Two tests, both alike in validity?

A concurrent validity study of a two academic proficiency tests

CNaVT
CNaVT CIEAUT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL

ITNA

KU LEUVEN
Universiteit Antwerpen
UNIVERSITEIT GENT
Vrije Universiteit Brussel

Koen Van Gorp, Lucia Luyten, Sabine Steemans & Lieve De Wachter

CNaVT, Centre for Language & Education, University of Leuven
ILT, University of Leuven
Linguapolis, University of Antwerp

1 Two academic proficiency tests

ITNA = Inter University Test of Dutch as an L2

Organized by consortium of language institutes of the main Flemish universities

PTHO = Profile Language Proficiency for Higher Education

Centrally organized by the Certificate of Dutch as a Foreign Language (CNaVT) (= KU Leuven & Fontys Tilburg)

*This research is a joint project of ITNA & CNaVT

1

ITNA: Dutch as L2

1 PTHO: Dutch as FL

1 different operationalisations

ITNA	PTHO
Written part: Language in Use, R & L Exclusion	Written part: Pen & paper Task-based and integrated
Linguistic criteria	Content & linguistic criteria

Face-to-face oral part:
1. Presentation 2. Argumentative speaking

Linguistic criteria	Content & linguistic criteria

Target level: B2

Goal: university admission

Ability	Can Do

Bachman & Palmer 2010,
Bachman 2013

1 but one consequence.

To enter or not to enter higher education

2 RQ: Concurrent validity

To what extent does test Y (ITNA) correlate with a previously validated test X (PTHO)?

To what extent are they both measures of the same underlying skill?

Taking into account that ...

- Written part of ITNA and PTHO is quite distinct.
- Oral part of both tests is quite similar.

2 Live test data

		ITNA		Total
		0	1	
PTHO	0	16	13	29
	1	3	32	35
Total		19	45	64

		Value	Approx. Sig.
Measure of Agreement	Kappa	.450	.000
	Pearson	.51**	.000
N of Valid Cases		64	

2 Quantitative study

Population:	77 prospective L2 students
Location:	3 universities (Ghent, Leuven & Antwerp)
Timing:	1 week apart, different orders

2 Quantitative study (part 1)

ITNA Computer-based	**PTHO** Paper-based
Language in Use Closed vocabulary Closed grammar Gap filling	Receptive listening Integrated writing (lecture summary)
Reading Re-arrange sentences MC reading texts	Receptive reading Integrated writing (text summary) Semi-independent writing (argumentation)
Listening MC listening texts Dictation	

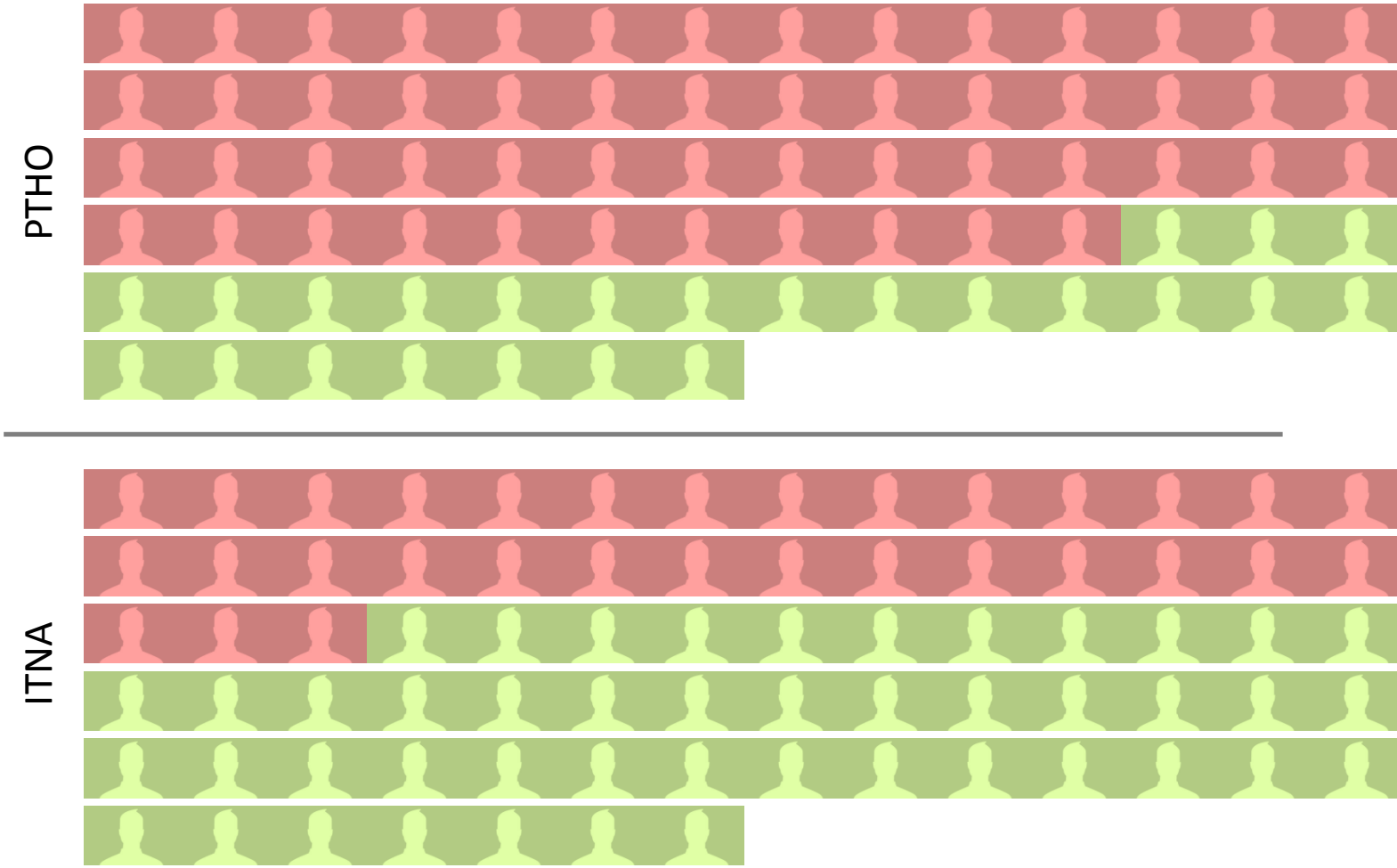
2 Quantitative study (part 1)

