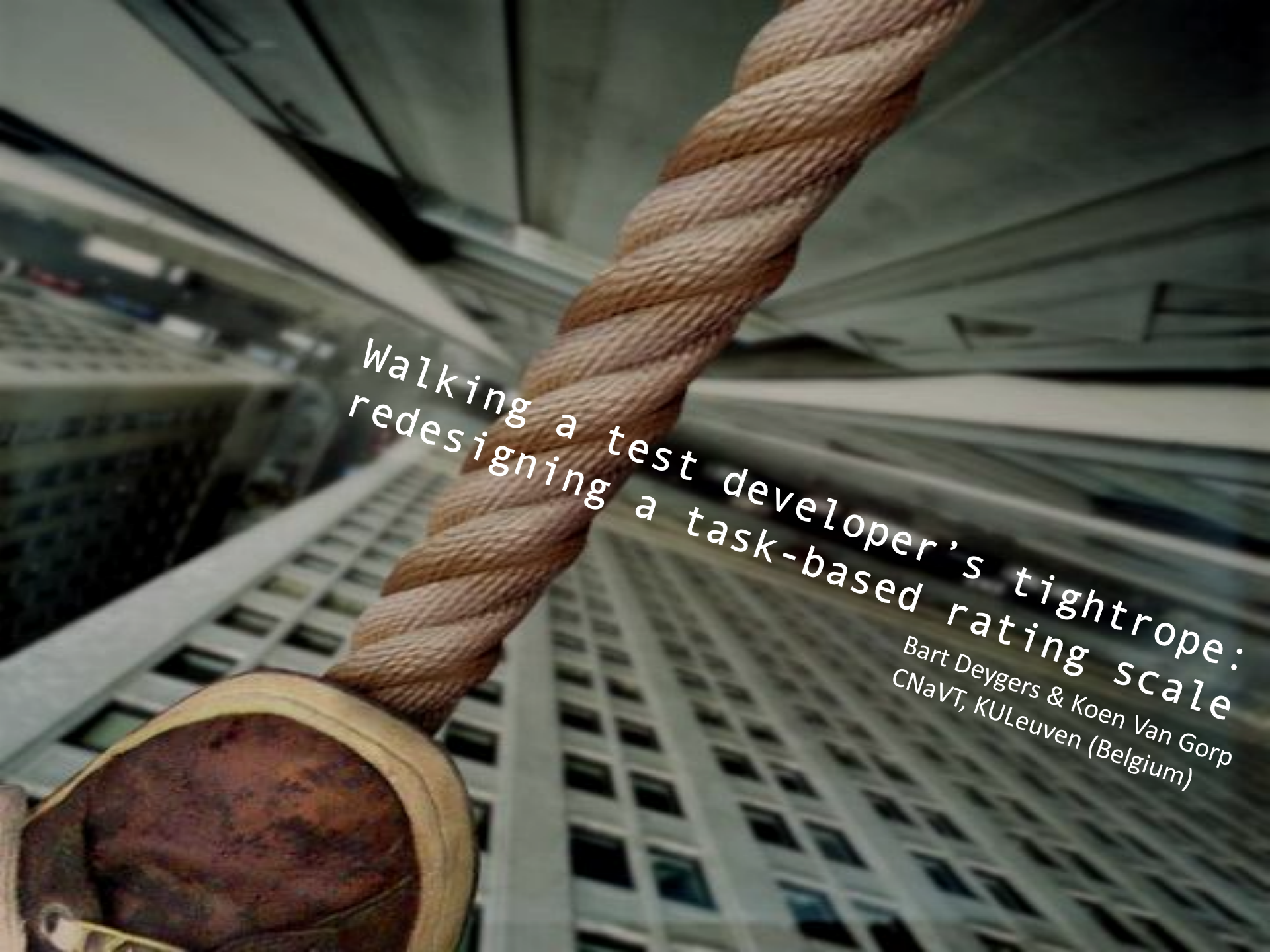


Walking a test developer's tightrope: redesigning a task-based rating scale

Bart Deygers & Koen Van Gorp
CNaVT, KULeuven (Belgium)



The CNaVT uses a functional, task-based approach

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

CNaVT: six profiles

- "Can-do" approach

(Bachman 2011)

- Construct is formulated in terms of task performance

(Colpin and Gysen 2006, Long and Norris 2000)

- Starting point = functional use

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



The CNaVT uses a functional, task-based approach

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Focus: Dutch for Academic Purposes

- "Can-do" approach

(Bachman 2011)

