Descriptors for reading: Which one(s) make the difference?

Dr. Monique Reichert

Luxembourg Centre for Educational Testing
University of Luxembourg

12th EALTA Conference, University of Copenhagen
Policy and Practice in Language Testing and Assessment
26th – 31st May 2015
1. INTRODUCTION: Setting the context

The Épreuves Standardisées (ÉpStan)

National school monitoring program (regulated by the national law from 6th Feb. 2009)

Aim:

- Provide an external, longitudinal and standardized evaluation of the competency level reached by the pupils of the Luxembourg educational system in key competency domains.

Link with educational standards (*Socles de compétences*)
1. INTRODUCTION: Setting the context

The Épreuves Standardisées (ÉpStan) (cntd.)

<table>
<thead>
<tr>
<th>KIGA 1</th>
<th>KIGA 2</th>
<th>1st grade</th>
<th>2nd grad.</th>
<th>3rd grad.</th>
<th>4th grad.</th>
<th>5th grad.</th>
<th>6th grad.</th>
<th>7th grad.</th>
<th>8th grad.</th>
<th>9th grad.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYCLE 1</td>
<td>CYCLE 2</td>
<td>CYCLE 3</td>
<td>CYCLE 4</td>
<td>7e</td>
<td>8e / 6e</td>
<td>9e / 5e</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. INTRODUCTION: Setting the context
The Épreuves Standardisées (ÉpStan) (cntd.)

Item development

• Multidisciplinary working groups
  – Psychologists / Psychometrieciens (UL)
  – Teachers (EF)
  – Authors of reference documents (MENFP)
  – Didactics experts (UL)
  – Collaborators (MENFP)
  – External partners (Illustrators, Designers, Sound ingenieurs, Speakers)
1. INTRODUCTION: Setting the context

The Épreuves Standardisées (ÉpStan) (cntd.)

Item development

Reading comprehension in secondary schools

French

What is being measured?
1. INTRODUCTION: Setting the context

The Épreuves Standardisées: Item development

Reading comprehension in secondary schools
1. INTRODUCTION: Setting the context
The Épreuves Standardisées: Item development
Reading comprehension in secondary schools
1. INTRODUCTION: Setting the context
The Épreuves Standardisées: Item development

Reading comprehension:
What is being measured?
1. INTRODUCTION: Setting the context

The Épreuves Standardisées: Item development

Reading comprehension:
What is being measured?

1. INTRODUCTION: Setting the context

The Épreuves Standardisées: Item development

Reading comprehension:
What is being measured?

Problem 1:

Descriptors are not sufficiently precise in order to guide teachers involved in item development specify the construct measured by their test, and to help understand the differences between adjacent levels

→ Difficult to clearly specify the difficulty of the items, and to understand what makes items more or less difficult
1. INTRODUCTION: Setting the context

The *Épreuves Standardisées*: Item development

**Reading comprehension:**

*What is being measured?*

**Problem 2:**

Empirical evidence to justify the levels – particularly with regard to a theoretical model of reading – is not very strong

→ Disadvantage with regard to what can be provided as a feedback to (language) teachers
1. INTRODUCTION: Setting the context

The Épreuves Standardisées: Item development

Reading comprehension: What is being measured?

Problem 3:

Tests are administered to **highly heterogeneous groups of test candidates**

(High share of immigrants: 44.5%; Largest group of immigrants: Portuguese: 37% of all immigrants)

Students tested in **German and French**
1. INTRODUCTION: Setting the context
The Épreuves Standardisées: Item development

Reading comprehension:
What is being measured?

Problem 3:
→

Construct validity of the reading tests may strongly differ depending on the target language, the items used, or the language background of the test candidates.

More crucially: Even if we know that the construct of the test differs from one test or group to another, it remains difficult to determine what makes the test / items function differently.
1. INTRODUCTION: Setting the context

The Épreuves Standardisées: Item development

Reading comprehension:

What is being measured?

Better knowledge of the attributes that affect the difficulty of reading comprehension items would lead to a more theory-driven test construction and item analysis, and support a more systematic comparison of tests (e.g., French and German test) and study results.
2. THEORETICAL BACKGROUND

2.1. Factors known to affect reading item difficulty

→ Review of scientific literature (e.g., Buck, Tatsuoka & Kostin, 1997; Freedle & Kostin, 1999)

→ Consideration of existing ‘guiding’ tools, such as the Dutch Grid (Alderson, Figueras, Kuijper, et al., 2006), or the CEFESTIM grid (Tardieu, Hildén, Lehman & Reichert, 2011)

→ A selection of different descriptors was taken, discussed and used during item development meetings, which led to a minimal number of descriptors guiding the item developers throughout the item development
2. THEORETICAL BACKGROUND

2.1. Factors known to affect reading item difficulty (i.a.)

- Text vocabulary (how frequent is the vocabulary used in the text?)
- Concreteness of the language used in the text
- Grammatical complexity of the text
- General text structure (e.g., does it follow a known, narrative structure) (cf. Levy-Hillerich et al., 2008)
- Position of necessary information in the text
- Familiarity of the students with the subject of the test
2. THEORETICAL BACKGROUND