		ITNA		Total
		0	1	
PTHO	0	30	23	53
	1	1	23	24
Total		31	46	77

		Value	Approx. Sig.
Measure of Agreement	Kappa	.419	.000
	Pearson	.77**	.000
N of Valid Cases		77	

Correlation PTE / IELTS = .73

2 Quantitative study



2 Quantitative study (S)

ITNA	**PTHO**
Face-to-face	Face-to-face
Presentation Argumentative speaking	Presentation Argumentative speaking

2 Quantitative study (S)

		ITNA		Total
		0	1	
PTHO	0	4	6	10
	1	7	21	28
Total		11	27	38

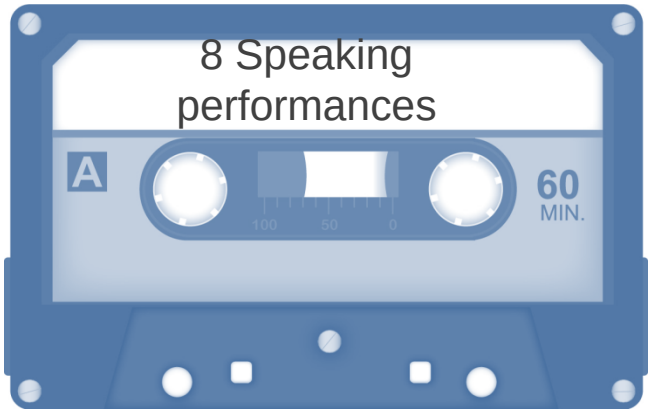
		Value	Approx. Sig.
Measure of Agreement	Kappa	.145	.369
	Pearson	.15	.136
N of Valid Cases		38	

2 Quantitative study

Part 1:
Highly dissimilar operationalization, but moderate agreement and .77 correlation

Speaking:
Parallel operationalization, but slight agreement and .15 correlation

2 Qualitative study



9 raters

Council of Europe
Conseil de l'Europe

COMMON
EUROPEAN
FRAMEWORK

2 Qualitative study

B2? Independent user, abstract language, academic domain

Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

CEFR: 24

2 Qualitative study

Ordinal level on an intuitive scale

Cf. Fulcher 2012, Little 2007, Alderson 2007

→ Asymmetry in the attention to productive vs receptive skills

Alderson 2004, Fulcher 2004, Weir 2005, Alderson 2007, Davidson & Fulcher 2007, Staehr 2008, Milton 2010

→ Vagueness and inconsistencies in level descriptors

Fulcher 2004, Alderson 2007

"Relatively high degree of grammatical control [without] mistakes which lead to misunderstanding"

(lower end B2)

"Generally good control […] errors occur, but it is clear what he/she is trying to express"

(higher end B1)

2 Qualitative study

CEFR Linking

Rater	Test	Performance							
		1	2	3	4	5	6	7	8
1	ITNA								
2	ITNA								
3	ITNA								
4	ITNA								
5	PTHO								
6	PTHO								
7	PTHO								
8	PTHO								
9	PTHO								