- Construct is formulated in terms of task performance

(Colpin and Gysen 2006, Long and Norris 2000)

- Starting point = functional use

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Using domain-related tasks has its consequences

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Subject specialists

- Input & Feedback

(Douglas 2000, 2001)

- Boundary Problem

(Davies 2001)

- 2009: rating scale concerns raised

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Using domain-related tasks has its consequences

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Subject specialists

- Input & Feedback

(Douglas 2000, 2001)

- Boundary Problem

(Davies 2001)

- 2009: rating scale concerns raised

↓

Dichotomous
Least able user
Getting the message across

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Using domain-related tasks has its consequences

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Subject specialists

- Input & Feedback

(Douglas 2000, 2001)

- Boundary Problem

(Davies 2001)

- 2009: rating scale concerns raised
  - Dichotomous
  - Least able user
  - Getting the message across

↓

Rating process: fundamental impact on test validity and test fairness

(Weigle 2002; Barkaoui, 2010)

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Using domain-related tasks has its consequences

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Subject specialists

- Input & Feedback

(Douglas 2000, 2001)

- Boundary Problem

(Davies 2001)

- 2009: rating scale concerns raised

↓

Need to verify
"indigenous" criteria

(Jacoby & McNamara 1999)

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



We had to have a thorough look at the rating scales

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

… and pretty much everything else

- Tasks
- Criteria
- Rating process

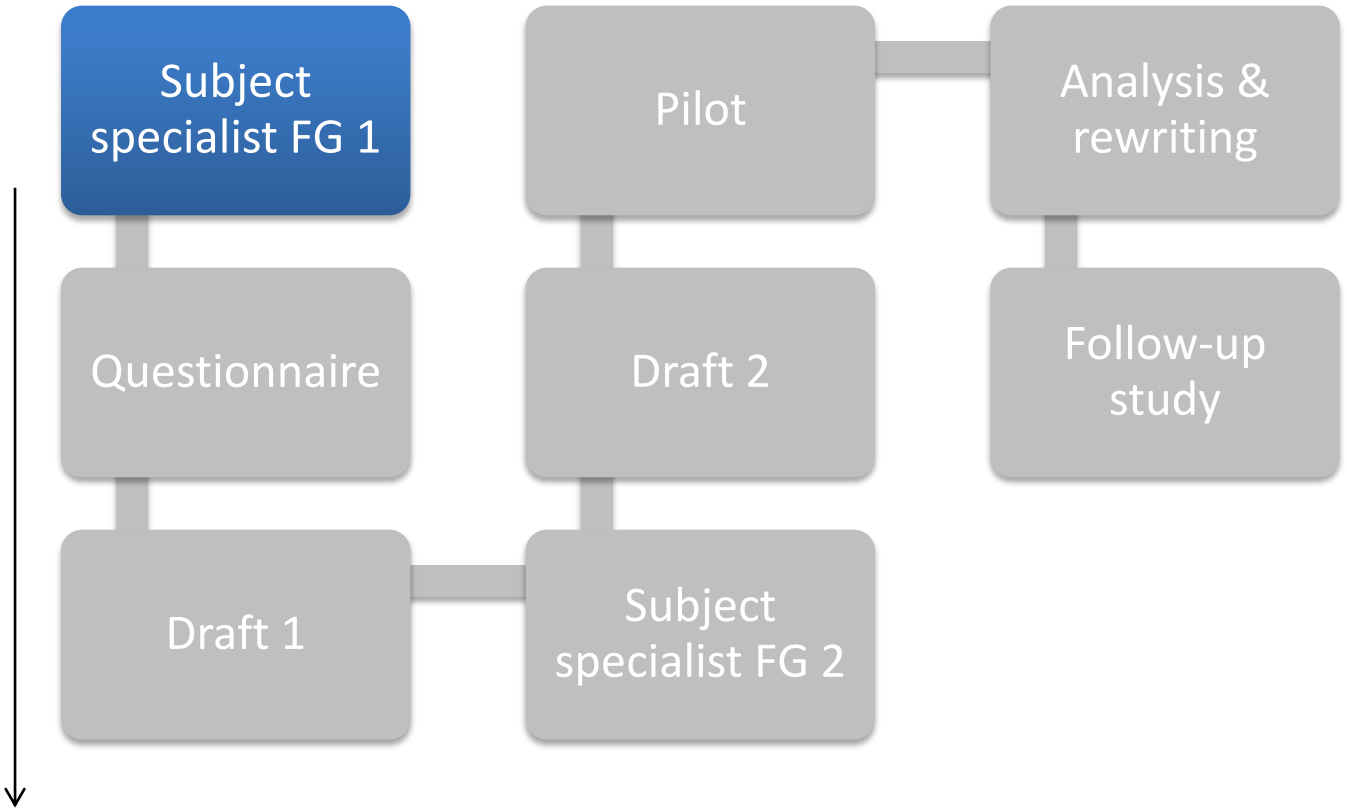
CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Reexamining the rating process

