What is the impact of diagnostic language tests?

J Charles Alderson, Lancaster University
Ari Huhta, Lea Nieminen, Riikka Ullakonoja, Eeva-Leena Haapakangas
University of Jyväskylä
Outline

• Diagnostic testing
• Stakes
• Wasback
• Washback from certain diagnostic tests or studies
• What could diagnostic feedback be like?
• Challenges
• We ask a lot of questions ...
Diagnostic testing

• Very under-developed and under-theorised in language testing and teaching
• Focus on learners’ strengths and weaknesses; on their prediction, even explanation
• Diagnosis requires a better understanding of the nature of strengths and weaknesses in particular language skills
• There are very few diagnostic SFL tests
Diagnostic tests

NOT Proficiency
NOT Achievement
NOT Placement
NOT Aptitude

BUT all the above *could* yield useful diagnostic information

HOWEVER, better is diagnosis by design

Diagnostic vs formative testing / assessment?
High-stakes tests

• Tests whose results are seen – rightly or wrongly – by students, teachers, administrators, parents or the general public, as being used to make important decisions that immediately and directly affect them. (Madaus, 1988)

• Are diagnostic tests (seen as) high-stakes? In what circumstances?
Washback

• Relates to the effects of tests on classroom practices – particularly teaching and learning (impact = a more general term)

• Can be positive or negative, to the extent that it either promotes or impedes the accomplishment of educational goals of learners and/or programme personnel.

• Should diagnostic tests have an effect on teaching and learning?
Negative Washback

• Mismatch between the stated goals of instruction and the focus of assessment
• May lead to the abandonment of instructional goals in favour of test preparation
• Forces teachers to do things they would not normally do
• Would diagnostic tests do this?
Positive Washback

• If a test has positive washback, ‘there is no difference between teaching the curriculum and teaching to the test’. (Weigle & Jensen, 1997, p. 205)

Do diagnostic tests lead to teaching to the test?
Positive views

Tests can be a powerful, low-cost means of influencing the quality of what teachers teach and what learners learn at school. (Heyneman & Ransom, 1992)

Do diagnostic tests influence the quality of what teachers teach and learners learn?
Areas of possible washback

Curriculum
  contents of curriculum, timetabling

Teaching materials
  choice of textbooks, use of past papers, teacher-made materials

Teaching methods
  choice of methods, teaching of test-taking skills

Attitudes and feelings, motivation
  of learners and teachers

Learning
  Do test results improve?
  Does learning improve?

Which of these do diagnostic tests affect?
Factors influencing washback

- Teacher beliefs
- Teacher attitudes
- Teacher training
- Learner beliefs, attitudes
- Resources
- The school
- Cultural factors

How do diagnostic tests relate to these factors, and vice versa?
Diagnostic testing (repeated)

• Very under-developed
• Focus on learners’ strengths and weaknesses;
• Diagnosis requires understanding of the nature of strengths and weaknesses in particular language skills
• There are very few diagnostic SFL tests
• Rare examples of diagnosis in action and in research into theory: DIALANG and DIALUKI
• But also DELTA, CohMetrix, Automated Writing Evaluation, ...
DIALANG

Diagnosis in action
Diagnosis by design
What is DIALANG?

• Computer-based diagnostic language testing system introduced in 2001 / 2004
• 14 European languages
• Delivers tests across the Internet
• Supports language learners
• Institutional or private use, free of charge
• Still widely used throughout Europe and beyond, 9 (or 12) years after launch
DIALANG feedback

Extensive feedback which learners can choose or ignore
Overall skill test performance + sub-skill + items
Match between self assessment and test result
Reasons for possible mismatch
Advice
DIALANG Experimental Items illustrate even further
types of feedback, e.g., on the number of attempts
= Task, Process and Self-regulation level feedback (see
Hattie and Timperley 2007)
Washback of DIALANG?

Little systematic research; Huhta’s (2010) survey of 550 test takers’ perceived usefulness:

• overall test result most useful, then item level feedback (reported learning about errors)
• however, also FB targeting self-regulation (e.g. on self-assessment) was useful for many

Awareness raising for both teachers and learners: language proficiency, CEFR, ...

Impact of DIALANG on an increasing interest in diagnostic testing
Other diagnostic (or potentially diagnostic) tests / assessment systems presented at EALTA 2013, such as:

- DELTA (Wong & Raquel)
- CohMetrix (Graesser)
- Automated Writing Evaluation (Link & Dursun)

... these presentations also included discussion, hypotheses and/or results of their impact / washback
DIALUKI

Understanding Diagnosis
Researching Diagnosis
• Diagnosing Reading and Writing in a Second or Foreign Language
• Research project 2010-2013: work in progress
  www.jyu.fi/dialuki
• Funded by the Academy of Finland, the University of Jyväskylä and the UK Economic and Social Research Council (ESRC)
• Cooperation between language testers, other applied linguists and psychologists (L1 reading)
The main research questions

– To what extent can different L1 and L2 linguistic, psycholinguistic, motivation and background measures predict SFL R/W performance, especially difficulties?

– How does SFL proficiency in R/W develop in psycholinguistic and linguistic terms?

