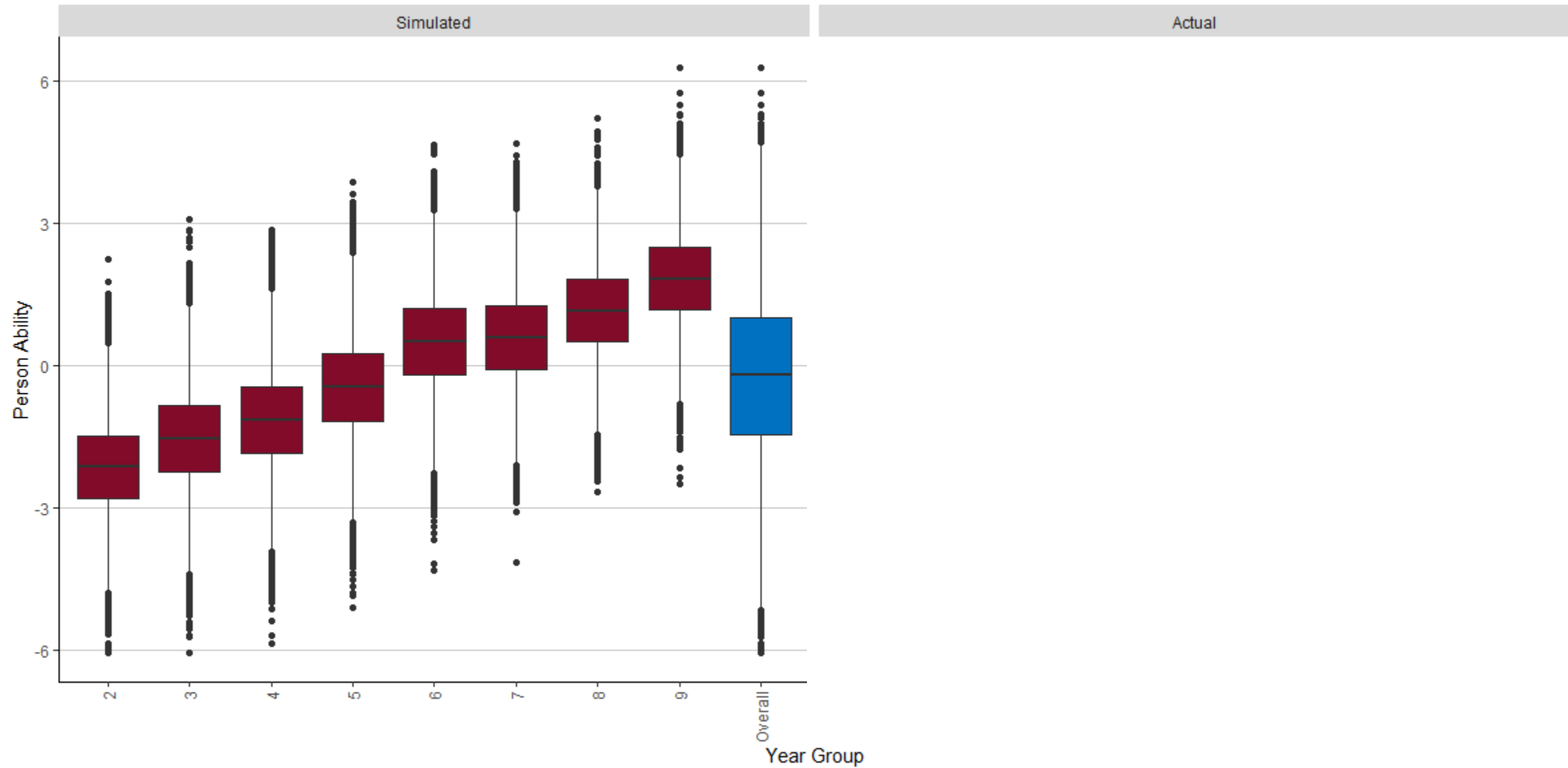
Complication

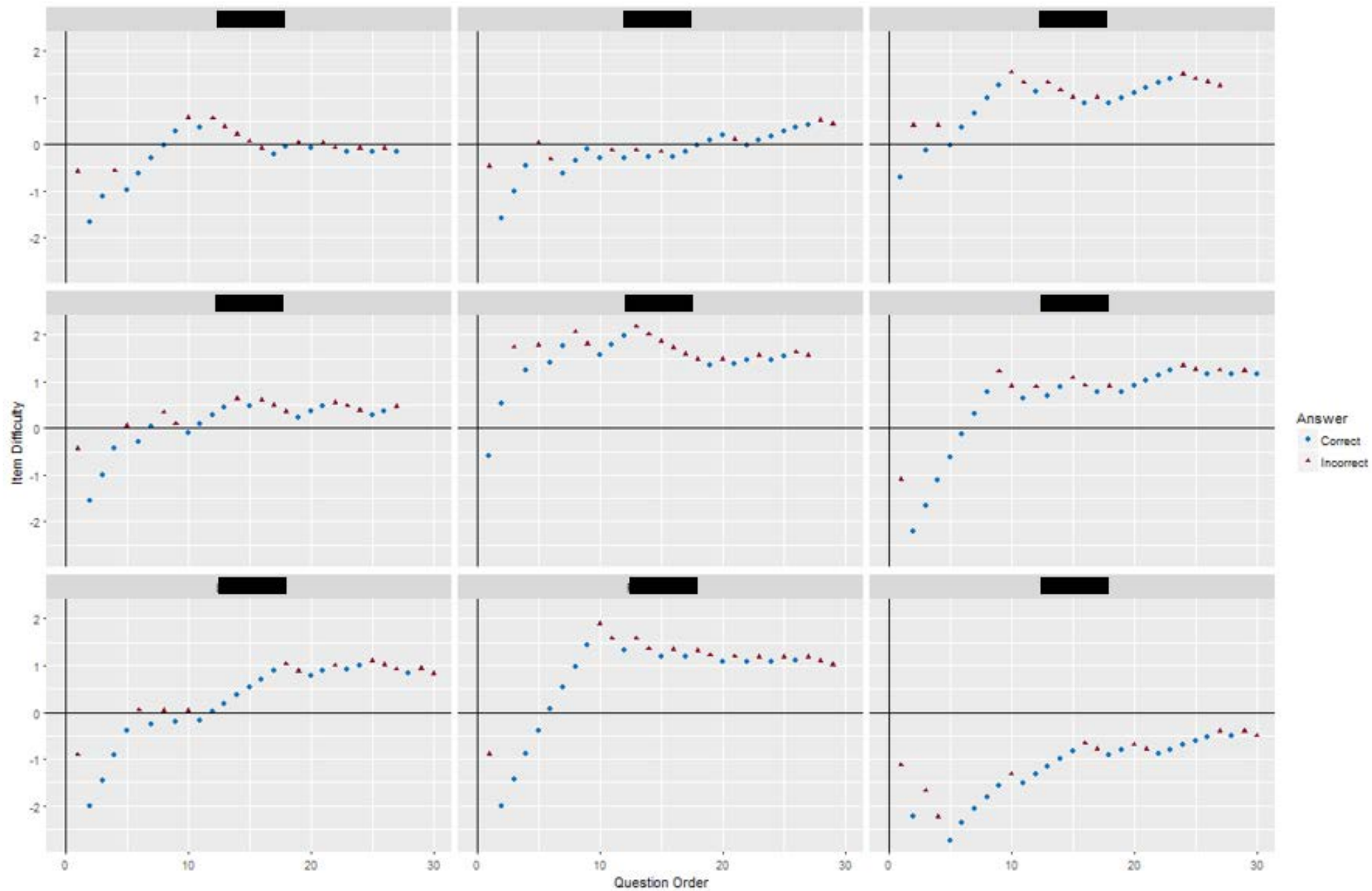
- CATs typically target a single age group
- I.e. we might have a Year 2 CAT with one set of items, a Year 3 CAT with another set of items, etc.
- The intent was to avoid floor and ceiling effects
- As such, the brief was for us to develop a single item bank for all year groups (for learners in Years 2-9)
- Learners can “roam” the bank
 - High ability learners are free to receive challenging questions that stretch them
 - Low ability learners are not demotivated, as they receive questions they can access, even if originally intended for lower year groups