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Research path



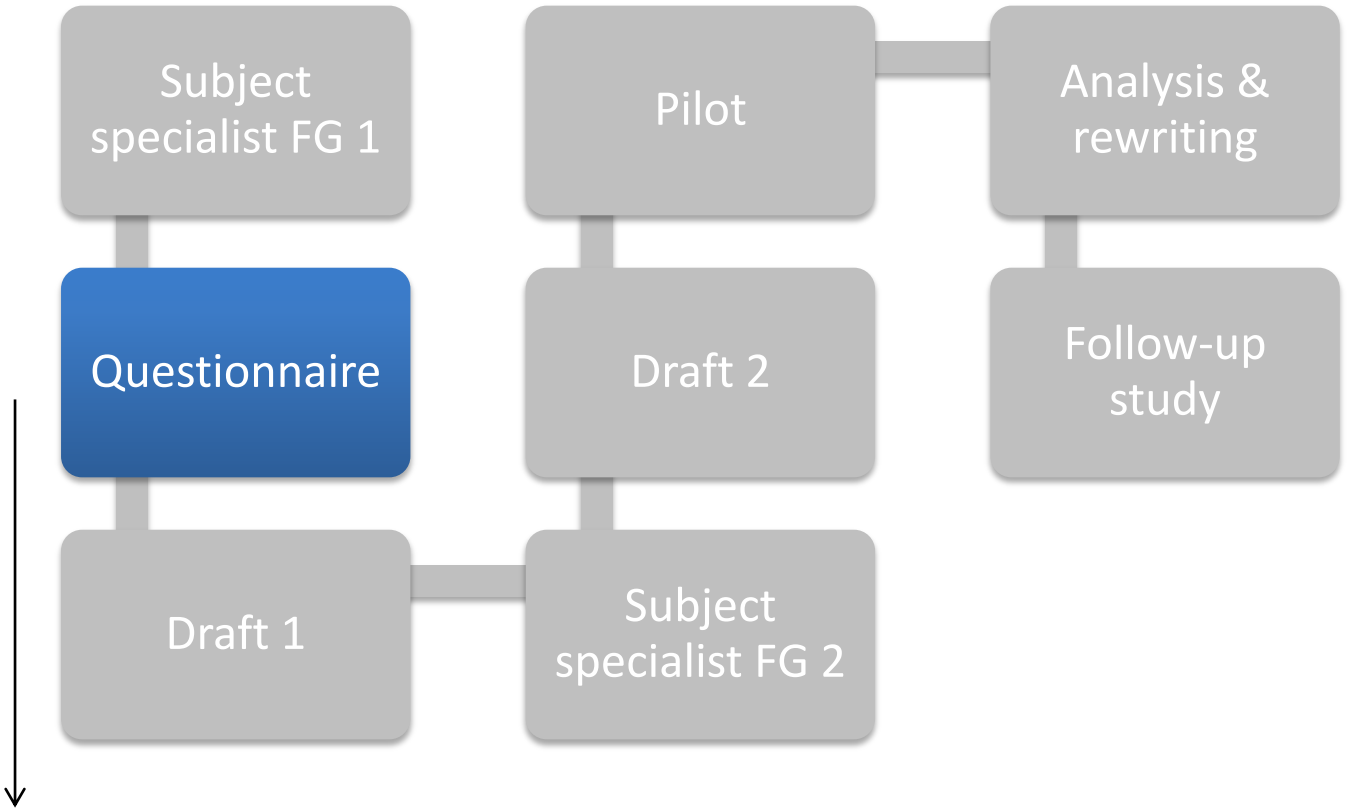
CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Reexamining the rating process

