Institut zur Qualitätsentwicklung im Bildungswesen

EALTA

Barcelona, June 2007

Validity of a Test of English as Second Language – The Test-Takers' Perspective

Hans Anand Pant & Miriam Vock

Team:
R. Green, C. Harsch, M. Leucht, D. Neumann, R. Oehler, S. Franke

Overview

- German National Educational Standards (NES) for English as Foreign Language
- Background of Test Development at the IQB
- Test-takers' Perspectives as a Source of Information during Test Development
- Associations between Test-takers' Perceptions and Test Results
- Conclusions

German National Educational Standards (NES) for English as a first foreign language (EFL)

- describe **core competencies** German students should have acquired at the end of grades 9 and 10 (**age 15 and 16**)
- will be **evaluated** in continuing and nationally representative assessments from 2009 on (conducted by IQB)
- focus on **communication skills** (e.g. reading comprehension, listening comprehension) and state cumulative learning processes
- describe expected proficiencies in terms of **can-do statements** derived from **levels of foreign language use** described in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001).

NES for communication skills – derived from CEFR levels of language use

Proficient User	C2	Can understand with ease virtually everything heard or read.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning.
Independent User	B2	Can understand the main ideas of complex texts on both concrete and abstract topics, including technical discussions in his/her field of specialisation.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.

Note. Table was simplified. From Council of Europe (2001, p. 24).

NES for reading comprehension – expected proficiencies

At the end of grade 9 (age 15) students are able to:

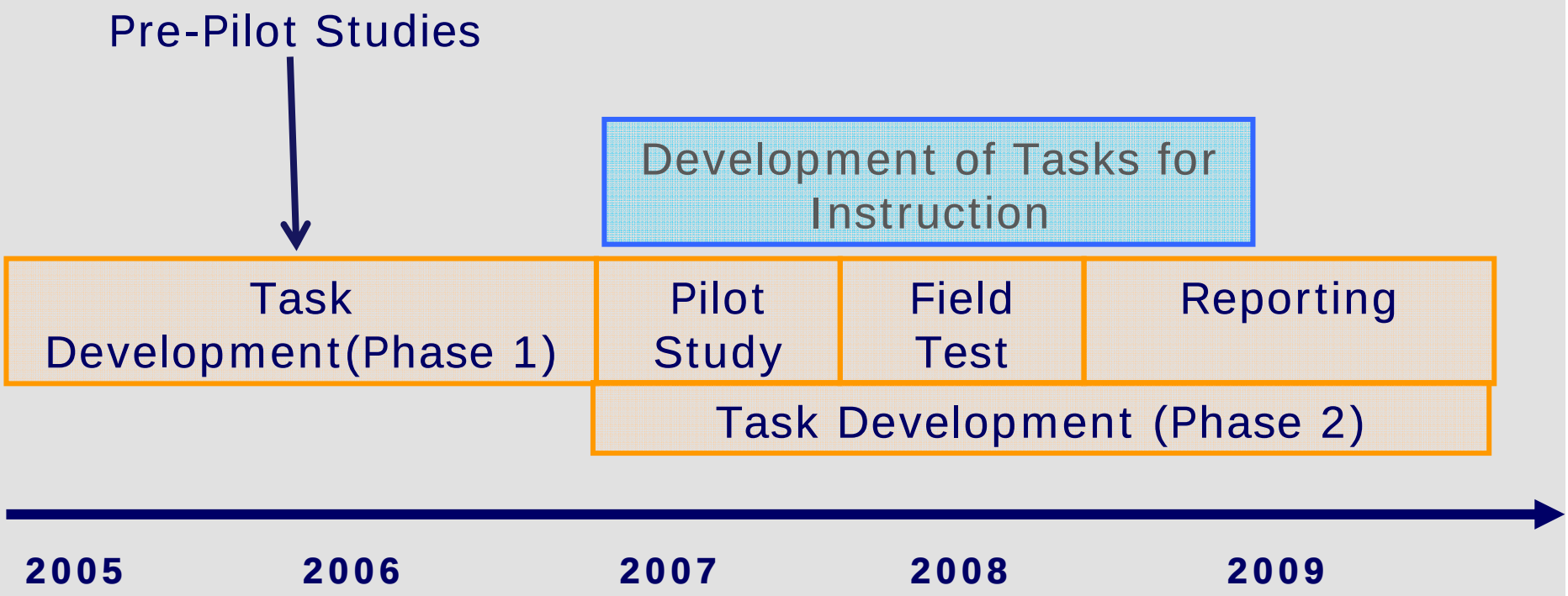
1. understand short, simple, personal letters and e-mails (A2),
2. find concrete, predictable information in simple everyday texts (e.g. adverts, program magazines) (A2),
3. find specific information in simple written materials (e.g. newspaper articles) or fiction in line with the level (A2).

