TNO-rapport

TNO-DV 2007 C053

Onderzoek naar de kwaliteit van het inburgeringexamen buitenland

Datum	11 april 2007
Auteur(s)	dr. ir. J.M. Kessens, drs. G. Jacobusse
Opdrachtgever	Ministerie van Justitie
Projectnummer	013.75144/01.02
Rubricering rapport	Vertrouwelijk
Titel	Onderzoek naar de kwaliteit van het inburgeringexamen buitenland
Samenvatting	4
Rapporttekst	41
Bijlagen	23
Aantal pagina's	64 (incl. bijlagen)
Aantal bijlagen	8

Lijst van afkortingen en termen

Afkorting	**Betekenis**
CEF	Common European Framework
KNS	Kennis Nederlandse Samenleving
TGN	Toets Gesproken Nederlands
WIB	Wet Inburgering Buitenland
DIF	Differential Item Functioning
FACETS	multi-facets Rasch-analyse FACETS-software [refA]

Term	**Betekenis**
subscript ‘FACETS’	duidt de FACETS-schaal aan
subscript ‘TGN’	duidt de TGN-schaal aan
subscript ‘i’	i=zinsbouw, woordenschat, uitspraak en vloeiendheid
CEF-grenswaarde	vaardigheid bij de overgang naar een hoger CEF-niveau
CEF-grensscore	TGN score bij de overgang naar een hoger CEF-niveau
$score_{TGN,i}$	$deelscore_i$ op de rapportageschaal van de TGN
$theta_{FACETS}$	taalvaardigheidschatting op de FACETS-schaal
$theta_{TGN,i}$	ongetransformeerde $vaardigheidschatting_i$ op de TGN-schaal

Voorwoord

In dit rapport beschrijft TNO een onderzoek naar de kwaliteit van het inburgeringexamen in het buitenland. Dit inburgeringexamen bestaat uit twee onderdelen: de Toets Gesproken Nederlands (TGN) en de toets Kennis van de Nederlands Samenleving (KNS). Het onderzoek bestaat uit de beantwoording van twee onderzoeksvragen:

1. *Zijn er substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert en menselijke beoordelingen?*
2. *Is de zak-/slaaggrens van de TGN op het goede niveau ingesteld?*

Het rapport is zodanig opgezet dat het op verschillende manieren leesbaar is:

- De snelste manier om het rapport te lezen, is het lezen van de korte (blz. 5) of uitgebreide samenvatting (blz. 6 t/m 8).
- Door het lezen van het rapport zonder appendices wordt een completer beeld van het onderzoek verkregen. Aan het einde van het hoofdstuk waar de onderzoeksvragen worden beschreven, wordt per onderzoeksvraag een korte samenvatting gegeven in de grijze kaders.
- Door het lezen van het rapport inclusief appendices wordt het meest complete beeld van het onderzoek verkregen.

Samenvatting kort

Onderzoeksvraag 1: Zijn er substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert en menselijke beoordelingen?

Resultaten
- Voor zowel de Toets Gesproken Nederlands (TGN) als de toets Kennis van de Nederlandse Samenleving (KNS) zijn er geen substantiële verschillen, er wordt voldaan aan de voorafgestelde kwaliteitsnorm.

Conclusie
- In deze steekproef van het inburgeringexamen buitenland zijn er geen substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert en menselijke beoordelingen.

Consequentie
- Het is voor de TGN en KNS acceptabel om de menselijke beoordeling te vervangen door automatische beoordeling.

Onderzoeksvraag 2: Is de zak-/slaaggrens van de TGN op het goede niveau ingesteld?

Voor het bepalen van de schaal zijn twee datasets gebruikt. De methode om deze datasets gezamenlijk te analyseren heeft TNO op twee punten verbeterd.

Onderzoeksvraag 2 is daarom opgesplitst in twee deelvragen:
a) Vinden we dezelfde zak-/slaaggrens met de twee verbeteringen?
b) Wordt dezelfde schaal gevonden voor de data uit het binnen- vs. buitenland?

Resultaten vraag 2a
- Met de verbeteringen vinden we dezelfde afstanden tussen de zak-/slaaggrenzen.
- De gehele schaal is echter verankerd volgens een criterium dat ruim een vijfde van de schaal soepeler is dan beoogd.

Resultaten vraag 2b
- De zak-/slaaggrens geschat op basis van navertelde verhaaltjes lijkt in het buitenland lager te liggen dan in het binnenland.

Conclusie
- De zak-/slaaggrens is ingesteld volgens een criterium dat ruim een vijfde van de schaal soepeler is dan beoogd.

Consequentie
- Bijstelling van de huidige instelling van de zak-/slaaggrens dient overwogen te worden.

Samenvatting uitgebreid

1. Inleiding

Het inburgeringexamen buitenland is door een consortium bestaande uit het CINOP[1] , LTS[2] en Ordinate[3] in opdracht van het Ministerie van Justitie ontwikkeld. Er is gekozen voor een geautomatiseerd toetssysteem dat gebaseerd is op de Phonepass-technologie ontwikkeld door Ordinate. De geautomatiseerde toets bestaat uit twee onderdelen:

- De 'Toets Gesproken Nederlands (TGN)' die mondelinge taalvaardigheid toetst. Aan de hand van de antwoorden op vragen die gesteld worden bepaalt de computer of de mondelinge taalvaardigheid van een kandidaat voldoende is. Voor het inburgeringexamen Buitenland wordt een minimaal taalvaardigheidniveau 'A1min' geëist, de eis voor het examen Binnenland is 'A2'.
- De 'toets Kennis van de Nederlandse Samenleving (KNS)' meet de kennis die een examenkandidaat heeft over een vooraf gedefinieerde verzameling aspecten van de Nederlandse samenleving. Op basis van antwoorden die een kandidaat geeft op een dertigtal vragen beoordeelt de computer of een examenkandidaat voldoende kennis heeft over de Nederlandse samenleving. De minimale eis voor de KNS is een score van 70% correct.

TNO heeft in een eerder onderzoek een 'second opinion' gegeven ten aanzien van de validatie van de spraaktechnologie die wordt gebruikt bij het inburgeringexamen. TNO concludeerde dat er geen bewijs gevonden is dat de TGN voldoende precies meet wat de toets moet meten, aangezien kwaliteitsmaten ontbreken. Om deze reden beval TNO aan om voor de TGN en de KNS voldoende representatieve data te verzamelen, succescriteria (normen) voor deze toetsen te formuleren, en op basis van deze data de toetsen te hervalideren.