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Research path



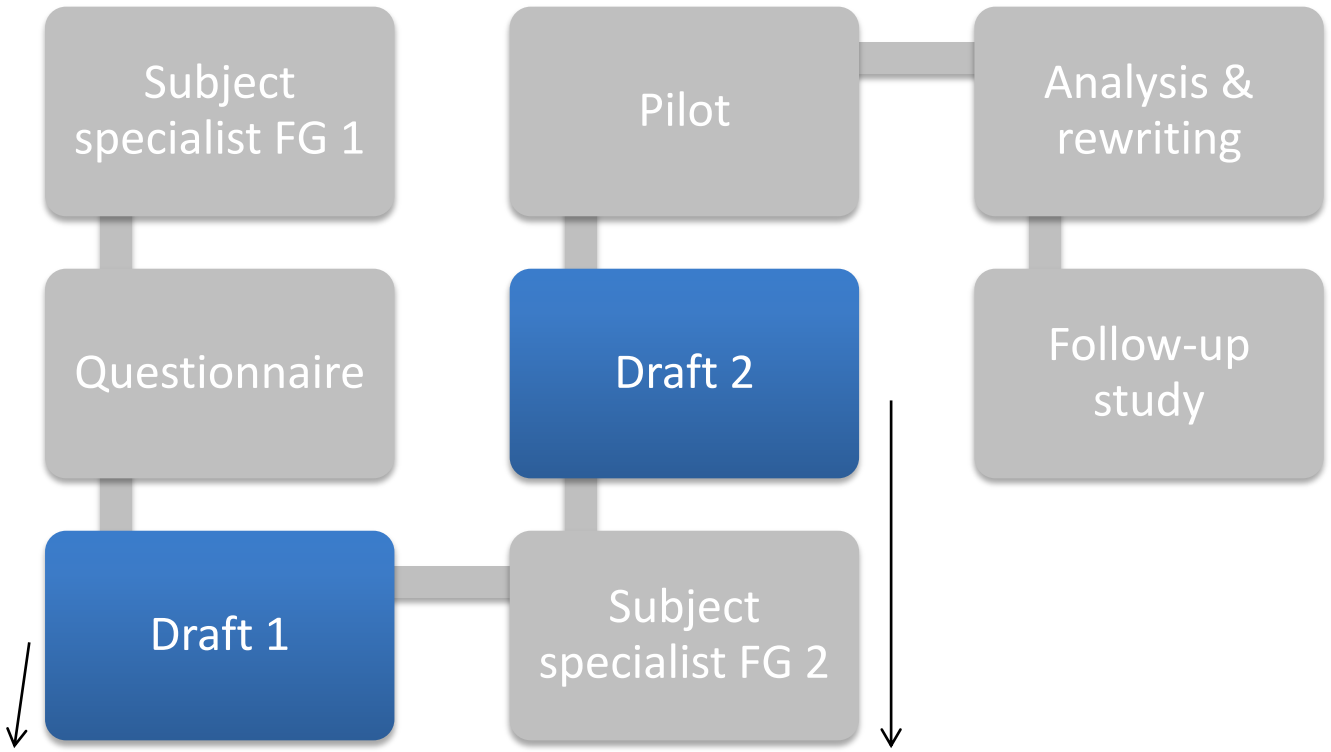
CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Reexamining the rating process