At the end of grade 10 (age 16) students are able to:

1. read correspondence relating to their own sphere of interest and understand the main information (B2),
2. search longer texts for required information and gather information from several texts to solve a particular task (B1+),
3. recognize the main conclusions in clearly written, argumentative texts on familiar topics, for example in newspaper articles (B1/ B1+).

IQB Project Timeline

Task Development Process



Method: Sampling Design

Pre-Pilot Study	Booklet	# Tasks	# Items	N
Reading Comprehension	1	5	31	82
	2	5	27	81
	3	5	29	81
	4	5	35	80
	5	5	36	105
	6	5	31	108
	7	5	36	103
	8	5	36	100
	Total RC	**40**	**261**	**740**
Listening Comprehension	1	4	25	200
	2	4	25	86
	3	4	24	157
	4	4	22	122
	Total LC	**16**	**96**	**565**

Test development princples

- Authenticity of input material (non-didacticized, no text books, original layout, etc.)
- Diversity of test methods (e.g., MC, sequencing, multiple matching, gap filling, table completion, short answers, true-false-not given, ...)
- Diversity of text/input sources, text types, and topics

Reading Comprehension Stimuli

Text Sources	Text Types	Topics
• Internet	• Timetables	• Personal identification
• Leaflets	• Programs	• House and home
• CD/DVD covers	• Adverts	• Environment
• Youth magazines	• Menus	• Daily life
• Magazines	• Blurbs	• Free time
• Newspapers	• E-mails	• Entertainment (media, sports, music)
• Menus	• Simple instructions	• Travel
• Books	• Complex instructions on machines or procedures	• Relations with other people
• Adverts	• Manuals	• Health and body care
• Blurbs	• Newspaper/ magazine articles	• Education
• Manuals	• CD/DVD covers	• Shopping
• Signs	• (Personal) Letters	• Food and drink
• Brochures	• Articles	• Services (museums, libraries, hospitals)

Reading Comprehension Stimuli (continued)

Text Sources	Text Types	Topics
• Posters	• Postcards	• Places
• Reference books/ encyclopaedias	• Leaflets	• Weather
• Comics	• Text messages	• Work
• Reviews	• SMS/text messages	• Multicultural society
• Correspondences	• Lyrics	• Celebrities
	• Literature (extracts)	• Animals/animals and wildlife
	• (Short) Stories	• History
	• Interviews	• Festivals/customs
	• Reports	• Languages
	• Recipes	• Crime
	• Cartoons	• Global problems
	• Signs	• English speaking countries
	• Reviews	• Aspects of society
	• Folk stories	• Adventure (and challenges)
	• Memos	• Science and technology