De aanbevelingen van TNO zijn overgenomen en het inburgeringexamen is in de praktijk getest, waarvan deze rapportage de weergave is. Het onderzoek betreft de eerste 500 examenkandidaten die het inburgeringexamen buitenland hebben afgelegd. Om te voorkomen dat kandidaten werden benadeeld, kregen kandidaten tijdens de duur van het onderzoek een herbeoordeling van menselijke beoordelaars.

Het onderzoek bestaat uit het beantwoorden van twee onderzoeksvragen:

1. *Zijn er substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert en menselijke beoordelingen?*
2. *Is de zak-/slaaggrens van de TGN op het goede niveau ingesteld?*

De kwaliteitseisen waaraan de toets moest voldoen zijn vooraf vastgelegd in een onderzoeksprotocol dat opgesteld is door TNO.

---

1 Onderdeel van CINOP Advies b.v., 's Hertogenbosch, Nederland, www.cinop.nl
2 Language Testing Services, Velp, Nederland
3 Ordinate Corporation, Menlo Park, California, USA, www.ordinate.com

2. Onderzoeksvraag 1

Zijn er substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert en menselijke beoordelingen?

Om te onderzoeken of er verschillen zijn tussen menselijke en machinale scoring van de toets voeren mens en machine exact dezelfde taak uit. Voor de TGN gaat dit om een inhoudelijke en kwalitatieve beoordeling van de antwoorden op de toetsitems, bij de KNS gaat het alleen om inhoudelijke beoordeling. De scores verkregen op itemniveau worden voor mens en machine op exact dezelfde wijze gecombineerd tot één totaalscore. Tenslotte wordt de correlatie bepaald tussen de menselijke en machinale totaalscores. Het kwaliteitscriterium dat gesteld wordt is dat de Pearson correlatie minstens 0.90 is.

Resultaten

- De ongecorrigeerde correlatiecoëfficiënt voor de TGN en KNS is resp. 0.90 en 0.91. Zowel de TGN als de KNS voldoet dus aan de vooraf gestelde kwaliteitsnorm (Pearson correlatie ≥ 0.90).

Conclusie

- Er zijn geen substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert en de menselijke beoordelingen.

Consequentie

- Het is voor de TGN en KNS acceptabel om de menselijke beoordeling te vervangen door automatische beoordeling.

Aanbevelingen n.a.v. onderzoeksvraag 1

Voor een substantieel deel (12%) van de kandidaten kon geen automatische toetsuitslag verkregen worden[4]. Achterhaal de oorza(a)k(en) hiervoor en probeer deze weg te nemen/reduceren.

Voor de steekproef van kandidaten is de verdeling van toetsscores zodanig dat er relatief weinig kandidaten onder de zak-/slaaggrens zitten. Monitor daarom het percentage[5] kandidaten met een score onder de zak-/slaaggrens. Mocht een substantieel deel (≥10%) van de kandidaten een score onder de zak-/slaaggrens hebben, dan verdient het aanbeveling om verschillen tussen menselijke en machinale scoring nogmaals te onderzoeken.

[4] wegens technische problemen of de aanwezigheid van ruis/achtergrondlawaai
[5] bijvoorbeeld per maand of kwartaal

3. Onderzoeksvraag 2

Is de zak-/slaaggrens van de TGN op het goede niveau ingesteld?

Voor het bepalen van de schaal zijn twee datasets gebruikt. De methode om deze twee datasets gezamenlijk te analyseren heeft TNO op twee punten verbeterd:

De koppelingsmethode is op een correcte manier uitgevoerd en er is een andere aanname gemaakt over de data (gelijkheidsassumptie voor items).

Verwijder de ongekoppelde interviewdata, zodat er geen ongekoppelde subsets zijn binnen de tweede dataset.

Onderzoeksvraag 2 is om deze reden in twee deelvragen opgesplitst:

a) Vinden we dezelfde zak-/slaaggrens met de twee verbeteringen?
b) Wordt dezelfde schaal gevonden voor de data verzameld in het binnen- vs. buitenland?

Resultaten vraag 2a

- We vinden dezelfde afstanden tussen de CEF-grenswaarden. Alleen voor het B2-niveau vinden we een klein verschil.
- De gehele schaal is verankerd volgens een criterium dat ruim een vijfde van de schaal soepeler is dan beoogd.

Resultaten vraag 2b

- De A1min-zak-/slaaggrens geschat op basis van navertelde verhaaltjes lijkt in het buitenland lager te liggen dan in het binnenland.

Conclusies

- De A1min-zak-/slaaggrens is ingesteld volgens een criterium dat ruim een vijfde van de schaal soepeler is dan beoogd.

Consequentie

- Bijstelling van de huidige instelling van de zak-/slaaggrens dient overwogen te worden. Dit kan bereikt worden door de schaal te verankeren t.o.v. een strenger criterium (bijvoorbeeld gemiddelde itemmoeilijkheid). De omvang van de bijstelling dient op grond van inhoudelijke en beleidsmatige argumenten gekozen te worden.

Inhoudsopgave

1 Inleiding

De uitgangspunten van het huidige inburgeringbeleid zijn vastgelegd in het Hoofdlijnenakkoord van het kabinet Balkenende II (16 mei 2003). Wie zich duurzaam in Nederland wil vestigen moet zich de Nederlandse taal eigen maken en dient actief deel te nemen aan de Nederlandse samenleving. Om in aanmerking te komen voor een Machtiging tot Voorlopig Verblijf (MVV) dienen nieuwkomers[6] in het eigen land een examen af te leggen, het *inburgeringexamen Buitenland*. Om in aanmerking te komen voor een vergunning voor permanent verblijf moet in Nederland nog een keer een examen gedaan worden, het *inburgingexamen Binnenland*.