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Research path



Iterative & dialogical process, influenced by CEFR & subject specialists

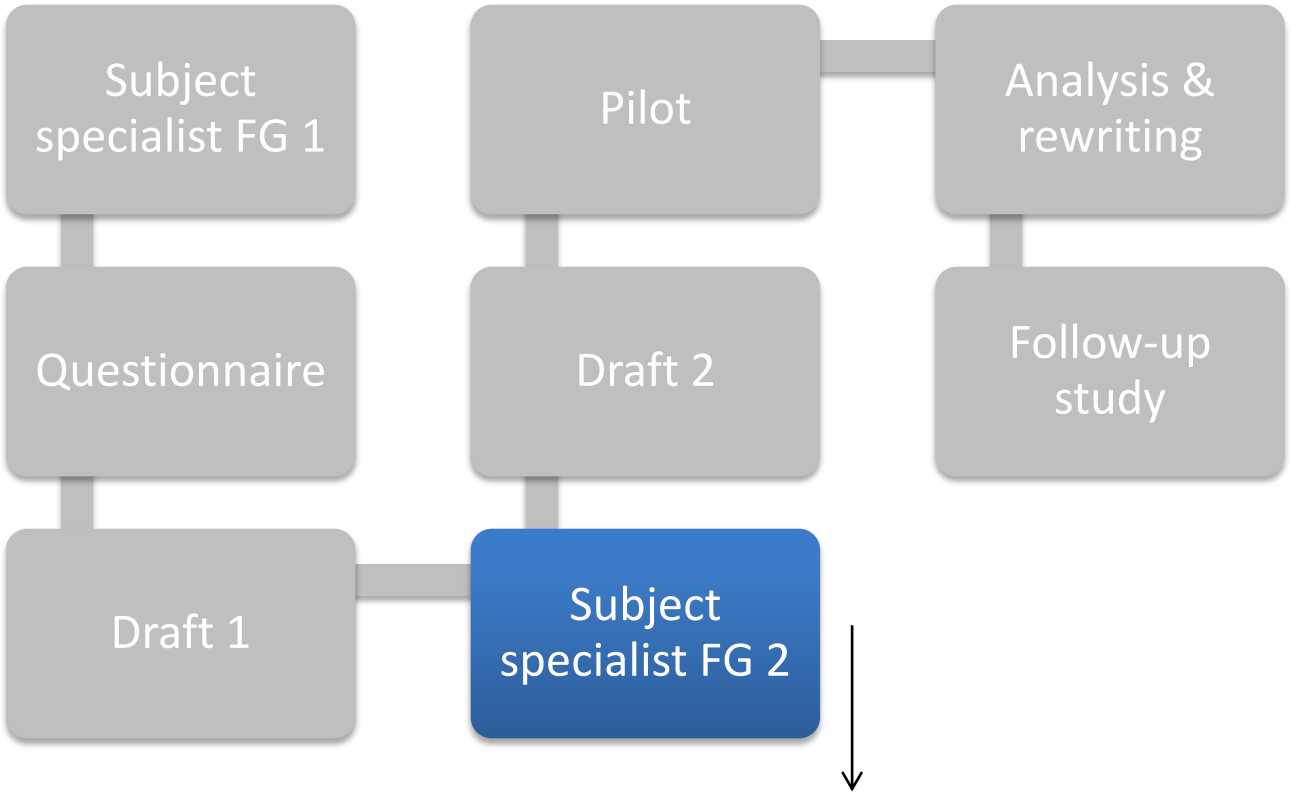
CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Reexamining the rating process

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Research path



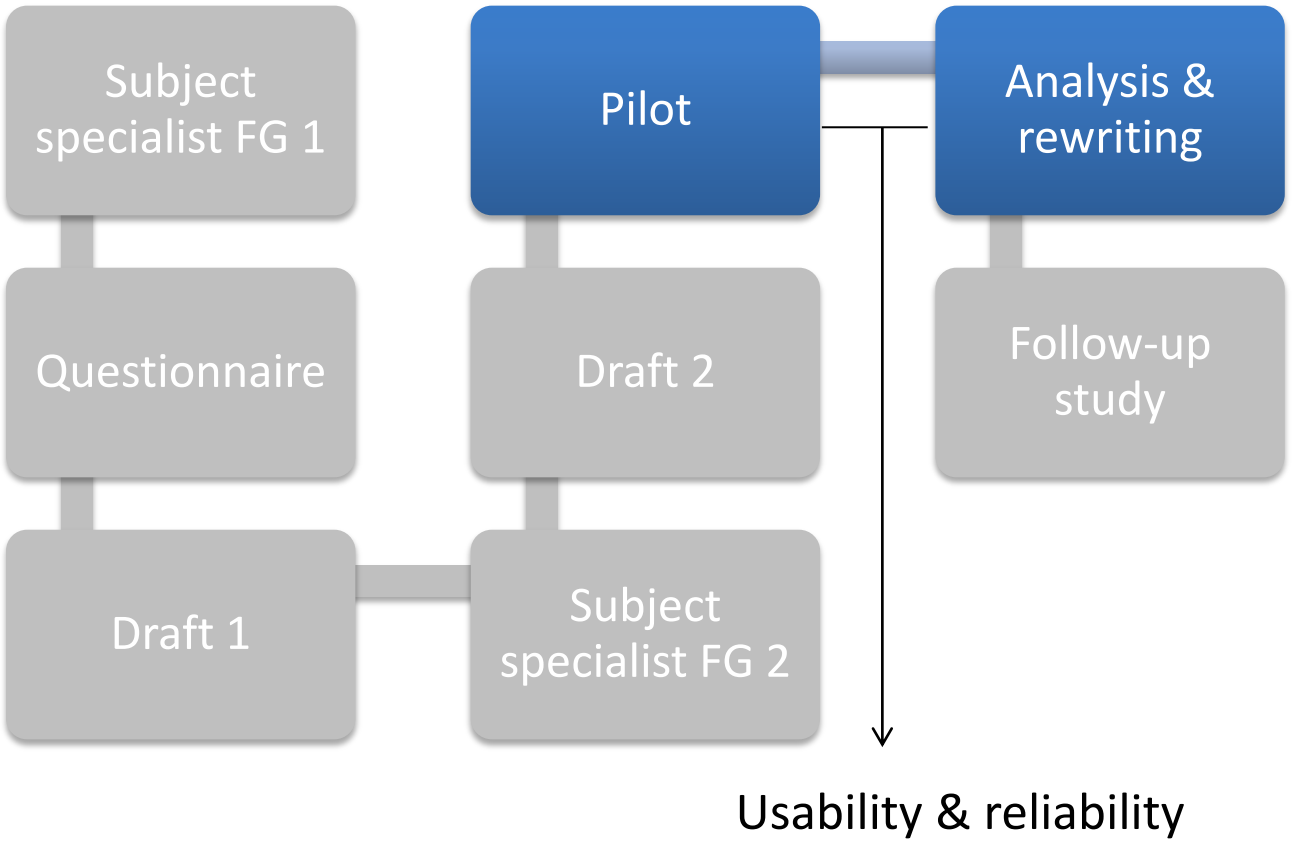
CNaVT CITO
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Reexamining the rating process