Validity Issues during Pre-Pilot Studies

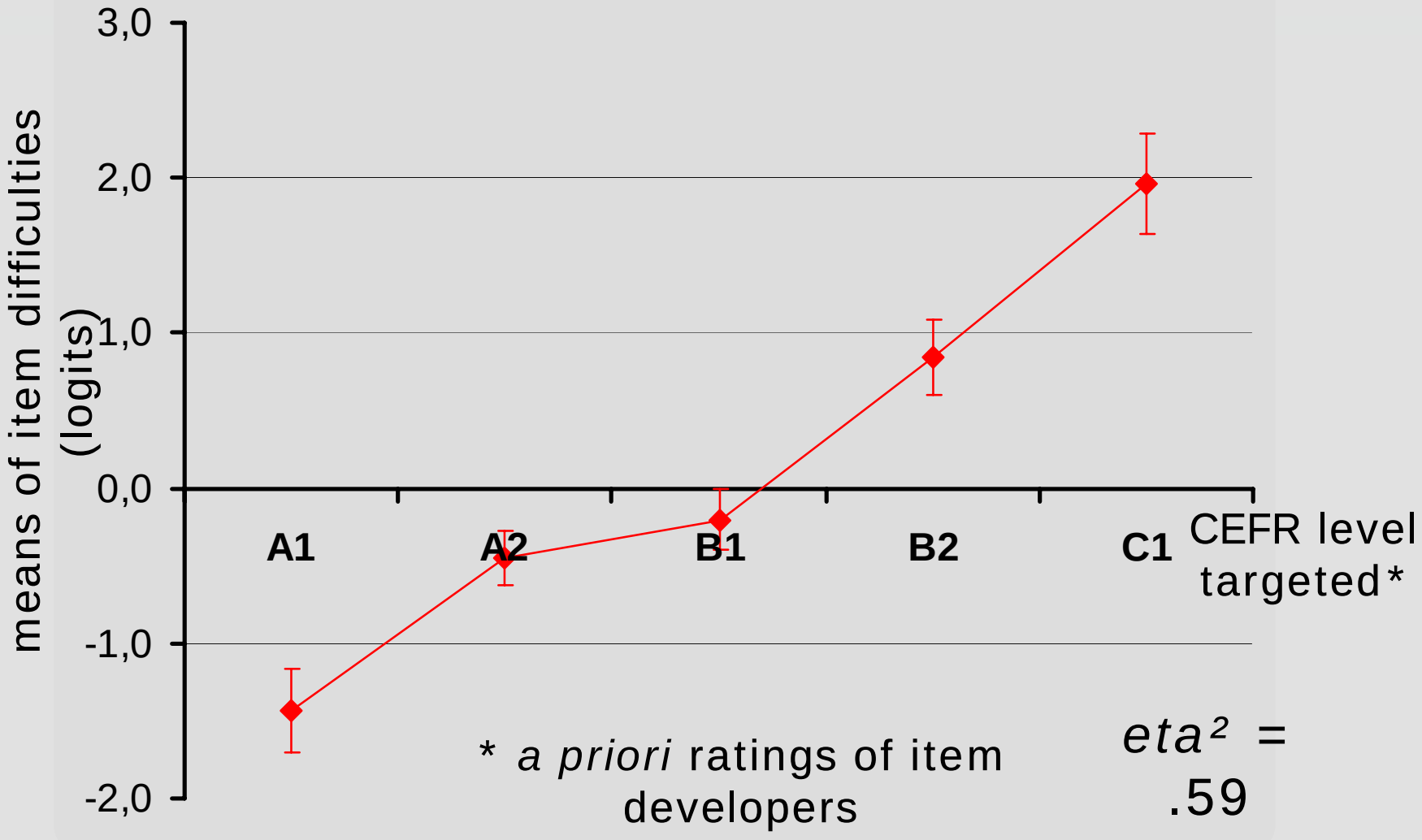
Test "usefulness" according to Bachman & Palmer (1996) is understood as ...

- argument-based
- comprising reliability, construct validity, authenticity, interactiveness, impact, and practicality
- important from the very beginning of test design.

Validity Issues addressd during Pre-Pilot Studies

- Cronbach's alphas for reading comprehension test booklets range between .76 - .90 and for listening comprehension booklets between .77 - .87
- Correlations with C-Test (r=.79) and TOEFL (r=.59) as well as with school report marks in English (r=.39)
- A-priori ratings of item developers in terms of the CEFR level of a task accounted for 60 percent of the variance in empirical item difficulties

Validity –
means of item difficulties by CEFR level targeted



Role of test-takers' perspectives

- Ethical aspect
- Practical aspect
- Familiarity of test content and its relatedness to test takers' interests is an integral part of the construct definition

Specific Goals and Research Questions

The **goal** of the study is to analyze...