In december 2003 kreeg een consortium bestaande uit het CINOP[7] , LTS[8] en Ordinate[9] van het Ministerie van Justitie de opdracht om een examen te ontwikkelen voor toetsing van mondelinge taalvaardigheid in het Nederlands. Deze toets was in eerste instantie bedoeld voor het inburgeringexamen Buitenland, maar werd ook ontwikkeld om ingezet te kunnen worden voor het examen Inburgering Binnenland. Vanwege de specifieke randvoorwaarden die aan de toets gesteld worden heeft het Ministerie van Justitie gekozen voor een geautomatiseerd toetssysteem dat gebaseerd is op de Phonepass-technologie ontwikkeld door Ordinate. De 'Toets Gesproken Nederlands (TGN, zie [refB])' die deze mondelinge taalvaardigheid toetst wordt afgenomen via de telefoon. De examenkandidaat moet op de juiste manier reageren op de vragen die de computer stelt. De antwoorden worden door een speciaal ontwikkeld automatisch scoringssysteem, dat bekend is onder de naam Phonepass, verwerkt. Met dit systeem wordt bepaald of de mondelinge taalvaardigheid van een examenkandidaat voldoet aan de gestelde minimumeisen. Voor het inburgeringexamen Buitenland is door de Commissie Franssen een advies uitgebracht over het minimale taalvaardigheidniveau [refC] dat gesteld zal worden aan de TGN. Een nieuw taalvaardigheidniveau 'A1min' is gedefinieerd, dat zich bevindt onder het laagste niveau (A1) beschreven in het 'Common European Framework' (CEF).

In het voorjaar van 2004 werd besloten dat Kennis van de Nederlandse Samenleving (KNS, zie [refD]) ook onderdeel werd van het inburgeringexamen Buitenland. Hiervoor werd ook een geautomatiseerde toets ontwikkeld die gebaseerd is op de Phonepass-technologie. De 'toets Kennis van de Nederlandse Samenleving (KNS)' meet de kennis die een examenkandidaat heeft over een vooraf gedefinieerde verzameling aspecten van de Nederlandse samenleving. De toets KNS bestaat uit vragen die aan de kandidaat gesteld worden. Op basis van de automatisch herkende antwoorden bepaalt het systeem of een examenkandidaat voldoet aan de norm van 70% correcte antwoorden.

Het ministerie van Justitie heeft in mei 2005 aan TNO gevraagd om een 'second opinion' te geven ten aanzien van de validatie van de spraaktechnologie die wordt gebruikt bij het inburgeringexamen. Deze 'second opinion' is gebaseerd op resultaten van validatiestudies (niet door TNO uitgevoerd) die aan TNO ter inzage zijn gegeven. De resultaten van de 'second opinion' zijn in oktober 2005 vastgelegd in een rapport [refE]. TNO concludeert dat er geen bewijs gevonden is dat de TGN voldoende precies meet wat de toets moet meten, noch dat de toets dit in ónvoldoende mate doet.

---

[6] geldt voor de nieuwkomers die onder de de doelgroepen van de Wet Inburgering vallen
[7] Onderdeel van CINOP Advies b.v., 's Hertogenbosch, Nederland, www.cinop.nl
[8] Language Testing Services, Velp, Nederland
[9] Ordinate Corporation, Menlo Park, California, USA, www.ordinate.com

Of de kwaliteit van de gebruikte spraaktechnologie voldoende is kon niet vastgesteld worden omdat geen kwaliteitsnormen gesteld zijn. Om deze reden heeft TNO aanbevolen om voor de TGN en toets KNS voldoende representatieve data te verzamelen, succescriteria (normen) voor deze toetsen te formuleren, en op basis van deze data de toetsen te hervalideren.

In november 2005 sprak minister Verdonk met de tweede kamer af om de aanbevelingen van TNO en het consortium over te nemen en de toetsen in de praktijk te testen. Dit praktijkonderzoek werd als een vervolgonderzoek op de 'second opinion' door TNO uitgevoerd en betrof de eerste 500 examenkandidaten die het inburgeringexamen buitenland hebben aflegt. Tijdens het onderzoek werden alle examenkandidaten naast de automatische toetsuitslag herbeoordeeld door vier menselijke beoordelaars. Hangende de uitkomst van het onderzoek - om te voorkomen dat kandidaten worden benadeeld - kan een gezakte kandidaat alsnog geslaagd zijn als de menselijke score hoger is dan de machine score. De herbeoordelingprocedure is opgezet en uitgevoerd door het consortium. De herbeoordeling zal voortduren tot het moment dat dit TNO-rapport in de Tweede Kamer besproken zal zijn. Op 15 maart 2006 is de Wet Inburgering Buitenland (WIB) in werking getreden en is begonnen met de dataverzameling.

Zoals beschreven in de brief aan de Tweede Kamer [refF], bestaat het onderzoek uit het beantwoorden van twee onderzoeksvragen die in overleg met Justitie en het consortium zijn gedefinieerd:

Onderzoeksvraag 1
Zijn er substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert voor de TGN en voor de Toets KNS en menselijke beoordelingen?

De eerste vraag van het onderzoek is erop gericht om te achterhalen of er verschillen bestaan tussen de beoordeling door de computer en door de menselijke beoordelaars. Indien die verschillen niet substantieel zijn, betekent dit dat de automatische scoring de kandidaten niet bevoordeelt of benadeelt, en behoeft het scoresysteem niet te worden aangepast. Als die verschillen wél substantieel zijn, zal bekeken worden hoe die verschillen geminimaliseerd kunnen worden.

Onderzoeksvraag 2
Is de zak-/slaaggrens van de TGN op het goede niveau ingesteld?

Voor de tweede vraag in het onderzoek wordt tevens de zak-/slaaggrens zoals die thans is vastgesteld, met data die in het praktijkonderzoek worden verzameld, geconfronteerd. Het doel hiervan is om vast te stellen of met deze praktijkdata eenzelfde cesuurinstelling wordt gevonden als met de data die in de eerdere analyse zijn gebruikt. Mochten de resultaten van het onderzoek hiertoe reden geven, dan zal bezien worden of en op welke wijze tot een aanpassing van de cesuurinstelling kan worden besloten.

In een eerste fase van het onderzoek zijn de twee bovenstaande onderzoeksvragen nader uitgewerkt. Hierbij is de onderzoeksmethodiek nauwkeurig beschreven en zijn vooraf kwaliteitsnormen opgesteld. Het onderzoeksprotocol is opgezet in onderling overleg tussen TNO, het consortium en PMI. Om de kwaliteit en onafhankelijkheid van TNO extra te waarborgen heeft TNO zich laten adviseren door een externe expert [refG].

