KU LEUVEN CTO

CENTRUM VOOR TAAL & ONDERWIJS

Naar een valide, betrouwbare en eerlijke NT2-toets

NETWERKMOMENT NT2-TEST, 20 april 2023

Wie zijn we?

KU LEUVEN CTO
CENTRUM VOOR TAAL & ONDERWIJS

Goedele Vandommele

Lies Strobbe

Annelies Jehoul

Mariet Schiepers

KU LEUVEN CTO
CENTRUM VOOR TAAL & ONDERWIJS

Wetenschappelijk onderbouwde, gestandaardiseerde toets

Valide

- Meten we wat we willen en moeten meten?
- Op welke manier en voor welke specifieke doelen worden de toetsresultaten gebruikt?
- Worden toetsresultaten van een individuele taalleerder correct geïnterpreteerd?
- Weerspiegelt de toets wat taalleerders moeten kunnen in de doeltaal in een levensechte situatie?
- …

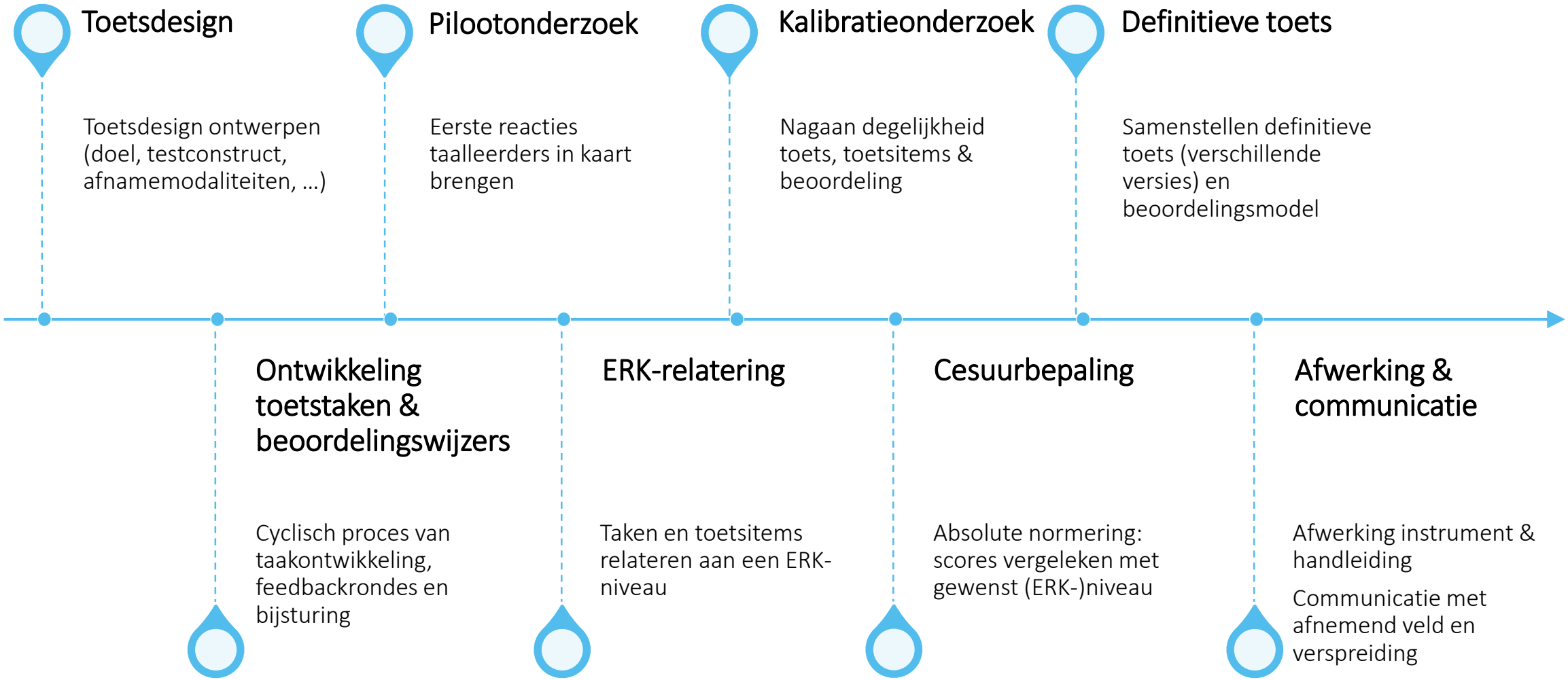
Betrouwbaar

- Meet de toets consistent?
- Werkt de toets voor alle doelgroepen in alle centra, alle regio's en voor alle docenten?
- Zou de kandidaat eenzelfde resultaat krijgen op een ander moment, voor een andere versie van de toets?
- …

Eerlijk & rechtvaardig

- Eerlijke gevolgen voor belanghebbenden en sociale waarde van toetsen
- Is de toets eerlijk en rechtvaardig voor alle taalleerders?
- Liggen de gevolgen in lijn met de doelstellingen van de toets?
- Aandacht voor positieve en negatieve nevenaffecten, vb. washback
- …

Stappen naar een wetenschappelijk onderbouwde toets



Als de laatste fase valide moet zijn, moet elke voorafgaande fase van het ontwikkelingsproces goed doorlopen worden en een bevredigend resultaat opleveren

Bij afwijkingen, risico op

- Problemen met validiteit
- Problemen met betrouwbaarheid
- Problemen met eerlijk gebruik

Problemen met validiteit (‘construct’ | ‘ecological’ | ‘face’)

- Construct validiteit
  - Oorzaken: meten voorkennis, aandachtspanne, digitale geletterdheid, … eerder dan taalcompetenties
  - Gevolgen: geen accuraat beeld taalcompetenties en ongegronde uitspraken