(1) test-takers' acceptance of task content in terms of
- task familiarity,
- interest in task content (appeal)
- perceived difficulty of the test and
- 'subjective validity' of the test.

(2) associations between the test-takers' perceptions and their actual **test performance**.

(3) associations between the test-takers' perceptions and **'subjective validity'**.

Method: Measures
Test-Taker Feedback Questionnaire

Familiarity:

How familiar did you find the topics of the reading texts?

(1= not at all familiar, [...] 4 = very familiar)

Appeal:

How interesting did you find the texts?

(1= not at all interesting, [...] 4 = very interesting)

Perceived Difficulty:

How difficult to understand did you find the texts?

(1= not at all difficult, [...] 4 = very difficult)

Subjective / Face Validity:

How well does this test measure your English reading ability?

(1= very poorly, [...] 4 = very well)

Method: Sample Characteristics

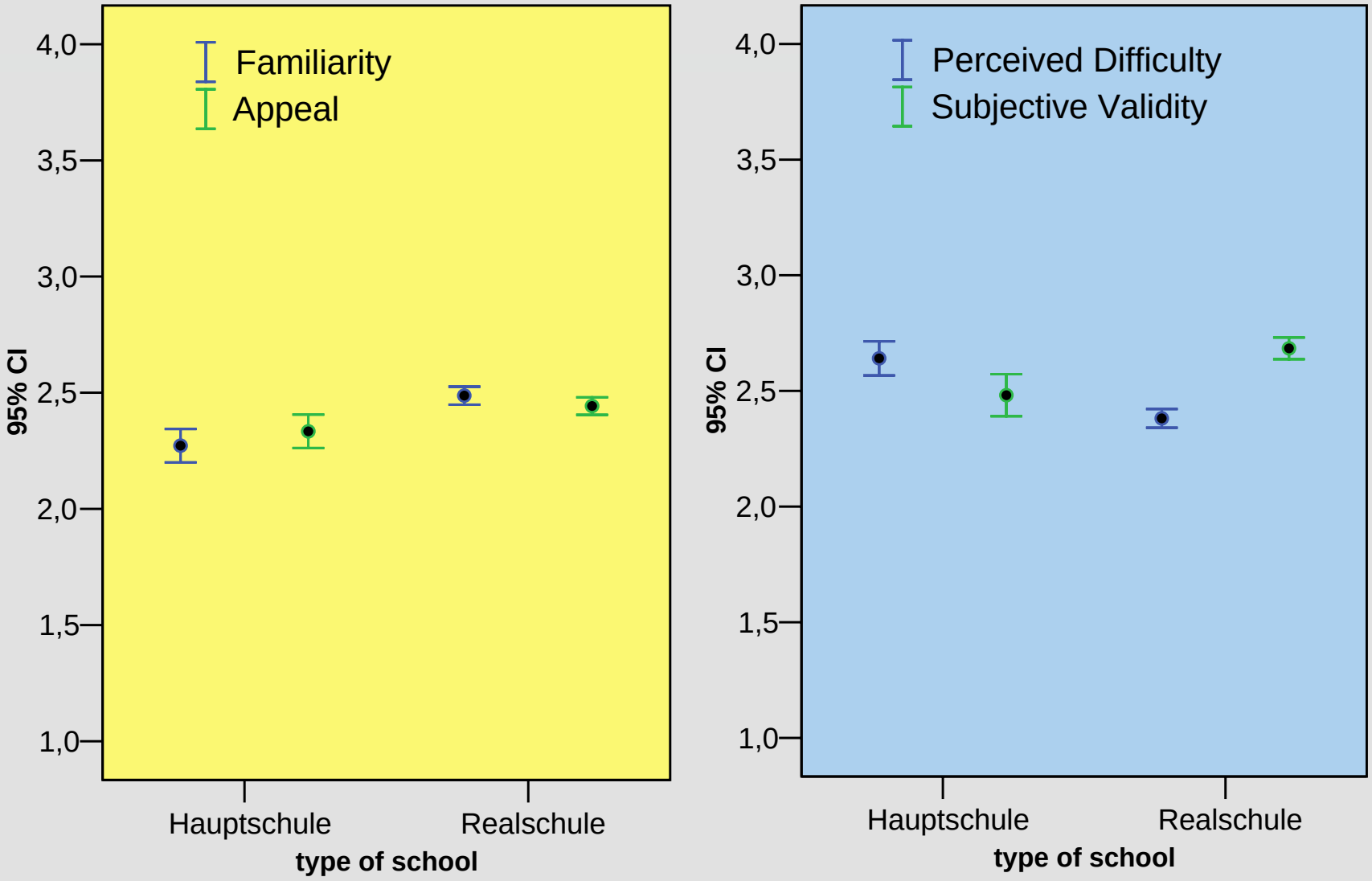
***N* = 1283**

Students in grade 9 (55%) and 10 (45%)

Secondary School tracks:

- Lower track (‘Hauptschule’): 23%
- Intermediate track (‘Realschule’): 30%
- Higher track (‘Gymnasium’): 43%
- Comprehensive school (‘Gesamtschule’): 5%