De uitgebreide uitwerking van het onderzoeksprotocol is aan dit rapport toegevoegd als bijlage A.

De tweede fase bestond uit het uitvoeren van het onderzoek. Hiervan wordt in dit rapport verslag gedaan. De afkortingen en de terminologie die in dit rapport gebruikt worden, zijn op de tweede pagina samengevat.

2 Verantwoording

Voor de opzet van het onderzoek is TNO verantwoordelijk. Voor de uitvoering van het onderzoek is TNO echter afhankelijk geweest van andere partijen. Om praktische redenen is het consortium verantwoordelijk geweest voor de dataverzameling. Daarnaast betrof een deel van het onderzoek bedrijfsgevoelige informatie. Ter bescherming van de intellectuele eigendomsrechten van Phonepass-technologie is een aantal delen van het onderzoek uitgevoerd door Ordinate. In onderstaande tabellen is aangegeven welke partij verantwoordelijk was voor welk onderdeel van het onderzoek.

Tabel 1 Verantwoordelijke per onderdeel van onderzoeksvraag 1 voor de TGN.

Training en selectie van transscribenten voor het maken van woordelijke transcripties	CINOP/LTS
Training en selectie van beoordelaars van vloeiendheid- en uitspraakoordelen	CINOP/LTS
Omzetting van woordelijke transcripties naar ruwe inhoudsscores	Ordinate
Omzetting van ruwe deelscores naar menselijke totaalscore	TNO
Berekening van correlatiecoëfficiënt + grafische weergave van de data	Ordinate
Controle berekeningen en interpretatie	TNO

Tabel 2 Verantwoordelijke per onderdeel van onderzoeksvraag 1 voor de KNS.

Training en selectie van transscribenten voor het maken van woordelijke transcripties	CINOP/LTS
Omzetting van woordelijke transcripties naar ruwe inhoudsscores	Ordinate
Berekening van correlatiecoëfficiënt	Ordinate
Controle berekeningen en interpretatie resultaten	TNO

Tabel 3 Verantwoordelijke per onderdeel van onderzoeksvraag 2a.

Aanleveren van oorspronkelijke dataset voor herhaling FACETS-analyse	LTS
Herhalen van oorspronkelijke analyse met verbeteringen uit vooronderzoek	TNO
Transformatie FACETS-grenswaarden naar rapportageschaal	LTS
Interpretatie resultaten	TNO

Tabel 4 Verantwoordelijke per onderdeel van onderzoeksvraag 2b.

Training en selectie van beoordelaars CEF-oordelen verhaaltjes navertellen	CINOP/LTS
Facets-analyse nieuwe dataset	TNO
Berekening percentage kandidaten per CEF-categorie	LTS
Transformatie FACETS-grenswaarden naar rapportageschaal	LTS
Interpretatie resultaten	TNO

3 Onderzoeksvraag 1

Onderzoeksvraag 1

Zijn er substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert voor de TGN en voor de Toets KNS en menselijke beoordelingen?

Het automatische scoringssysteem van de TGN en de toets KNS rapporteert een totaalscore die voor de TGN ligt tussen 10 en 80 en voor de KNS tussen 0% en 100%. Deze totaalscores vormen respectievelijk een schatting van de mondelinge taalvaardigheid van een kandidaat en van diens kennis over (een gedefinieerde verzameling aspecten van) de Nederlandse samenleving. Op basis van deze totaalscore en de zak-/slaaggrens wordt de toetsuitslag van een kandidaat bepaald: een kandidaat is gezakt of geslaagd. Het automatische scoringssysteem is zó getraind dat de automatische scoring zo goed mogelijk de menselijke scoring benadert. Dit impliceert dat het menselijke oordeel het criterium is waartegen een computerbeoordeling gemeten wordt. Er zal rekening gehouden worden met het feit dat dit criterium niet perfect betrouwbaar is.

Het doel van dit onderzoek is om te achterhalen of er verschillen bestaan tussen de beoordeling door de computer en het criterium (de menselijke beoordelaars).
Indien die verschillen klein zijn (kleiner dan de vooraf vastgestelde norm), betekent dit dat de automatische scoring de kandidaten niet bevoor- of benadeelt ten opzichte van het menselijke oordeel, en hoeft het scoringssysteem niet te worden aangepast. Als die verschillen wél substantieel zijn (groter dan de norm), kan door PMI besloten worden tot een apart onderzoek, waarin de oorzaken van de mogelijke verschillen achterhaald worden en zal bekeken worden hoe die verschillen geminimaliseerd kunnen worden.

Voor de dataverzameling en de training en selectie van de beoordelaars en transscribenten was het consortium verantwoordelijk. Voor de opzet en uitvoer van het onderzoek was TNO verantwoordelijk.

3.1 Dataverzameling

3.1.1 Praktijkdata betrokken in het onderzoek

De praktijkdata bestaan uit de set van toetsafnames waarvan zijn uitgesloten:

1. toetsafnames van ambassadepersoneel (gebruikt voor training van het ambassadepersoneel);
2. toetsafnames waarbij het automatische systeem geen toetsuitslag heeft gegeven wegens technische problemen of de aanwezigheid van ruis/achtergrondlawaai;
3. onvolledige toetsafnames.

Na uitsluiting van bovengenoemde toetsafnames ontstaat een dataset van 500 kandidaten waarvan zowel de TGN als de KNS is afgenomen in de praktijk, namelijk op ambassades in het buitenland. Het onderzoek is zo opgezet dat de eerste helft van de praktijkdata gebruikt is om de validiteit van het scoringssysteem vast te stellen en de tweede (onafhankelijke) helft van de praktijkdata (kandidaten 251-500) gebruikt kan worden om het effect van eventuele hertraining te kunnen valideren.