- Verschillende versies van een toets representeren niet hetzelfde construct.
  - Oorzaken: geen duidelijke afbakening testconstruct, geen behoefteanalyse, ongebalanceerde toetssamenstelling (type taken, verwerkingsniveaus, etc.)
  - Gevolgen: geen accuraat beeld taalcompetenties, ongegronde uitspraken, oneerlijke toetsen.
- Ecologische validiteit | face validiteit
  - Oorzaken: geen duidelijke afbakening testconstruct, geen behoefteanalyse, onderwijsveld onvoldoende betrekken en ondersteunen in feedbackrondes, piloot-en kalibratieonderzoek
  - Gevolgen: geen goede weerspiegeling van de context waarin mensen Nederlands leren| geen herkenbare, breed gedragen toets

Problemen met betrouwbaarheid

- Meetfouten en bias
  - Oorzaken: gebruik van niet-gekalibreerde items, geen consistente beoordeling, onvoldoende onderzoek naar verschillende doelgroepen
  - Gevolgen: geen accuraat beeld van taalcompetenties, groepen systematisch bevoor- of benadeeld

- Verschillende versies van toets zijn niet met elkaar gelinkt
  - Oorzaken: gebruik van niet-gekalibreerde items
  - Gevolgen: geen consistente resultaten over regio's, exameninstellingen en tijd heen. Sommige taalleerders krijgen een gemakkelijke toets, anderen een moeilijke toets.

Problemen met eerlijk gebruik (‘use’ | ‘fairness’ | ‘justice’)

- Onjuist gebruik van toetsen
  - Oorzaken: onduidelijkheid over doelgroep, doelstellingen en testconstruct; misalignment tussen doelstellingen toets en onderwijsdoelstellingen, misalignment tussen doelstellingen toets en sociale impact op individuele kandidaten…
  - Gevolgen: conclusies en gevolgen verbonden aan de toets liggen niet in de lijn met het doel van de toets

- Ongewenste effecten
  - Oorzaken: draagvlak ontbreekt, onduidelijke handleidingen, onduidelijke doelstellingen en gevolgen van de toets
  - Gevolgen: teaching to the test (rollen weglaten, enkel meerkeuze vragen inoefenen,…)