Results: Mean task perception

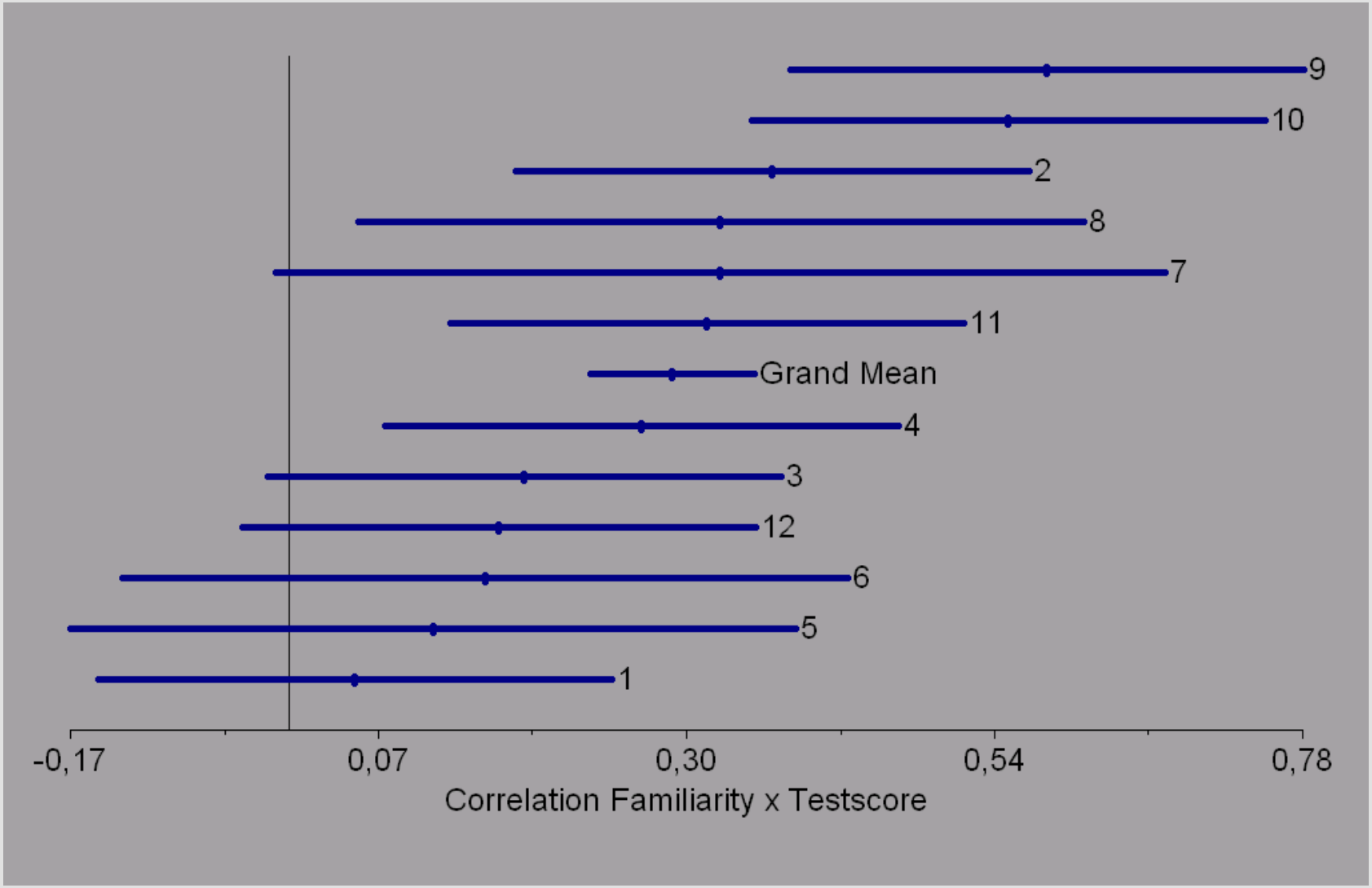


Results: Associations

Steps in meta-analysis

- Use effect sizes (here: *r*)
- Test for homogeneity of effect sizes
- In case of homogeneity: weighted aggregation indicates "true correlation"
- In case of heterogeneity: weighted aggregation indicates "average correlation" acknowledging substantive "true" differences between samples
- Search for moderators on study level (here: construct type; proportion of lower-track students) to account for differences

Results: Example Effect Size Distribution



Results: Aggregated Associations

	Familiarity	Appeal	Perceived Difficulty	Subjective Validity	Test Score
Familiarity					
Appeal					
Perceived Difficulty					
Subjective Validity					
Test Score	.23 (RC) .40 (LC)	.13	-.49	.15 (RC) .40 (LC)	

Results: Aggregated Associations

	Familiarity	Appeal	Perceived Difficulty	Subjective Validity	Test Score
Familiarity					
Appeal	.41 (het)				
Perceived Difficulty					
Subjective Validity	.24	.29	-.13 (RC) -.36 (LC)		
Test Score	.23 (RC) .40 (LC)	.13	-.49	.15 (RC) .40 (LC)	

Results: Summary

- On average, indicators of test acceptance are located in the "semantic middle" of the scale
- Lower-track students are less familiar and interested in the chosen task content
- Associations between test scores and indicators of acceptance are low to medium and differ significantly between test constructs