Om volledige toetsafnames van in totaal 250 kandidaten te verkrijgen was het nodig om voor 283 kandidaten de toetsantwoorden op te nemen.. Het percentage kandidaten dat niet in het onderzoek betrokken kon worden vanwege uitsluitingscriterium 2 was dus 12%

3.1.2 Type beoordelingen gemaakt voor het onderzoek

Om het automatische scoringssysteem te valideren voeren zowel mens als machine dezelfde taak uit op exact dezelfde taaluitingen van de kandidaten. De machine genereert zowel scores voor taalkwaliteit als woordelijke transcripties. Om deze scores en transcripties te kunnen evalueren moesten de menselijke beoordelaars ze onafhankelijk herhalen:

- Woordelijke transcripties
  - Voor de 'kort-antwoord' items en de 'tegenstelling'-items van de TGN (22) en voor alle items van de KNS (30 items) wordt op basis van de woordelijke transcriptie en het antwoordmodel een dichotome score verkregen; namelijk 0=fout, of 1=goed.
  - Voor de 'zinnen herhalen'-items van de TGN (23) wordt op basis van de woordelijke transcriptie een polytome score verkregen; variërend van 0 tot het maximaal aantal correct herhaalde woorden in de herhaalde zin.
- Scores voor taalkwaliteit
  - Voor de 'zinnen herhalen'-items van de TGN (23) worden CEF-beoordelingen gemaakt voor uitspraak en vloeiendheid. Dit zijn polytome scores lopend van 0 (≤ A1min) t/m 7 (= C2).

In tabel 7 zijn de typen beoordelingen en de daaruit afgeleide typen scores samengevat. Er wordt gewerkt met een pool van beoordelaars (meer dan vier). Om halo-effecten te voorkomen wordt iedere kandidaat-respons aan 4 willekeurig uit de pool getrokken beoordelaars voorgelegd. Voor de inhoudelijke en kwalitatieve beoordelingen bestond de pool uit respectievelijk 11 en 12 beoordelaars.

Tabel 5 Typen beoordelingen gebruikt in het onderzoek.

Toets	Aspect	Type score	menselijk oordeel per item	aantal beoordeelde items per kandidaat[10]
TGN	woordenschat	dichotoom: 0/1[11]	woordelijke transcriptie	22
	zinsbouw	polytoom: 0-max.	woordelijke transcriptie	23
	vloeiendheid	polytoom: 0-7	geheel getal tussen 0-7	23
	uitspraak	polytoom: 0-7	geheel getal tussen 0-7	23
KNS	kennis	dichotoom: 0/1[11]	woordelijke transcriptie	30

3.1.3 Training en selectie van de beoordelaars en transscribenten

Training en selectie van de beoordelaars en transscribenten werd uitgevoerd door het consortium. TNO is bij de training aanwezig geweest. Om inzicht te kunnen krijgen in de beoordelings- en transcriptieprocedure heeft TNO een aantal beoordelingen en transcripties gemaakt. Deze beoordelingen zijn niet in het onderzoek opgenomen.

[10] Het eerste item van iedere nieuwe opgavensoort wordt niet in de score meegenomen
[11] Deze scores worden automatisch afgeleid op basis van de door mensen vervaardigde woordelijke transcripties

Voor de inhoudelijke TGN items en voor de KNS zijn woordelijke transcripties van de opgenomen responsen gemaakt. Hierbij is gebruik gemaakt van een transcriptieprotocol opgesteld door Ordinate, zie bijlage B. Beluisteren en beoordelen vond plaats via internet waar de audio kon worden afgespeeld en beoordeeld. Waarborging van de kwaliteit van de transcripties is gerealiseerd door aan het einde van de training de transcribenten transcripties te laten maken van 50 willekeurig uit de buitenlanddata getrokken responsen (afkomstig van verschillende kandidaten) die verder niet in het hoofdonderzoek zijn betrokken. De transcripties zijn geëvalueerd op grond van vergelijking met modeltranscripties. Deze modeltranscripties zijn opgesteld door een expert. Als norm werd gesteld dat 90% van het aantal getranscribeerde woorden over de totale set van 50 responsen moesten overeenstemmen met het model.
Uitsluitend transcribenten die aan de norm voldeden zijn in het hoofdonderzoek betrokken. In totaal hebben 13 personen deelgenomen aan de training. Bij de eerste evaluatie zijn 8 transcribenten afgewezen. Bij de tweede evaluatie zijn 7 transcribenten alsnog geslaagd en is 1 transcribent definitief afgewezen. De gemiddelde overeenstemming met de modeltranscripties van de 12 geslaagde transcribenten was 90,9%.

Voor de kwalitatieve TGN items zijn beoordelingen gegeven over de kwaliteit van de spraak op de CEF-schaal. Hierbij is gebruik gemaakt van CEF-descriptoren voor uitspraak en vloeiendheid, zie bijlage D en E. Beoordeling vond plaats door in te bellen op een speciaal telefoonnummer, zie bijlage F. Waarborging van de kwalitieve beoordelingen is gerealiseerd door aan het einde van de training beoordelaars kwalitatieve beoordelingen te laten maken over 50 willekeurig uit de buitenlanddata getrokken responsen (afkomstig van verschillende kandidaten) die verder niet in het hoofdonderzoek zijn betrokken. Alle beoordelaars kregen dezelfde set van 50 responsen voorgelegd. De beoordelaarsovereenstemming is geschat met een intraklasse-correlatiecoëfficiënt [refH]. Als norm wordt gesteld dat de ondergrens van het 90% betrouwbaarheidsinterval [refI, refJ] voor de geschatte beoordelaarsovereenstemming bij gebruikmaking van vier willekeurige beoordelaars hoger is dan 0,90.
In totaal hebben 11 personen deelgenomen aan de training. Bij de eerste evaluatie zijn 3 beoordelaars afgewezen. Bij de tweede evaluatie zijn de eerder afgewezen beoordelaars allemaal geslaagd.

3.2 TGN

3.2.1 Methode TGN

Het automatische scoringssysteem van de TGN rapporteert een totaalscore tussen 10 en 80 die een beoordeling vormt van de mondelinge taalvaardigheid van een kandidaat. De validiteit van het automatische scoringssysteem wordt bepaald ten opzichte van menselijke beoordelaars die eveneens de vaardigheid van de kandidaten schatten (respectievelijk de mondelinge taalvaardigheid of kennis van de Nederlandse samenleving). De vaardigheidschattingen worden gemaakt door iedere deelvaardigheid een multi-facets Rasch-analyse uit te voeren met drie facets:

- facet 1 = de kandidaat;
- facet 2 = het item;
- facet 3 = de beoordelaar.