Stappen naar een wetenschappelijk onderbouwde toets

Piloot

- Doel: eerste reacties in kaart brengen
- Deelnemers: een kleinere groep taalleerders en testafnemers
- Opzet:
  - Taalleerders maken een eerste versie van de toets
  - Kwantitatieve en kwalitatieve analyses geven eerste indicaties van de validiteit van de toets, de geschiktheid voor verschillende doelgroepen, afnamemodaliteiten, …
- Na deze fase: herwerking van toetsmatrijs, toetsspecificaties, taken, beoordelingsmodellen, handleiding
  - Minimale herwerkingen → ERK-relatering en kalibratieonderzoek
  - Grotere herwerkingen → opnieuw kleinschalige piloot

ERK-relatering

- Doel: een toets op het correcte ERK-niveau
  - Input
  - Verwerking
  - Output
  - Beoordeling
- Deelnemers: ERK-experten uit toets- en onderwijsveld, niet betrokken in de toetsontwikkeling
- Opzet:
  - Familiarisatie ERK-schalen en geijkte taken
  - Linken van toetsitems en taken aan descriptoren uit ERK 2001 & Companion volume 2022
  - Individuele voorbereiding en gezamenlijke beslissing
- Na deze fase: eventuele herwerkingen of schrappen van taken & items

Kalibratieonderzoek

- Doel: bepalen van de schaal van de toets en het inschatten van de moeilijkheidsgraad van toetsitems
- Deelnemers: een grote groep taalleerders
- Opzet:
  - Taalleerders maken afgewerkte toetstaken.
  - Psychometrische analyses, o.a. moeilijkheidsgraad toetsitems, spreiding van moeilijkheidsgraad, fit van de toetsitems, discriminerende waarde van toetsitems, betrouwbaarheids- en validiteitsanalyses, biasonderzoek (DIF), samenhang inschatting docenten – prestaties op de toets

Cesuurbepaling

- Doel: bepalen waar de grens tussen slagen en niet slagen ligt
- Deelnemers: ERK-experten uit toets- en onderwijsveld, niet betrokken in de toetsontwikkeling
- Opzet
  - Absolute normering: de scores van het kalibratieonderzoek worden vergeleken met het gewenste ERK-niveau
  - Verschillende methodes mogelijk, vb. aantal voorbeeldprestaties per deciel vergelijken met het gewenste ERK-niveau

Naar een wetenschappelijk onderbouwd NT2-toetskader

Stappen naar een wetenschappelijk onderbouwde toets (2): rolverdeling

Toetsdesign

Examencommissie:

- Bepalen doel van de toets, definitie doelgroep, behoefte-analyse, testconstruct, afnamemodaliteiten, toetsmatrijs

CTO:

- Feedback op toetsdesign

Ontwikkeling toetstaken & beoordelingswijzers

Examencommissie & toetsontwikkelaars:

- Ontwikkelen en bijsturen toetstaken

CTO:

- Feedback op toetstaken
- Inhoudelijke professionaliseringssessies

Pilootonderzoek

Examencommissie & toetsontwikkelaars

- Logistieke organisatie, contactname scholen, …
- Taken beoordelen

CTO:

- Testinstrumenten, protocollen, dataverzameling & -analyse

ERK-relatering

Examencommissie & toetsontwikkelaars:

- Logistieke organisatie
- Bijsturen taken en toetsitems

CTO:

- Design ERK-relatering, dataverzameling &-analyse

Kalibratieonderzoek

Examencommissie & toetsontwikkelaars:

- Steekproeftrekking en contactname scholen
- Taken beoordelen

CTO:

- Testinstrumenten, dataverzameling & data-analyse

Cesuurbepaling

Examencommissie & toetsontwikkelaars:

- Logistieke organisatie

CTO:

- Design cesuurbepaling, dataverzameling &-analyse

Definitieve toets

Examencommissie & toetsontwikkelaars:

- Feedback op definitieve samenstelling

CTO:

- Samenstelling toets (verschillende versies) volgens diverse parameters (o.a. moeilijkheidsgraad, onderscheidend vermogen, type items…)

Afwerking & communicatie

Examencommissie & toetsontwikkelaars

- Afwerking instrument & handleiding
- Communicatie met afnemend veld en verspreiding

CTO:

- Feedback

Onderzoeksvragen

- Algemeen
  - In welke mate voldoet de toetsbatterij aan de randvoorwaarden voor **valide en betrouwbaar** toetsen? Hoe kan de kwaliteit nog verhoogd worden?