- 'Subjective validity' is moderately correlated with both test scores and indicators of acceptance and, hence, might be reflecting hindsight-bias and "true" evaluation of the test content

Conclusions

- The good news:
  Test-takers to a great deal evaluated tasks as being interesting and familiar, thus corroborating the validity argument for an import part of the construct

- The ambivalent news:
  Have we as test developers accomplished enough in terms of construing equally interesting and familiar tests for EVERYONE?
  Is 15 percent of variance in test results accounted for by indicators of acceptance too much, too little, or...?

Institut zur Qualitätsentwicklung im Bildungswesen

Dr. Hans Anand Pant
Dr. Miriam Vock
phone +49[30]2093-5501
fax +49[30]2093-5336
Anand.Pant@IQB.hu-berlin.de

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Kultusminister KONFERENZ

Thank you!

Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

German National Educational Standards (NES) for English as a foreign language (EFL)

Functional communication competencies	
Communication skills	Application of linguistic resources
A) Reading comprehension B) Listening comprehension C) Writing	A) Vocabulary B) Grammar C) Pronunciation and intonation
Intercultural competencies	
A) Sensitive approach to cultural diversity	
Methodological competencies	
A) Text reception B) Learning strategies	

Note. Table was simplified. From KMK (2003, 2004, S. 8).