– Which features or combinations of features characterise different CEFR proficiency levels?
# Three major studies

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A cross-sectional study with 850 students</td>
<td>Longitudinal study with over 200 students</td>
<td>Intervention study</td>
</tr>
<tr>
<td>Data collection: 2010-11</td>
<td>Data collection 2010-13</td>
<td>Data collection 2012-13</td>
</tr>
<tr>
<td>Exploring the value of a range of L1 &amp; L2 measures in predicting L2 reading &amp; writing</td>
<td>The development of literacy skills, and the relationship of this development to the diagnostic measures.</td>
<td>The effects of training on SFL reading and writing</td>
</tr>
</tbody>
</table>
Informants in DIALUKI Study 1

Finnish-speaking learners of English as FL

- Primary school 4\textsuperscript{th} grade
- Lower secondary school, 8\textsuperscript{th} grade
- Gymnasium 2\textsuperscript{nd} year students

- 200+ students in each group

Russian-speaking learners of Finnish as SL

- primary school (3-6th grade; N= 186)
- lower secondary school (7-9th grade; N= 78)
Possible areas of impact of the DIALUKI study

For teachers involved in the study:

• Hightened awareness about ’diagnostic assessment / testing’ and about reading in L2
  — variable approaches to reading, not always specifically taught

• More and earlier samples for teachers of learners’ reading and writing (English)

• External view to learners’ skills; support to grading and (formative) knowledge about students

For learners: Feeback on performance, awareness of proficiency levels
Examples of feedback to learners and teachers from DIALUKI studies

• Cross-sectional: after each data collection phase

• Longitudinal (grade 4-6): after the final data collection round
  – Reading & writing in English at 4-5 points in time, vocabulary at 2 points
Longitudinal FB on reading
- sample FB for one student (translation)

<table>
<thead>
<tr>
<th></th>
<th>Grade 4</th>
<th>Grade 5 autumn</th>
<th>Grade 5 spring</th>
<th>Grade 6 autumn</th>
<th>Grade 6 spring</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>READING IN ENGLISH</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Sofia and the cat</td>
<td>3 / 5 points</td>
<td>5 / 5 points</td>
<td>5 / 5 points</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>• Missing words - gap-fill text</td>
<td>5 / 7 points</td>
<td>7 / 7 points</td>
<td>7 / 7 points</td>
<td>-</td>
<td>7 / 7 points</td>
</tr>
<tr>
<td>• Letter to parents</td>
<td>-</td>
<td>6 / 9 points</td>
<td>8 / 9 points</td>
<td>-</td>
<td>7 / 9 points</td>
</tr>
<tr>
<td>• Healthy eating</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>• Story about sailing</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4 / 5 points</td>
</tr>
</tbody>
</table>
Longitudinal FB on writing

<table>
<thead>
<tr>
<th></th>
<th>Grade 4</th>
<th>Grade 5 autumn</th>
<th>Grade 5 spring</th>
<th>Grade 6 autumn</th>
<th>Grade 6 spring</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WRITING IN ENGLISH</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A message to a friend abroad</td>
<td>A2 (CEFR)</td>
<td>A2.2 (NCC)</td>
<td>A2.1 (NCC)</td>
<td>-</td>
<td>B1.1 (NCC)</td>
</tr>
<tr>
<td>in which the pupil tells</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what he/she likes and why</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A story about a memorable</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A2.1 (NCC)</td>
<td>B1.1 (NCC)</td>
</tr>
<tr>
<td>event</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

What are the intended (and) positive consequences of diagnostic testing?

What might be the possible unintended negative consequences of diagnostic testing?

The fact is that so far we have little research into the washback or impact of diagnostic tests.

Empirical research is urgently needed: How might such research be designed and conducted?
Conclusions

• Washback of diagnostic tests takes place via feedback; if FB is not relevant or understandable it won’t impact learning and thus does not have the intended washback.

• Diagnostic tests aim at positive washback, i.e., they are not ‘neutral’ in this respect(?).

→ Diagnostic tests that don’t have positive washback fail in their main purpose and, therefore, we can question their validity(?)
Some characteristics of feedback from diagnostic tests that are likely to increase tests’ positive washback 1

• Based on solid understanding of relevant constructs
• Immediate
• Detailed enough
• Relates to clear, achievable goals
Some characteristics of feedback from diagnostic tests that are likely to increase tests’ positive washback 2

• Computerization in general and automated analysis of learners’ responses / language in particular would enhance it (cf. Art Graesser’s plenary: conversational agents, CohMetrix type of analyses, speech recognition)

• Targets different FB levels (task, process, self-regulation); including time

• Understandable: use of L1 if possible, teacher interpretation

• Others?
Challenges for diagnosis, feedback and washback 1 → input for research agenda?

• Ensuring that the entire chain from diagnosis to feedback to action / intervention is not broken
• Understanding language proficiency & learning in enough detail (constructs)
  → useful / meaningful diagnosis
• Measurement / sampling (how much evidence is enough, how many items per aspect / point?)
Challenges for diagnosis, feedback and washback 2 ... 

• Designing / deciding on what FB to give, how & when
  – Intelligibility
  – Learners’ motivation, goal orientation, self-regulation & meta-linguistic knowledge
  – To what extend can learners interpret and act on FB without assistance from a teacher? Depends on type of FB?
  – Too much feedback? (e.g. from automated analyses)
  – Overemphasis on task feedback (correctness / product)?

• Action, intervention: self-study vs course requirement
  \[ \rightarrow \text{impact / washback – is it what we hope it to be?} \]
Do diagnostic tests have impact?

What do you think?

Positive or negative?
On teaching?
On learning?
On content?
On method?
On rate and sequence of learning?
On degree and depth of learning?
On attitudes and motivation?
On all teachers and learners?
On some teachers and learners?
On applied linguistics?
Thank you for your attention!