- Afnamemodaliteiten
  - In welke mate beïnvloedt de **afnamemodaliteit** (papier versus digitaal) de testresultaten?
  - In welke mate beïnvloedt de **decentrale afname** standaardisatie van afname en beoordeling? Hoe kan de kwaliteit nog verhoogd worden?

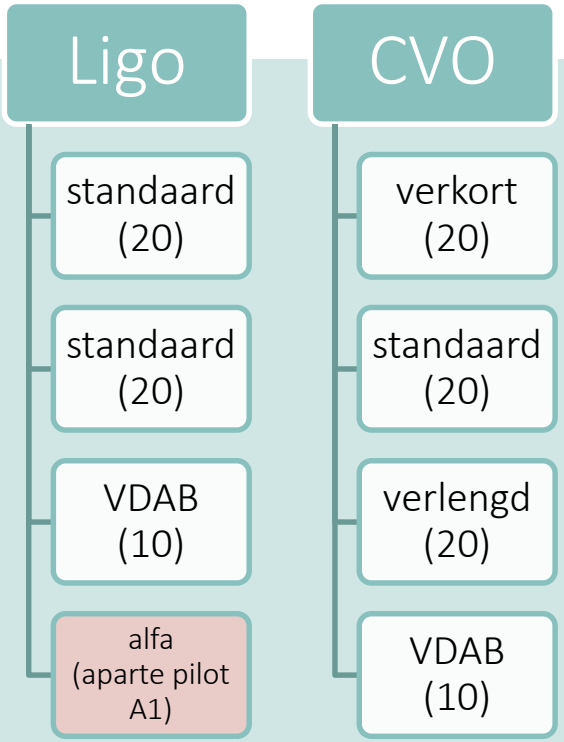
- Heterogene doelgroep
  - Is het mogelijk om valide en betrouwbaar te toetsen en hierbij rekening te houden met **behoeftes van individuele leerders**? En zo ja, hoe kan dit geconcretiseerd worden?
  - Is de testbatterij valide en betrouwbaar voor **laaggeletterde leerders**? En zo ja, welke ingrepen verhogen voor deze doelgroep de betrouwbaarheid?

- Samenhang andere evaluaties
  - Is er samenhang tussen de resultaten uit **permanente evaluatie** in het onderwijs en uit de testbatterij? En zo ja, hoe kunnen beide bronnen van informatie ingezet worden om een volledig beeld te krijgen van de taalcompetentie van een cursist?

Pilootstudie

Design

Doelgroepen & participanten



Binnen doelgroep variatie in

- Digitale vaardigheden
- Taalniveau

Kwantitatieve test

- 120-tal taalleerders
- Op locatie in eigen centrum

Kwalitatieve test

- 25-tal taalleerders (mix van de doelgroepen)
- Op locatie in eigen centrum
- Minimum 1 observatie per doelgroep door CTO, andere observaties door toetsafnemer uit ontwikkelteam (a.d.h.v. protocol & training)

Dataverzameling & analyse

Unieke identificatie van leerders:

- Traject
- Leerkracht
- Centrum
- Toetsafnemer
- Datum van afname
- Scholingsgraad
- Afnamemodaliteit (papier/digitaal)
- Digitale vaardigheid (inschatting leerkracht)
- Taalniveau (inschatting leerkracht)

Analyse van

- Toetsprestaties, met scores op itemniveau
- Input uit observaties, interviews en focusgesprekken

Wat wordt getest?

8 clusters van vragen, gebaseerd op de toets volgens de toetsmatrijs

NT2-test

1. Lezen: informatief
2. Lezen: prescriptief
3. Lezen: narratief
4. Lezen: persuasief

5. Schrijven: info geven & vragen
6. Schrijven: info geven & vragen
7. Schrijven: beschrijving geven

→ Cluster 1-6: 6 taken
→ Cluster 7-8: 7 taken

+ beoordelings-wijzers

+ handleiding

Alle taken werden door 30-tal taalleerders getest, verdeeld over de verschillende doelgroepen, zowel op papier, op laptop als op tablet.