Beispiel:

Text Sources	Text Types	Topics
• Internet	• Interviews	• Everyday life (e.g., hobbies, school, friends)
• Radio	• Conversations (phone-ins)	• Free time
• TV	• Commercials	• Travelling
• CD	• Ads	• Entertainment
• DVD	• Reports	• House and home
• Audio Books	• Announcements	• Environment
• Answering machines	• Directions	• Health and body care
• Public announcements	• Messages	• Education/academia
• (Private) Recordings	• Instructions	• Shopping
• Audio tours/guides	• Weather forecast	• Food and drink
• Live interviews	• Stories	• Services
• Telephone	• Monologues	• Places
	• News	• Weather
	• Limericks	• World of work
	• Dialogues	• Multicultural society
	• Lectures	• Celebrities
	• Speeches	• Animals
	• Talks	• Festivals/customs
	• Academic presentations	• Global problems
	• Jokes	• Crime
	• Anecdotes	• History
	• Radio/Audio programs	• Science and technology
	• News (radio, TV)	• Aspects of society
	• Sports commentaries	• Adventures and challenges
	• Films	
	• Movies	
	• Telephone conversations	
	• Talk shows	
	• Plays	
	• Discussions	
	• debates	
	• Technical information	
	• Documentaries	

Item development process

Item development by 20 German teachers that received skill related workshops in item writing. Monitoring of item development process by a panel of national and international experts in didactics and educational research.

Item development as a four stage process:

(1) Familiarization with a) German NES for EFL and b) CEFR.

(2) Conceptualization of item pools via *test specifications* (Alderson, Clapham & Wall, 1995)

(3) Retrieving information addressed by items via *text mapping* (Sarig, 1989)

(4) Small sample item evaluation and item revision

Five item pools resulted, targeting reading comprehension skills on five CEFR levels of foreign language use: A1 and A2 (basic language use), B1 and B2 (independent language use), C1 (proficient language use)

Sample Task at CEFR Level B1

Task: Read the text. Then decide if the statements are true, false or not given in the text. Tick (✓) the correct box. There is an example at the beginning (0).

Bear Safety

Backcountry hiking and camping are wonderful ways to see wildlife in the open. When travelling through bear country, however, it is important to follow a few rules for your own safety.

Safety rules while hiking

- Make your presence known. Singing, speaking loudly or making similar human-sounding noises as you hike will make grizzly bears aware of your presence and help ensure that any bears in the area have enough time to leave.
- While in grizzly country, hike in groups of three or more. This is another way to help ensure that bears are conscious of your presence.
- As much as possible, stay in open areas where you can be seen. Stay at least a quarter mile away from any bear you see.
- Avoid areas where there are clear signs of bears (for example, along streams where bears may be fishing, or places where bear tracks are visible).
- Be especially careful when hiking in the first or last daylight as bears may be more active these times.

Sample Task at CEFR Level B1 (continued)

Safety rules while camping

- Do not allow bears to find human food. Store all food (both cooked and uncooked) and trash in Bear Resistant Food Containers (BRFC). These hard-plastic containers are issued free of charge. The use of these containers keeps bears from associating campsites with food.
- Set up your area for food preparation at least 100 yards from your sleeping area. Never cook or eat in your tent.
- Smelling human will help the bear recognize that you are a person. Never sleep in clothes that you have worn while fishing or hunting.

		True	False	Not given
0	Singing and speaking loudly attracts bears.		✔	
Q14	Hiking in the early morning and late afternoon is safe.			
Q15	When leaving your camping area you may leave some food for the bears.			
Q16	Having your meals in the tent is dangerous.			
Q17	Make sure there are some high trees that you may climb if a bear attacks you.			
Q18	Don't wear red clothes when hiking.			

Methods – established EFL tests

test	description
C- Test	Overall EFL language ability (e.g. Coleman, Grothjahn, and Raatz, 2002).
'Curricular'	EFL reading comprehension as demanded in German curricula for secondary level (Nold & Rossa, 2006; cf. Beck & Klieme, 2006).
TOEFL	EFL reading comprehension as required at a pre- university level.

Note. WLE person separation reliabilities: 0.96 (C- Test), 0.80 ('Curricular'), and 0.63 (TOEFL).

Results – Correlations between Rasch model means for students' proficiencies and …

… proficiencies in established EFL tests			
	C- Test	'Curricular'	TOEFL
NES reading scale	.79**	.72**	.59**
… school report marks [1]			
	Math	German	English
NES reading scale	- .18**	- .28**	- .39**

Note. ** $p < .01$. [1] School report marks in German secondary school ranging from 1 ('very good') to 6 ('insufficient').