Voor de multi-facets Rasch-analyse is gebruik gemaakt van het programma Facets (v. 3.61.0) [refA]. Elke respons is altijd door vier verschillende beoordelaars beoordeeld. De multi-facets Rasch-analyse houdt bij de schatting van de kandidaatsvaardigheden rekening met zowel item-moeilijkheid als beoordelaarsstrengheid. Het menselijke

criterium is de kandidaatsvaardigheid die met de Rasch-analyse uit de beoordelingen wordt geschat. Voor de TGN worden vier verschillende deelvaardigheden geschat:

Tabel 6 Deelvaardigheden van de TGN.

i	Deelvaardigheid
1	Woordenschat
2	Zinsbouw
3	Vloeiendheid
4	Uitspraak

Om te garanderen dat de deelscores in de menselijke totaalscore in dezelfde verhouding meewegen als bij de machine, worden de vaardigheidsschattingen zo genormeerd dat de verhouding van standaarddeviaties van deelvaardigheden overeenkomt met die van de machinale subscores:

$$theta_i^* = theta_i \frac{A_{machine}}{A_{mens}} \tag{1a}$$

$$A = \frac{stdev_i}{stdev_1} \tag{1b}$$

waarbij:

$theta_i$ = de deelvaardigheid i (i=1, 2, 3, 4) en stdev = de standaarddeviatie

Vervolgens worden de deelvaardigheden volgens een ongewogen gemiddelde tot een menselijke totaalscore gecombineerd:

$$score_{mens} = \frac{\sum_{i=1}^{4} theta_i^*}{4} \tag{2}$$

Aangezien de machinale scores van de TGN worden ingeperkt tussen 10 en 80, moeten de menselijke scores voor de TGN ook ingeperkt worden. Om te schatten wat de afkappunten zijn voor de menselijke scores, wordt een lineaire regressie uitgevoerd. Op basis van de regressiefunktie wordt bepaald wat de menselijke afkappunten zijn. Vervolgens worden de menselijke scores onder/boven de afkappunten vervangen door de menselijke scores bij het afkappunt. Op deze manier worden de menselijke en machinale scores op exact dezelfde manier afgekapt.

De correlatiecoëfficiënt (Pearson's Product Moment Correlation) is als volgt berekend:

$$\rho_{x,y} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} \tag{3}$$

De berekening van de menselijke totaalscores is uitgevoerd door TNO. Ordinate heeft vervolgens de scores ingeperkt en de correlatiecoëfficiënt berekend.

3.2.2 Resultaten TGN

In Tabel 7 staan de aantallen kandidaten, items, beoordelaars en beoordelingen per deelvaardigheid.

Tabel 7 Aantallen kandidaten, items, beoordelaars en beoordelingen per deelvaardigheid.

	Woordenschat	Zinsbouw	Vloeiendheid	Uitspraak
kandidaten	295	295	295	295
items	388	493[1]	494	494
beoordelaars	12	12	11	11
beoordelingen	22000	22956	19448[2]	23000
verwijderd	-	44[1]	-	-

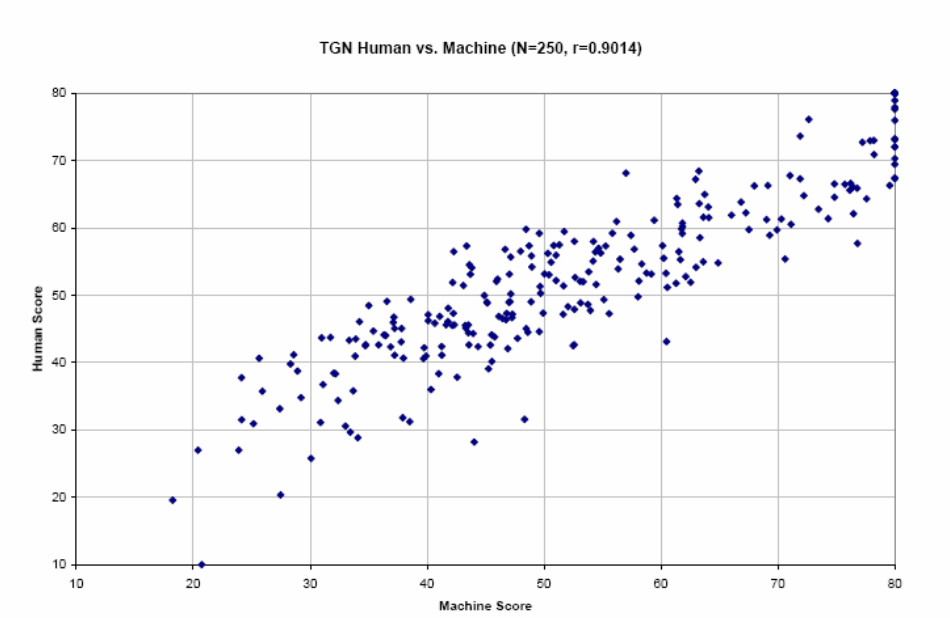
[1] er is één item verwijderd aangezien hiervoor maar één antwoordcategorie was gescoord
[2] niet voor alle kandidaten werden vier beoordelingen verkregen

Voor iedere deelvaardigheid is een aparte multi-facets Rasch-analyse uitgevoerd met het programma Facets [refA]. Uit Facets volgen schattingen van de kandidaat-vaardigheden. Deze kandidaatvaardigheden worden met formule (1) en (2) genormeerd en gecombineerd tot een menselijke totaalscore. Vervolgens worden de menselijke totaalscores op dezelfde manier als de machinale scores ingeperkt (zie paragraaf 2.1).

De ongecorrigeerde correlatiecoëfficiënt tussen de menselijke en machine scores is 0,9014. Zoals vermeld in het onderzoeksprotocol (zie appendix A) dient de correlatiecoëfficiënt gecorrigeerd te worden voor de onbetrouwbaarheid van het menselijke criterium. Om deze correctie te kunnen toepassen dient de test-hertest betrouwbaarheid van de menselijke beoordelaars geschat te worden. In de verzamelde dataset zijn echter te weinig test-hertest oordelen aanwezig om de test-hertest betrouwbaarheid nauwkeurig te schatten. Er kan wel een schatting gemaakt worden voor de ondergrens van de betrouwbaarheid[12]. Aangezien de ongecorrigeerde correlatiecoëfficiënt al voldoet aan de gestelde kwaliteitsnorm van $r \geq 0{,}90$, zal dit ook gelden voor de gecorrigeerde correlatiecoëfficiënt, die altijd hoger is dan de ongecorrigeerde coëfficiënt. Het uitvoeren van de correctie is daarom nagelaten.