- Verdeling:
  - 2 clusters per klasgroep
  - Binnen cluster: de helft start op papier, de andere helft digitaal (laptop/tablet). Halverwege wordt er van modaliteit gewisseld.

- Ook andere documenten bij toets worden getest:
  - Beoordelingswijzers
  - Handleiding toetsafname

Keuze voor overlappend, cross-over, counterbalanced design

Overlappend

- Wat? Elke taak wordt in verschillende clusters opgenomen, en wordt zowel digitaal als op papier uitgetest.
- Waarom? Om items aan elkaar te relateren. Zo moet één cursist niet alle taken testen en krijg je zicht op hoe afnames zich ten opzichte van elkaar verhouden (verschillende doelgroepen).

Cross-over

- Wat? elke persoon maakt een gedeelte digitaal én een gedeelte op papier
- Waarom? Zo sluiten we individuele effecten uit en kan je met een kleine groep het effect van de modaliteit nagaan

Counterbalanced:

- Wat? sommige cursisten starten op papier en anderen starten digitaal
- Waarom? Effecten van volgorde, vermoeidheid, aandachtspanne, gewenning, etc. uitsluiten)

Waarom kozen we voor dit design?

- Heel efficiënt: minimum aan kandidaten - maximale informatie
- Veel **parameters**:
  - Verschillende types taalleerders (doelgroepen)
  - Twee afnamemodaliteiten
  - Verschillende types digitale toestellen, etc
  - Verschillende schrijftaken voor Ligo en CVO
  - Verschillende types taken (meerkeuze meerdere antwoorden, meerkeuze één antwoord, vergrootglas, juist/fout, open veld)
  - Verschillende tekstlengte, teksttypes en doelstellingen leesteksten
  - …
  - Pilootafname moet gebeuren binnen de lestijden in Ligo en CVO: mag geen te grote ruimte binnen onderwijstijd innemen
- Werd succesvol toegepast binnen andere projecten: Koala, VDAB screening laaggeletterdheid, pilootonderzoek centrale toetsen

Protocol kwalitatieve & kwantitatieve piloot

1 Vóór de pilootafnames

- CTO, Examencommissie & toetsontwikkelaars bereiden de piloot samen voor:
  - CTO: opzet piloot & testinstrumenten
  - Examencommissie: praktische organisatie & planning
  - Toetsontwikkelaars: ontwikkelen toetstaken, beoordelingswijzers & handleidingen
- Centra:
  - Doorlopen van protocol (wat moet je voor, tijdens en na de piloot doen?) (samen met CTO & EC)
  - Lokaal reserveren (voldoende ruimte per cursist)
  - Reserveren van PC's, laptops of tablets + installeren AssessmentQ

2 Vóór of tijdens de pilootafnames

- Leerkrachten: inschatting cursisten maken
- Examencommissie: achtergrondgegevens pilootkandidaten verzamelen

3a Tijdens de pilootafnames (kwantitatief)

- De test duurt 1,5 uur (incl. inleiding & afsluiting).
- Halverwege wisselen de deelnemers van modaliteit
- De toetsafnemers volgen de handleiding bij de toetsbatterij
- Extra maatregelen piloot:
  - Aanvang toets: geruststellen, korte kadering piloot
  - Cursisten toestemmingsformulier laten ondertekenen
  - Invullen beperkt observatieschema
  - Afsluiten: bedanken voor deelname

3b Tijdens de pilootafnames (kwalitatief): extra taken

- Aanvang toets: cursisten stimuleren om vragen te stellen & onduidelijkheden te verwoorden
- Invullen uitgebreider observatieschema
- Achteraf: kort interview met cursist

4 Na de pilootafnames

- Focusgroepgesprek met toetsafnemers
- Toetsontwikkelaars: verbeteren taken
- Examencommissie: extractie data uit AssessmentQ
- CTO: data-analyse