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Research path



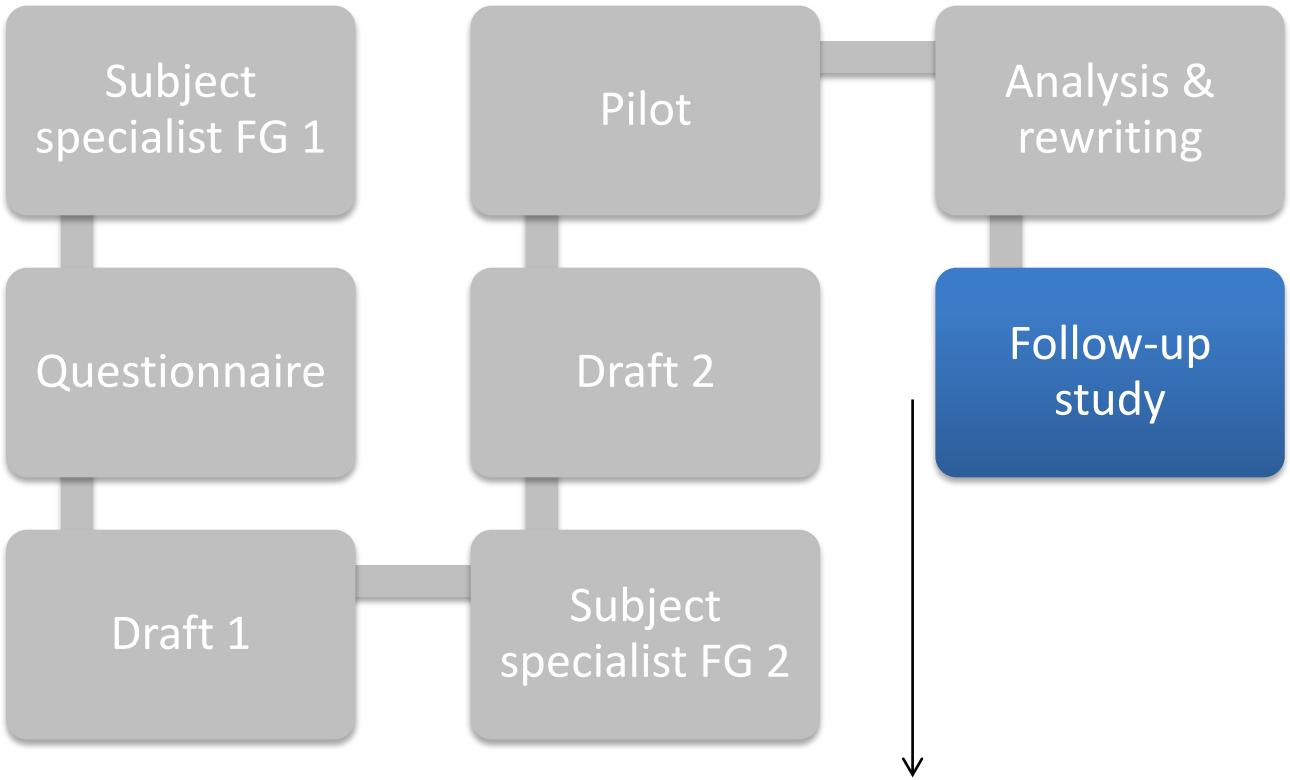
CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Reexamining the rating process

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Research path



CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Choosing a rating scale

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

- Holistic
- Analytic

(Alderson 1995, Weigle 2002, Knoch 2009, Barkaoui 2011)