2.1. Factors known to affect reading item difficulty

Test results should reflect as far as possible target construct, i.e., reading comprehension → Take into account theoretical models of reading / comprehension

2 theoretical models:

Kintsch’s Construction-Integration model (1998)

Khalifa & Weir (2009): Cognitive processing model for reading comprehension, notably distinction between different types of reading (Urquhart & Weir, 1998)
2. THEORETICAL BACKGROUND

2.2. Kintsch’s Construction-Integration model

(Kintsch, 1998; Kintsch & Rawson, 2005)

- Text-based comprehension (students have to find explicitly stated information in text) (text surface representation)

- Type 1 Inference, based on information from the text (paraphrased information)

- Type 2 Inference, based on information from the text (information paraphrased and distributed across more sentences)

- Type 3 Inference: automatically made inference (part of situation model: readers go beyond what is explicitly stated)

- Type 4 Inference: controlled retrieval (inference based on knowledge)

- Situation model (text has to be understood globally)
2. THEORETICAL BACKGROUND


-Careful reading versus Expeditious reading

- **Careful reading local**: processing at the sentence level, until the basic meaning of a proposition is established (e.g. resolution of lexical ambiguity, matching similar words)

- **Careful reading global**: reader is trying to identify the main idea(s) by establishing the macro-structure of a text: how the ideas in the whole text relate to each other and to the author’s purpose.
2. THEORETICAL BACKGROUND


Distinction between different types of reading:

- Careful reading versus Expeditious reading
  - **Expeditious reading**- **Scanning**: basically a visual recognition process
  - **Expeditious reading**- **Search reading**: Search for information is guided by predetermined topics, so the reader does not necessarily have to establish a macro-propositional structure for the whole of the text; search is for words *in the same semantic field* as the target information; reader is sampling the text.
2. THEORETICAL BACKGROUND


Distinction between different types of reading:

- Careful reading versus Expeditious reading
  - **Expeditious reading - Skimming**: reading is highly selective, with particular parts of the text omitted or given little attention; attempt to build macro-structure of the text with as few details as possible
3. Research objectives

Exploration of the following questions:

- Are there descriptors of reading comprehension that affect item difficulty (in the ÉpStan reading test) more strongly than others?
- How well do Urquhart and Weir’s reading types explain item difficulties as compared to Kintsch’s Construction-Integration model?
- Are there items for which the empirical difficulty can not sufficiently be predicted by the descriptors?
3. Research objectives

The responses to these questions aim at elaborating the existing descriptors and providing items that illustrate well different degrees of difficulties.
4. METHOD

4.1. Instruments: Items

Computer-based reading comprehension items used in the frame of the Épreuves Standardisées

For the current analyses:
Focus on the French secondary school reading comprehension items (9th grade), used in November 2013 (N=6027)

Among closed format item, only Multiple-Choice items were taken (42 items) for the current analyses
4. METHOD

4.2. Item Pool calibration

- Verification whether the item pool fits the Rasch model
- analysis of Differential Item Functioning (DIF), depending on
  - Score (median-split)
  - Gender
  - Migration background

DIF found for 8 items, resulting in a final item pool of 34 items
4. METHOD

4.3. Rating of the items, based on theoretical attributes

- theoretical attributes or descriptors formulated as questions, e.g.:
  - Kintsch’s *Textbased comprehension* and Kintsch’s Type 2 Inference (based on information from the text):
26. **Textbasiertes Verstehen:** Die Frage bezieht sich auf eine Information, *(mehrere Angaben, d.h. Kombinationen möglich)*

26.1. die **explizit** im Text enthalten ist:

- die relevanten Wörter innerhalb der Frage sind wortwörtlich im Text wiederzufinden
- die korrekte (vorgeschlagene) Antwortoption ist wortwörtlich im Text wiederzufinden
- die Wörter innerhalb der Frage und die Antwortoption sind wortwörtlich im Text wiederzufinden

26.2. die **in anderen Worten** (paraphrasiert) im Text wiedergegeben ist

- die relevanten Wörter innerhalb der Frage sind in anderen Worten im Text wiederzufinden
- die korrekte (vorgeschlagene) Antwortoption ist in anderen Worten im Text wiederzufinden
- sowohl die Wörter innerhalb der Frage als auch die Antwortoption sind in anderen Worten im Text wiederzufinden

26.3. **Nicht zutreffend:** Die Frage ist auf einer anderen Ebene als derjenigen des Textes gestellt *(z.B. Frage nach passendem Titel, der Intention des Autors, etc.)*
4. METHOD

4.3. Item ratings, based on the theoretical attributes

- Within item specification workshops: each item rated by between 2 and 6 experts (among whom at least 50% were French language teachers)