Onderzoeksmethoden

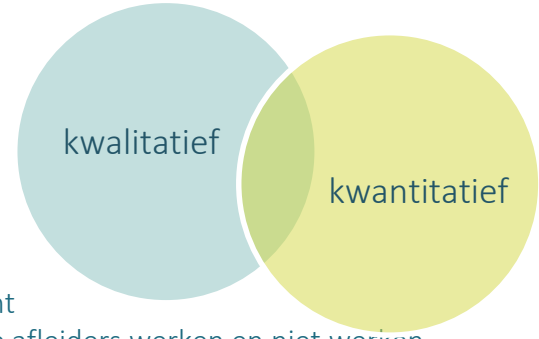
1. Observaties
   - Participerende observatie: cursist wordt aangemoedigd om opmerkingen te maken en vragen te stellen
   - Observatieschema: tijdsbesteding, inhoud (instructies, taalniveau, …), digitale vaardigheden
   - Kwantitatieve test: kort observatieschema (praktische organisatie, reacties cursisten, ervaring toetsafnemers)

2. Inschatting cursisten door leerkrachten
   - Inschatting digitale vaardigheden en taalniveau
   - Drie niveaus: groen - geel - rood
   - Ranking o.b.v. taalniveau

3. Interview met cursisten
   - Mondeling: kort interview met neerslag in formulier
   - Focus op o.a. begrip, aansluiting bij behoeftes, authenticiteit, digitale vaardigheden, …
   - Door toetsafnemer, met behulp van interviewleidraad

4. Focusgesprek voor toetsafnemers
   - In 1-2 groepen, begeleid door CTO
   - Focus op inhoud test & praktische randvoorwaarden

Wat willen we te weten komen?



Onderzoeksvragen	Focus pilootstudie
• **Algemeen** • In welke mate voldoet de toetsbatterij aan de randvoorwaarden voor **valide en betrouwbaar** toetsen? Hoe kan de kwaliteit nog verhoogd worden?	• Indicatie betrouwbaarheid screeningsinstrument • Grove indicatie van welke vragen werken, welke afleiders werken en niet werken • Overeenstemming inschatting docenten – prestaties op toets

• **Afnamemodaliteiten**
  • In welke mate beïnvloedt de **afnamemodaliteit** (papier versus digitaal) de testresultaten?
  • In welke mate beïnvloedt de **decentrale afname** standaardisatie van afname en beoordeling? Hoe kan de kwaliteit nog verhoogd worden?
	• Invloed van de modaliteit (papier of digitaal, laptop of tablet)
	• Drempels bij de (digitale/papieren) afname
	• Noodzakelijke condities voor de afname (locatie, tijd, voorzieningen, …)
	• Richtlijnen voor de toetsafnemers

• **Heterogene doelgroep**
  • Is het mogelijk om valide en betrouwbaar te toetsen en hierbij rekening te houden met **behoeftes van individuele leerders**? En zo ja, hoe kan dit geconcretiseerd worden?
  • Is de testbatterij valide en betrouwbaar voor **laaggeletterde leerders**? En zo ja, welke ingrepen verhogen voor deze doelgroep de betrouwbaarheid?
	• Duidelijkheid van de instructies voor de verschillende doelgroepen
	• Haalbaarheid van toetsing bij verschillende doelgroepen
	• Indicatie van benodigde tijd voor de verschillende doelgroepen
	• Herkenbaarheid voor verschillende doelgroepen

• **Samenhang andere evaluaties** • Is er samenhang tussen de resultaten uit **permanente evaluatie** in het onderwijs en uit de testbatterij? En zo ja, hoe kunnen beide bronnen van informatie ingezet worden om een volledig beeld te krijgen van de taalcompetentie van een cursist?	• Overeenstemming resultaat pilottest & inschatting van de leerkracht

! Nog geen psychometrische analyses mogelijk: te kleine steekproef, veel verschillende doelgroepen

En het doel…

Een valide, betrouwbare en eerlijke NT2-test voor de NT2-cursist