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.

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.

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.

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

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.

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.

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

Technical challenges arising from development of computer adaptive tests for Wales

Ben Smith & Angela Verschoor

November 2018
Welsh National Tests (WNTs)

- 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. Currently traditional paper tests.

Map data: GeoBasis-DE/BKG, Google
Personalised assessments

• Over the next three 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, is due to go live shortly

• Reading (English and Welsh) due to go live next year, numerical reasoning in 2020
CATs (Computer Adaptive Tests)

- If we have a vast bank of items (hundreds to thousands) and know all their difficulty values (via pre-testing and application of IRT)...
- ...we can design a test that adapts to the ability level of the learner.
  - If you get a question wrong, you get an easier question
  - If you get a question right, you get a harder question

- A CAT algorithm does this by estimating a learner’s ability (using IRT) after every single question they answer
  - It then “picks” the item from the item bank with the closest difficulty to their current ability estimate (if we’re aiming for them to get 50% of the items right)

- The CAT is stopped once a learner’s SEM has dropped below a set threshold
Advantages of CATs

• The test is tailored
  • High ability learners don’t get bored after racing through an easy test
  • Low ability learners don’t get disheartened by being unable to answer most items

• You can sit the test whenever you are ready
  • Due to having thousands of items in the bank, security is much less of a concern
  • In a very high stakes situation there probably would be a time window imposed
  • Controlling item exposure can address security concerns, however

• The test can generally be much shorter (here, potentially 75% the length of a normal test)
  • Because of the tailoring – we maximise information about the learner with fewer items
Stages of CAT development

1. Build item bank (write items)
   • But how do you know how many items to write?
   • Of which item type/for which curriculum area?

2. Pre-test & calibrate items
   • How do you go about pre-testing thousands of items?

3. Trial CAT
   • How do you ensure the algorithm functions in the way you want?
     • I.e. item exposure, test length, other more esoteric behaviours...
   • How do you check the initial calibrations have not changed?
Simulations

• Before item writers go to the effort of producing thousands of questions, we need to check what we are planning will work = simulations

• 3 main components:
  • Simulated item bank (that mirrors the characteristics of the actual bank)
  • Simulated learners (with abilities similar to those we expect in the actual cohort)
  • CAT parameters (that are those we plan to use in the live CATs)
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 from age 6-14)
- 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
CAT parameters

• First item(s):
  • A random item
  • From a defined difficulty band (straightforward for that year group)

• Target success rate:
  • 50%

• Curriculum controls
  • No more than x% or less than y% of a test can be from any one content type
  • Values of x and y depend on how many content areas there are

• Test end conditions
  • Minimum & maximum test length (items)
  • Stopping SEM

• Item exposure control
Why control exposure?
CAT example
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE</th>
<th>P value</th>
<th>Odds ratio</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.56</td>
<td>0.13</td>
<td>&lt;0.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Item target year</td>
<td>-0.19</td>
<td>0.04</td>
<td>&lt;0.01</td>
<td>0.83</td>
<td>0.45</td>
</tr>
<tr>
<td>pCorr</td>
<td>0.84</td>
<td>0.03</td>
<td>&lt;0.01</td>
<td>2.32</td>
<td>0.70</td>
</tr>
<tr>
<td>Learner Year</td>
<td>0.13</td>
<td>0.02</td>
<td>&lt;0.01</td>
<td>1.13</td>
<td>0.53</td>
</tr>
<tr>
<td>Item target year * pCorr</td>
<td>-0.09</td>
<td>0.01</td>
<td>&lt;0.01</td>
<td>0.92</td>
<td>0.48</td>
</tr>
<tr>
<td>Item target year * Learner Year</td>
<td>0.00</td>
<td>0.01</td>
<td>0.59</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>pCorr * Learner Year</td>
<td>-0.03</td>
<td>0.00</td>
<td>&lt;0.01</td>
<td>0.97</td>
<td>0.49</td>
</tr>
<tr>
<td>Item target year * pCorr * Learner Year</td>
<td>0.01</td>
<td>0.00</td>
<td>&lt;0.01</td>
<td>1.01</td>
<td>0.50</td>
</tr>
</tbody>
</table>

**Learner year group**
- 9
- 8
- 7
- 6
- 5
- 4
- 3
- 2
Moving forwards

- Concurrent calibration
  - If new items need to be added to the bank, can we avoid pre-testing in a linear fashion
- Updating parameters
  - Checking for item parameter drift
- First live standardisation due in Summer 2019
Thanks for listening!