- Item attribute ratings represented within a weight matrix
## 4. METHOD

### 4.3. Item ratings, based on the theoretical attributes

Item attribute ratings represented within a weight matrix

<table>
<thead>
<tr>
<th>item code</th>
<th>long text</th>
<th>text characteristics</th>
<th>Textbased comprehension</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>low frequency words</td>
<td>high grammatical complexity</td>
</tr>
<tr>
<td>f90c511</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f90c531</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f90c541</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f90c551</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f90c561</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f952113</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>f952143</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>f952151</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>f95d121</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f95d131</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
4. METHOD

4.4. Statistical analyses
• Rasch-model (Rasch, 1960)
4. METHOD

4.4. Statistical analyses

• **Linear logistic test model (LLTM)** (Fischer, 1972) extension of the Rasch model with linear constraints on item parameters

  *Also see Purya Baghei & Klaus Kubinber (2015)*

With the LLTM, the fit of the weight matrix (the specified model) can be tested similar to testing the fit of the Rasch model:

• a **graphical model check** can be used to compare the estimated item parameters of the Rasch model to the parameters reproduced by the LLTM.

• Items **reducing the respective correlation deliver valuable information**, which can be used to adapt the weight matrix, or to help to identify problems in item construction.
4. METHOD

4.4. Statistical analyses

• LLTM with Q-Matrix as specified based on Kintsch’s C-I model;
• LLTM with Q-Matrix as specified based Khalifa & Weir’s Cognitive processing model
4. RESULTS: C-I Model (Kintsch)

Correlation Rasch-Model - LLTM: .8782
4. RESULTS: Cognitive processing model (Khalifa & Weir)

Correlation Rasch-Model – LLTM: .8793
4. RESULTS: C-I Model (Kintsch)

- Reading of continuous texts
- Vigilant reading required
- Text contains many low frequency words
- Necessary information is scattered across the text
- Text is characterized by frequent changes in perspectives
- Using background knowledge
- Situation model
- Make an inference that goes beyond the text
- Reading of literary texts
- Reading texts with a low readability index
- Reading long texts

Textbased comprehension: Necessary information is paraphrased
4. RESULTS: Cognitive Processing Model (Khalifa & Weir)

Careful reading global: overall text

Text is characterized by frequent changes in perspectives

Text contains many low frequency words

Vigilant reading required

Reading of continuous texts

Careful reading local

Search reading

Reading of literary texts

Reading texts with a low readability index

Using background knowledge

Careful reading global: across sentences

Search reading: Information scattered

Expeditious reading: Scanning

Reading long texts

← Difficult

← Easy
4. RESULTS: C-I Model (Khalifa & Weir)

Careful reading global: overall text
Text is characterized by frequent changes in perspectives
Text contains many low frequency words
Vigilant reading required
Reading of continuous texts
Careful reading local
Search reading
Reading of literary texts
Reading texts with a low readability index
Using background knowledge
Careful reading global: across sentences
Search reading: Information scattered
Expeditiously reading: Scanning
Reading long texts

Difficult

Easy
5. Summary and conclusions

• Good explanation of the difficulty of reading comprehension items by a restricted number of cognitive, textual and contextual descriptors.
• Further model adjustments will however be required in order to fully understand the test’s construct validity.
• Added value of combining attributes from both models will need to be explored.
• Additional analyses of items that are more difficult to evaluate by human raters will be necessary.
• Analyses provide a good support for a theory-driven item development.
THANK YOU FOR YOUR ATTENTION
Descriptors for reading: Which one(s) make the difference?

Acknowledgements

Students involved in the Luxembourg school monitoring program
Teachers involved in the item development

most notably:
Philipp Sonnleitner
2. The Luxembourg test setting (cont.)
Comparing reading performance in German and French for students with different mother languages
1. INTRODUCTION

The Épreuves Standardisées: Item development Reading comprehension: Classification of items

Item classification according to the framework

–1ᵉʳ preliminary classification, based on:
  • Text characteristics: length, word frequencies, familiarity, complexity of syntactical structures, etc.
  • Item characteristics: length of item stem and options, cognitive process
  • Text-item interaction: length of text/proposition to be read, number of propositions / information units to be considered or combined...

–Pretest: 2ᵉʳ preliminary classification, based on:
  • Empirical item parameters (difficulty, discrimination) (from pretest)
  • Pupils’ responses (option chosen)
2. The Luxembourg test setting

High share of immigrants: 44.5% (Statec, 2013)
Largest group of immigrants: Portuguese (37% of all immigrants)
Three official languages (Luxembourgish, German, French)
Multilingual school system; languages used:
• Luxembourgish
• German
• French
Further language(s) taught: English, other
1. INTRODUCTION: The Luxembourg test setting

High share of immigrants: 44.5% (Statec, 2013)

Largest group of immigrants: Portuguese (37% of all immigrants)

Three official languages (Luxembourgish, German, French)

Multilingual school system; languages used:

• Luxembourgish
• German
• French

Further language(s) taught: English, other
Median-Split RM (French)