In figuur 1 zijn de ingeperkte menselijke totaalscores (‘Human scores’) en de score die het automatische scoringssysteem heeft gegenereerd (‘Machine scores’) weergegeven.

[12] Dit kan door middel van een split-half methode. Hiertoe worden de items van elke kandidaat in twee sets gesplitst, set A en set B. De correlatie tussen A en B is dan de ondergrens voor de betrouwbaarheid van het criterium (bij halve testlengte, te corrigeren door middel van de Spearman Brown formule). Nadeel van het gebruik van de split-half betrouwbaarheid voor het berekenen van attenuatiecorrectie is dat hiermee de betrouwbaarheid wordt onderschat en er dus wordt overgecorrigeerd.



Figuur 1 Relatie tussen menselijke scores en machine scores voor de TGN.

3.3 KNS

3.3.1 Methode KNS

Het automatische scoringssysteem van de TGN rapporteert voor een kandidaat een totaalscore tussen 0 en 100% die een schatting vormt van zijn/haar kennis over (een gedefinieerde verzameling aspecten van) de Nederlandse samenleving. Er bestaan tien sets van 30 KNS-vragen die zo zijn samengesteld dat de moeilijkheid per set ongeveer gelijk is. Aangezien de moeilijkheid van de items (sets van 10 KNS-vragen) overeenkomstig zijn hoeft dit niet als facet in de analyse opgenomen te worden. De totaalscore is het percentage correcte antwoorden.

De menselijke beoordelaars voerden exact dezelfde taak uit als de machine, zij dienden een woordelijke transcriptie te maken van de antwoorden die kandidaten gaven. Validiteit van het automatische scoringssysteem wordt bepaald ten opzichte van een menselijk criterium dat gebaseerd is op de meerderheidtranscriptie van de vier beoordelaars (twee of meer beoordelaars zijn het eens over deze transcriptie). Indien de stemmen staken wordt willekeurig één van de vier transcripties als meerderheids-transcriptie gekozen. Vergelijking met een meerderheidstranscriptie is alleen zinvol als er een grote mate van overeenstemming is tussen de beoordelaars. Uit analyse van de transcripties blijkt dat dit inderdaad het geval is, zie Tabel 8

Tabel 8 Mate van overeenstemming.

overeenstemming	vier	drie	twee	Geen
percentage (%)	98.6	0.9	0.4	0.1

Op basis van de transcripties van een groot aantal antwoorden van kandidaten op de KNS vragen is in een eerder onderzoek een antwoordmodel gemaakt, zie [refD]. Dit antwoordmodel specificeert wanneer een antwoord ‘correct’ is. Voor het bepalen van de

correctheid van het antwoord wordt voor zowel de mens als de machine van dit antwoordmodel gebruik gemaakt. Op deze manier wordt voor ieder antwoord een score verkregen: ‘correct’ of ‘incorrect’. De menselijke score is - net als de machine score - het percentage inhoudelijk correcte antwoorden.

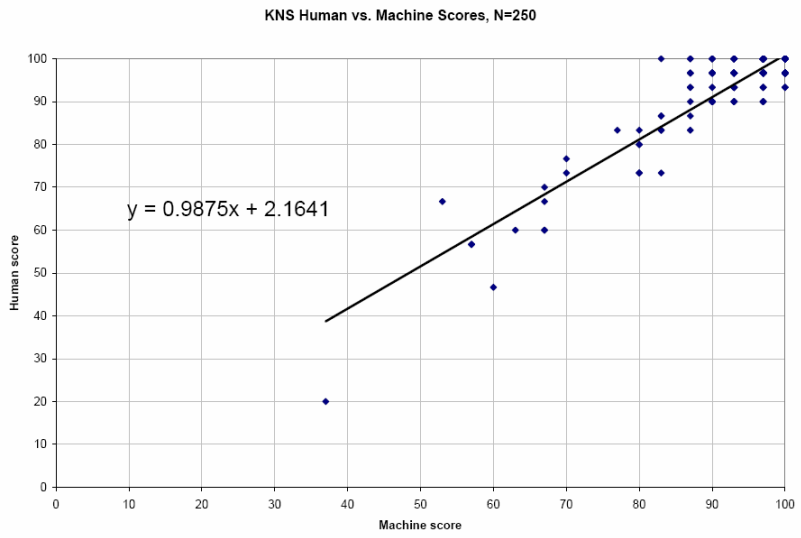
De correlatiecoëfficiënt tussen de menselijke en machinale scores wordt op dezelfde manier berekend als voor de TGN (zie formule 3). De berekening van deze correlatiecoëfficiënt is uitgevoerd door Ordinate.

3.3.2 Resultaten KNS

In totaal werden de antwoorden van 250 kandidaten woordelijk getranscribeerd. Ieder antwoord wordt door vier verschillende beoordelaars getranscribeerd. In totaal zijn dus (250x30x4=) 30000 transcripties gemaakt en worden op basis hiervan 7500 correct/incorrect scores bepaald.