- Measurement driven
- Performance driven

(Fulcher 2003, 2010)

‘scales are inevitably of somewhat limited validity, because of their inability to describe texts adequately’

Lumley (2002: 268)

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Choosing a rating scale

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

L2 Feedback

Novice rater reliability → Holistic / Analytic

Draws attention to scales

(Alderson 1995, Weigle 2002, Knoch 2009, Barkaoui 2011)

CEFR → Measurement driven / Performance driven

- Crucial point of reference even though the CEFR '*was not designed specifically for test specifications and language testing contexts*' (North 2004 in Papageorgiou 2010: 273)

(Fulcher 2003, 2010)

CNaVT CINAT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Two rating scales compared: Dichotomous

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Focus: getting the message across

Profielexamen 2011 PPT	Code kandidaat:	11 / … … … … … … / PPT / … … … …
	Naam kandidaat:	……
	Code beoordelaar:	A …… B …… C ……

DEEL A

Taak 1 Platenfirma			
Preconditie	A - B		
Globaal	1 - 2 - 3 - 4		
1a			
1b			
1c			
1d			
1e			
1f			
1g			
1h			

Taak 2 Kantoorwinkel.com			
Item	0 of 1	8	F
1 (B)			
2 (C)			
3 (B)			
4 (A)			
5 (B)			
6 (D)			
7 (C)			
8 (D)			
9 (D)			

CNaVT CNaVT CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Two rating scales compared: Polytomous

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Focus: getting the message across *adequately*

Structuur en samenhang (Ss)	
De prestatie vormt een samenhangend geheel met een duidelijke structuur die de inhoud foutloos ondersteunt. De presentatie is perfect te volgen. *De prestatie toont een adequaat en gevarieerd gebruik van ordeningspatronen en verbindingswoorden, waardoor ze vlot te volgen is. De algemene structuur en het interne ordeningspatroon ondersteunen de inhoud.*	A
De prestatie vormt een samenhangend geheel met een grotendeels logische structuur en grotendeels voldoende samenhang. Af en toe verliest de luisteraar het overzicht. *Cohesiebevorderende elementen worden doorgaans correct gebruikt. Grotendeels voldoende samenhang, hoewel bepaalde onderdelen van de communicatie niet steeds effectief zijn door occasionele tekortkomingen.*	B
De prestatie wordt gekenmerkt door een zekere springerigheid en/of een soms onduidelijke samenhang. *Cohesiebevorderende elementen zijn beperkt tot vaak voorkomende, lineaire verbindingswoorden. De samenhang is onvoldoende wegens fout gebruik van verbindingswoorden of wegens afwezigheid ervan.*	C
De prestatie vertoont nauwelijks structuur of samenhang. *Zodanig fragmentarisch of inadequaat opgebouwd dat de bedoelde boodschap moeilijk te begrijpen is.*	D
Woordenschat en woordgebruik (W)	
De lexicale variatie in de prestatie is breed. De woordkeuze is precies en nagenoeg foutloos. *De prestatie bevat een breed lexicaal repertoire en getuigt van een correcte, genuanceerde woordkeuze. Slechts per uitzondering kunnen kleine vergissingen optreden.*	A

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Still, it isn't the scale but the rater who decides

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

The rater

- Interprets the scale
- Is influenced by the linguistic complexity of the descriptors

(Alderson 1995, Shaw 2002, Fulcher 2010)

Focus on scale use

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



The research design involves four paired novice raters

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Rater A/B Performance 1-125	
Polytomous ↓ Dichotomous	

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



The research design involves four paired novice raters

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

"Decision booklet"

(Knoch 2007)

2-day training /rating scale

(Shohamy 1992, Weigle 1994, Lumley 2002)

Rater A/B	Rater C/D
Performance 1-125	Performance 126-250
Polytomous	Dichotomous
↓	↓
Dichotomous	Polytomous

- Clustered in 25 tasks
- Preselected:
  - Linguistic ability
  - Geographical dispersion
  - L1

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



The research design involves four paired novice raters

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Rater A/B Performance 1-125	Rater C/D Performance 126-250
Polytomous r = .47 K = .30 α = .76	

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



The research design involves four paired novice raters

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Rater A/B Performance 1-125	Rater C/D Performance 126-250
Polytomous r = .47 K = .30 α = .76 ↓ Dichotomous r = .82 K = .59 α = .77	Sequence effect?