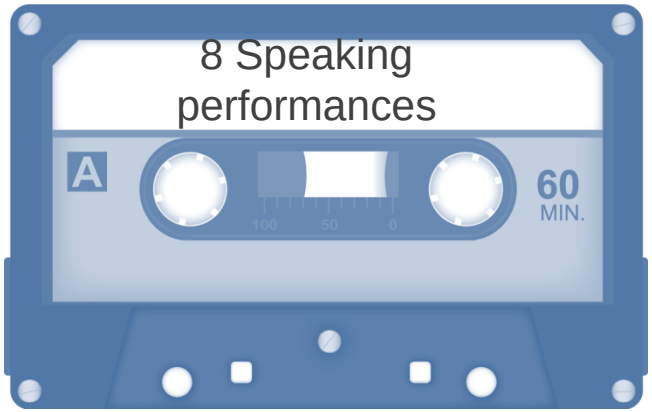
2 Qualitative study

CEFR Linking

Rater	Test	Performance							
		1	2	3	4	5	6	7	8
1	ITNA	B1+	B1+	A2	A2	B2-	B2	B2+	B2-
2	ITNA	B1	B1	A2+	A2+	B1+	B2	B2+	B2+
3	ITNA	B1	A2	A2	A2	B1	C1	C1	B2
4	ITNA	B1+	B1	B1	A2+	C1	B2	C2	B2
5	PTHO	B1+	B2	A2	A2	B2	B2	B1+	B2
6	PTHO	B2-	B2	B1	B1	B2	B2-	B1	B2
7	PTHO	B1	B2	B1	A2	B2	B2	B2	B2
8	PTHO	B1	B2	B1	B1	B2	B2	B2	B2
9	PTHO	B1	B1	A2	A2	B2	B2	B2	B2

Raters: similar understanding of CEFR levels

2 Qualitative study



9 raters

Council of Europe
Conseil de l'Europe
COMMON EUROPEAN FRAMEWORK

ITNA rated

PTHO rated

2 Qualitative study

Judgment

Performance	Rating team	Common criteria				
		V	G	Pa	S	Pr
1	ITNA					
B1	PTHO					
2	ITNA					
B1+	PTHO					
5	ITNA					
B2	PTHO					
6	ITNA					
B2	PTHO					
8	ITNA					
B2	PTHO					

2 Qualitative study

Judgment

Performance	Rating team	Common criteria				
		V	**G**	**Pa**	**S**	**Pr**
1	ITNA	0	0	0	1	*0*
B1	PTHO	0	0	0	1	*1*
2	ITNA	1	0	0	0	1
B1+	PTHO	1	0	0	0	1
5	ITNA	1	1	1	1	0
B2	PTHO	1	1	0	1	0
6	ITNA	1	1	1	1	1
B2	PTHO	1	1	1	1	1
8	ITNA	0	1	1	1	0
B2	PTHO	0	1	1	1	1

Form-focused rating criteria interpreted and used in the same way (but not all criteria are form-focused)

2 Causes for mismatch?

Test format

ITNA: Computer-based

PTHO: Paper-based

> Test mode influences test-taker's motivation

Endres 2012, Piaw 2012

2 Causes for mismatch?

Test format

Tasks

ITNA: No written performance tasks
PTHO: Summarizing and argumentative writing

> Problem of determining and maintaining difficulty in integrated writing tasks

Bachman 2002, Ross 2012

2 Causes for mismatch?

Test format
Tasks

Exclusion yes/no

ITNA: exclusion after failed part 1
PTHO: candidate can compensate for weaker written performance

> Truncated sample problem

Alderson, Clapham & Wall, 1995

2 Causes for mismatch?

Test format
Tasks
Exclusion yes/no

Spoken criteria

ITNA: Only linguistic criteria
PTHO: Linguistic and content-specific criteria

> Impact of topic choice in integrated tasks

Sawaki 2009, Yu 2009

2 Questions

Test format

Exclusion yes/no

Tasks

Does a proficiency test need writing tasks?

Does a LAP test need writing tasks?

Spoken criteria

Does a language test need content-specific criteria?

3 Future steps and research

Both ITNA and PTHO are currently investing in their rating scales.

ITNA: how are the rating scales interpreted by the different raters (interrater reliability)

PTHO: An iterative three-year rating scale construction and validation process; moving from a dichotomous scale towards a four-band scale based on the CEFR.

Moving towards more comparable rating scales: what is the impact on the overall rating of the speaking performances of both tests?

3 PhD research

To what extent do test scores on Dutch proficiency tests **predict** students' actual coping with academic language during their studies?

To what extent is the performance elicited by the test items/tasks **generalizable** to the broader field of academic language skills?

Thank you

koen.vangorp@arts.kuleuven.be
lucia.luyten@arts.kuleuven.be
sabine.steemans@uantwerpen.be
lieve.dewachter@ilt.kuleuven.be