Item difficulty



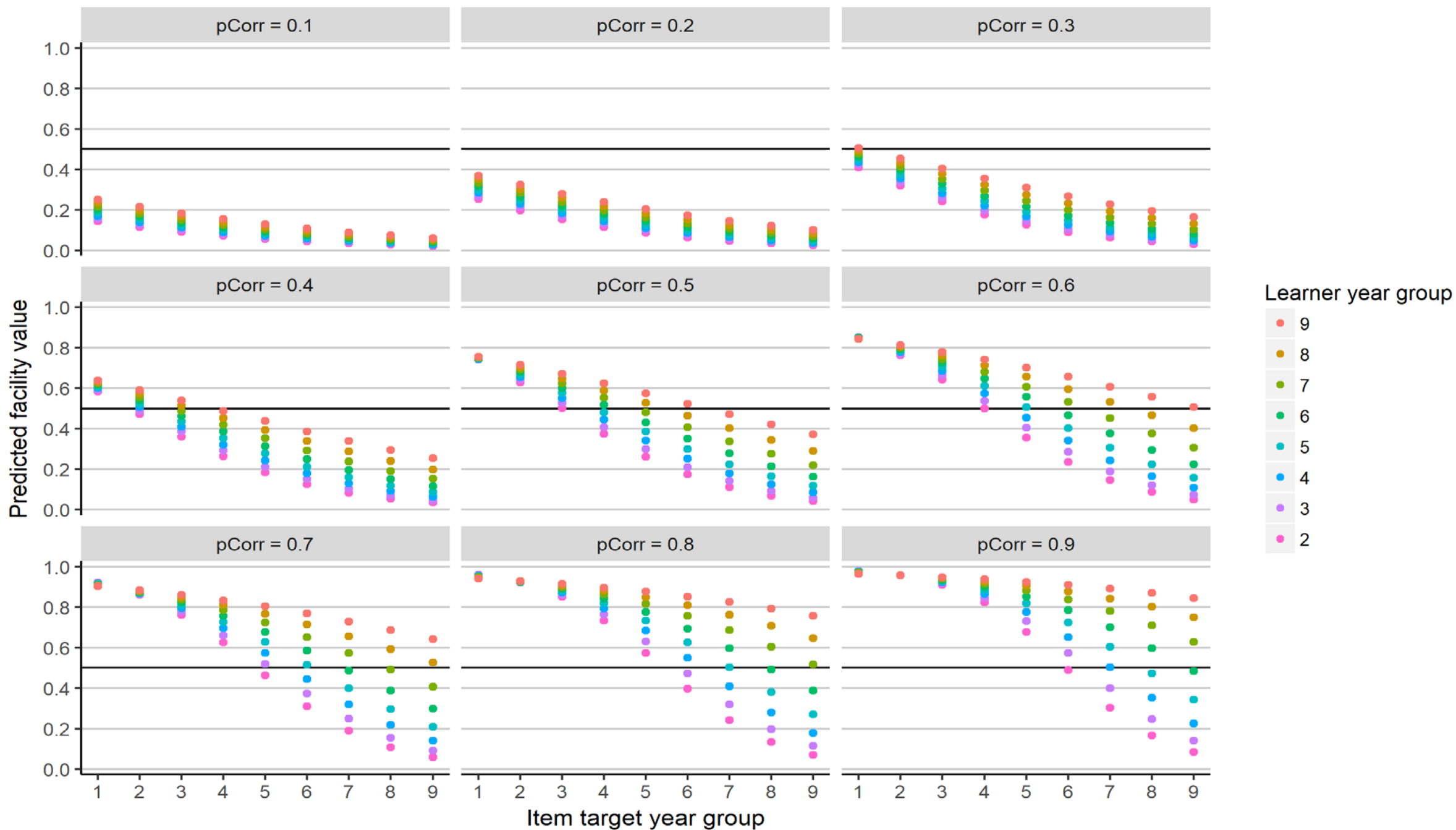
Learner ability





Multiple regression output

Variable	Coefficient	SE	P	Odds ratio	Probability
Intercept	-2.56	0.13	<0.01	-	-
Item target year	-0.19	0.04	<0.01	0.83	0.45
pCorr	0.84	0.03	<0.01	2.32	0.70
Learner Year	0.13	0.02	<0.01	1.13	0.53
Item target year * pCorr	-0.09	0.01	<0.01	0.92	0.48
Item target year * Learner Year	0.00	0.01	0.59	1.00	0.50
pCorr * Learner Year	-0.03	0.00	<0.01	0.97	0.49
Item target year * pCorr * Learner Year	0.01	0.00	<0.01	1.01	0.50



On-the-fly calibration

- Conventional wisdom is that a linear pre-test must be used to calibrate an item bank for CATs
 - This is because learners receive the “full range” of items, not merely items tailored to their ability level
 - Otherwise we see a “stretching” of the item difficulty parameters & inflated reliability estimates
- But substantial advantages to being able to calibrate “on-the-fly”, i.e. on adaptive test data
 - No need for linear pre-tests
 - Can renew item bank on a rolling basis as and when needed, rather than having to pre-test a set amount of items

Approaches in the literature

- Fink, Borr, Spoden & Frey (2018):
 - Initial (effectively) linear calibration
 - Continuous phase with 3 sections in the test:
 - Adaptive cluster
 - Calibration cluster
 - Linking cluster (common across forms) – 20% of items
 - Checks for item parameter drift

- Makransky & Glas (2014):
 - 3 strategies:
 - Two-phase – random, then adaptive once $>X$ average item administrations
 - Multi-phase – Increasing proportions of adaptive as more administrations
 - Continuous updating – Similar to above, but prop of adaptive not fixed per phase but proportional to # of items with $>X$ administrations

Approaches in the literature

- Verschoor, Bergen, Moser & Kleintjes (in press):
 - Various calibration methods with or without reference items
 - Different biases present in ability/difficulty estimates in each case:
 - Corrected JML & ELO – deflates severely
 - MML – inflates somewhat
 - Corrected JML with reference items – deflates somewhat, but only on the extremes
 - **MML with reference items – negligible bias**
- Only Verschoor et al addresses the question of calibration techniques purely on an adaptive dataset (i.e. no prior information from linear tests)
- Without reference items, MML best option
- Reference items are preferable – but how to pick them? How many are needed?



References:

Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, 60(3), 327-346.

Makransky, G., & Glas, C. A. (2014). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, 11(1), 1-20.

Verschoor, A., Berger, S. Moser, U. and Kleintjes, F. (in press). On-the-fly calibration in computerized adaptive testing, in Sluijter, C. and Veldkamp B.P. (eds), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Springer.

Thanks for listening!

AlphaPlus Consultancy Ltd

Unit 109

Albert Mill

50 Ellesmere Street

Manchester

M15 4JY

+44 (0)161 238 4928

info@alphaplusconsultancy.co.uk

www.alphaplusconsultancy.co.uk