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



The research design involves four paired novice raters

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Rater A/B Performance 1-125	Rater C/D Performance 126-250
Polytomous	Dichotomous
r = .47	r = .94
K = .30	K = .54
α = .76	α = .80
↓	
Dichotomous	
r = .82	
K = .59	
α = .77	

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



The research design involves four paired novice raters

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Rater A/B Performance 1-125	Rater C/D Performance 126-250
Polytomous	Dichotomous
r = .47	r = .94
K = .30	K = .54
α = .76	α = .80
↓	↓
Dichotomous	Polytomous
r = .82	r = .79
K = .59	K = .35
α = .77	α =.86

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



The research design involves four paired novice raters

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Rater A/B	Rater C/D
Performance 1-125	Performance 126-250
Polytomous	Dichotomous
r = .47	r = .94
K = .30	K = .54
α = .76	α = .80
~~Sequence effect~~	
Dichotomous	Polytomous
r = .82	r = .79
K = .59	K = .35
α = .77	α =.86

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Irrespective of rating sequence, the dichotomous scale is more reliable

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

However

Robustness of correlations is linked to number of options

'Unweighted kappa coefficients decrease with the number of categories'

(Brenner and Kliebst 1996: 199)

Impressionistic descriptors result in lower inter-rater reliability

(Knoch 2008: 61)

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Results of the qualitative research (1)

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Focus group 1

	Dichotomous	Polytomous
Subjective duration of rating	☺	☹
General preference	☺	☹
Preference for assessing writing	☺	☹
Preference for assessing speaking	☹	☺

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Results of the qualitative research (1)

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Focus group 1

Dichotomous	Polytomous
Easy to memorise	In line with intuition
Induces conficence	Too much doubt
Superficial	Complicated
Too crude	Too vague

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Follow-up study

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative (2)
Conclusions
Challenges

Rewriting the scale together with the raters

Concrete insertion

The argumentation is convincing, clear and logical.

All the required information is used in order to build a sound argumentation

---

The argumentation is convincing, barring occasional logical flaws .

The argumentation can easily be followed even though content-related inconsistencies or argumentative leaps can occur.

The argumentation is unconvincing and cannot be maintained without the interlocutor's help.

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Follow-up study

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative (2)
Conclusions
Challenges

Rewriting the scale together with the raters

Concrete insertion

"Subjective" insertion

The presentation's structure is consistent and perfectly aligned with the content. The audience has no problem following the presentation.

…

The presentation … Every now and then, the audience may lose track of the presentation.

…

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Follow-up study

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative (2)
Conclusions
Challenges

Rewriting the scale together with the raters

Concrete insertion

"Subjective" insertion

Clarifying insertion

The lexical range in this performance can be noticeably limited, especially when it comes to abstract concepts.

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Follow-up study

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative (2)
Conclusions
Challenges

Rewriting the scale together with the raters

Concrete insertion

"Subjective" insertion

Clarifying insertion

Discriminating insertion

The performance frequently shows noticeable and disruptive pauses.

The performance is ~~largely~~ only partly understandable …

…

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Follow-up study

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative (2)
Conclusions
Challenges

Rewriting the scale together with the raters

Concrete insertion

"Subjective" insertion

Clarifying insertion

Discriminating insertion

Exemplary addition

…in complex structures grammatical flaws may occur even though common grammatical structures (conjugation, inversion, subclause) are mostly correct.

The performance shows mastery of basic grammatical patterns (simple clauses, main word order)

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Results of the qualitative research (2)

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative (2)
Conclusions
Challenges

Focus Group 2:

- Preference: polytomous
- No vagueness reported
- No mention of lengthy rating processes
- Intuitive, flexible

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



What can we conclude from all this?

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Two rating scales, unlike in reliability

After first pilot, polytomous scale:

considered too abstract

less reliable than dichotomous

↘ *'vagueness in task specification inevitably leads to vagueness in measurement'*

(Bachman 2002: 458)

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



What can we conclude from all this?

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Novice raters reported benefits of

Using more tangible terminology

Assisting in rating scale refinement

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



What can we conclude from all this?

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

Dichotomous scale

- Statistical robustness
- Limitations regarding authenticity & validity

Polytomous scale

- After revising: improved usability & representativeness
- Preferred for assessing speaking

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Future challenges

CNaVT
Rating concerns
Research path
Scale design
Quantitative
Qualitative
Conclusions
Challenges

- Quantitative analysis of speaking scale
- Refine writing scale
- The role of the rater
- Rating scales and validity reduction

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL



Thank you

Bart.Deygers@arts.kuleuven.be

CNaVT
CERTIFICAAT NEDERLANDS als VREEMDE TAAL