BACKGROUND

The NAFS Project – University of Gothenburg
- Production of mandatory national high-stakes tests for EFL in Sweden (N~500,000/year) assessing reception and oral & written production and interaction.
- Production of assessment support materials for SFL in Sweden (optional use): French, German and Spanish.
- Core values: an action oriented approach to languages; national language syllabuses textually aligned with CEFR levels (so far, no empirical alignment); validity and reliability; broad construct coverage; accessibility, equity and fairness.

The Swedish Context
- Decentralised school system (however, national curricula and subject syllabuses).
- National tests as grading support for teachers and schools.
- Test development delegated by the National Agency for Education to universities; foreign languages at the University of Gothenburg.
- Increasingly digital society; strong public demand for ICT and digital tools; growing political pressure to produce digital national tests.

CB PILOT TESTS

CB Pre-pilot Tests: prior to the large-scale pilot tests, several pre-pilots were carried out.
- Demo Tasks: all students and teachers had access to CB demo tasks, exemplifying platform functions and various item formats.
- Anchoring: previous course grades reported for reference; anchor items on paper included in all CB pilot tests.
- Instruments: a web-based test platform within a locked browser; a variety of devices used by students (laptops, tablets, etc.); questionnaire for all test-takers including Likert-scales and written comments; all year 9 tasks in both CB and PB large-scale pilot tests.

Tested Skills
- Reading comprehension: shorter and longer texts, selected and constructed response; all input material and items delivered in the platform.
- Written production: access to a standardised set of functions for text editing and a word count; no spell- or grammar-check functions available.

CB Large-scale Pilot Tests: autumn 2017–spring 2018; sample of 6,010 students of EFL, ages 12–17, randomly selected (cluster sampling).

PRELIMINARY FINDINGS

Large-scale Pilot Tests

<table>
<thead>
<tr>
<th>Year</th>
<th>CEFR</th>
<th>PBT</th>
<th>CBT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample</td>
<td>Completed</td>
<td>Sample</td>
</tr>
<tr>
<td>6</td>
<td>~A2.1</td>
<td>933</td>
<td>771</td>
</tr>
<tr>
<td>9</td>
<td>~B1.1</td>
<td>1139</td>
<td>896</td>
</tr>
<tr>
<td>10</td>
<td>~B1.2</td>
<td>1166</td>
<td>919</td>
</tr>
<tr>
<td>11</td>
<td>~B2.1</td>
<td>1657</td>
<td>1346</td>
</tr>
<tr>
<td></td>
<td>4895</td>
<td>3932</td>
<td>6010</td>
</tr>
</tbody>
</table>

A Biased Sample?
- Missing
  - PBT: ~20%
  - CBT: ~50%, mainly caused by difficulties with locked browser function and other technical issues.
- Anchor item, results for year 9 (completed)
  - PBT: .69
  - CBT: .77

Positive Test-taker Feedback from CBT
- Reading comprehension: appreciated by most students; few comments on delivery mode.
- Written production: comments on observed benefits relating to speed, output, word count function, comfort, appearance, organisation, structure, composition, and proofreading.
  - I liked that I was able to edit my text in such an easy way. If I did the test by hand, I would not even bothered.
  - My handwriting is the worse so writing on the computer was gift sent from God.

Negative Test-taker Feedback from CBT
- Reading comprehension: comments related to interface problems (scrolling) and physical effects (eye strain and headache).
  - I think reading on a computer is hard and I can’t really focus on the content.
- Written production: comments concerning tablets without external keyboard.

CBT: BENEFITS AND CHALLENGES

A number of important issues and questions related to the core values of the NAFS Project have been noted:
- Authenticity, motivation and level of computer knowledge: possible bias towards schools/students with better access to and familiarity with devices; how to interpret current and minimize future, self-selected samples?
- Technical possibilities and limitations: administration process; hardware/software related issues; how to best design future CBT and CB answer keys and assessment material.
- Test developing process: what is possible and what is valid? Continuity?