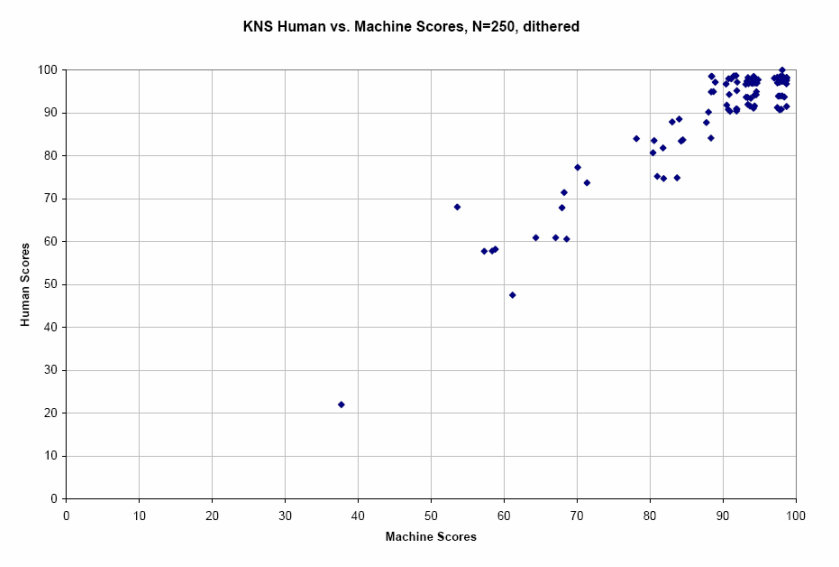
De ongecorrigeerde correlatiecoëfficiënt tussen de menselijke en machine scores is 0,914. De correctie voor onbetrouwbaarheid van het criterium wordt ook voor de KNS niet berekend, aangezien de ongecorrigeerde correlatiecoëfficiënt al voldoet aan de gestelde kwaliteitsnorm van $r \geq 0{,}90$.

In figuur 2a staan de menselijke en de machine scores weergegeven. De scores in figuur 2a liggen op een vast afstand van 1 item (3,3%). De verdeling van de scores is beter te zien in figuur 2b. In figuur 2b wordt het aantal scores per punt in de plot grafisch weergegeven door de scores een willekeurige kleine verschuiving[13] te geven in beide richtingen (‘dithering’).



Figuur 2a Relatie tussen menselijke (human) en machine KNS scores.

[13] Een willekeurig getrokken getal tussen -2% and +2% van de menselijke en machine score.



Figuur 2b Relatie tussen 'dithered' menselijke (human) en machine KNS scores.

In figuur 2a is ook de geschatte lineaire regressiefunctie weergegeven. Volgens deze regressiefunctie komt een menselijke score van 70% overeen met een machine score van 69%. Vertaald naar itemniveau betekent dit dat voor zowel de mens als de machine dezelfde zak-/slaaggrens van 21 items wordt gevonden.

3.4 Conclusie

Zowel het automatische scoringssysteem van de TGN als die van de KNS voldoen aan de vooraf gestelde kwaliteitsnorm. Dit betekent dat het acceptabel is om de menselijke beoordelaars door de computer te vervangen.

3.5 Overweging bij de conclusie

Een correlatiecoëfficiënt geeft een indicatie over de kwaliteit van het scoringssysteem over de gehele schaal. Een nadeel van een correlatiecoëfficiënt is dat deze afhankelijk is van de verdeling van de scores. De KNS scores zijn niet normaal verdeeld over het scorebereik. In het lage scoregebied zitten aanzienlijk minder kandidaten dan in het hoge scoregebied. Als er helemaal geen scores in het lage scoregebied zouden hebben gezeten, zou de correlatiecoëfficiënt, vanwege een restriction-of-range effect, lager zijn geweest. Als er meer kandidaten in het lage scoregebied zouden hebben gezeten, zou de correlatie hoger geweest kunnen zijn maar we weten niet hoeveel hoger.

Voor beide toetsen is uiteindelijk van belang of de verschillen tussen machinale en menselijke scoring tot een andere zak-/slaagbeslissing zouden leiden. In de huidige steekproef van kandidaten is het aantal gezakte kandidaten evenwel zeer klein. Voor de TGN zijn er helemaal geen gezakte kandidaten, alle kandidaten zijn volgens zowel menselijke beoordelaars als de machine geslaagd. Voor de KNS is 2,5% van de kandidaten zowel volgens de menselijke beoordelaars als de machine gezakt. In Tabel 9 staan de aantallen gezakte en geslaagde kandidaten voor de KNS volgens de mens en de machine. Voor slechts 1 van de 250 kandidaten wordt er door de KNS een andere toetsuitslag gegeven dan de menselijke toetsuitslag. Voor deze steekproef van kandidaten functioneert de machinale beoordeling van de KNS dus goed.

Tabel 9 Aantallen gezakte en geslaagde kandidaten voor de KNS volgens de menselijke beoordelaars (mens) en de machine.

		mens	
		geslaagd	gezakt
machine	geslaagd	239	0
	gezakt	1	10

Aangezien er in deze steekproef weinig kandidaten zijn met een score rond de zak-/slaaggrens, is het van belang dit percentage te monitoren. Als de verdeling namelijk substantieel gaat afwijken van de huidige steekproef is het noodzakelijk om een nader onderzoek te doen naar het functioneren van het automatische scoringssysteem rond de zak-/slaaggrens.

3.6 Aanbevelingen n.a.v. onderzoeksvraag 1

De kandidaten waarvoor geen machinale score kon worden verkregen zijn niet in dit onderzoek betrokken. TNO beveelt daarom aan om te onderzoeken wat de oorza(a)k(en) zijn waarom deze scores niet verkregen konden worden. Vervolgens dienen de oorza(a)k(en) weggenomen of gereduceerd te worden. Mocht het nodig zijn om hiervoor het automatische scoringssysteem aan te passen (bijvoorbeeld door het te hertrainen) dan dient geverifieerd te worden:

- dat nu meer kandidaten een toetsuitslag krijgen;
- dat voor deze kandidaten de correlatie tussen menselijke en machine scores ook hoog is;
- dat de in dit rapport gevonden hoge correlatie in stand blijft voor het aangepaste systeem.

Een tweede aanbeveling betreft hetgeen in de overweging is genoemd. Voor beide toetsen geldt dat in deze steekproef er weinig kandidaten zijn met een score onder de zak-/slaaggrens. Voor beide toetsen is het daarom van belang dit percentage te monitoren. Pas als een aanzienlijk percentage van de kandidaten een score onder de zak-/slaaggrens haalt (≥10%) is verdient het aanbeveling om verschillen tussen machinale en menselijke scores nader te onderzoeken.

3.7 Samenvatting

Onderzoeksvraag 1

Resultaten

- Het automatisch scoringssysteem van de TGN en de toets KNS voldoen aan de vooraf gestelde kwaliteitsnorm.

Conclusie

- Er zijn geen substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert en de menselijke beoordelingen.

Consequentie

- Het is voor de TGN en KNS acceptabel om de menselijke beoordeling te vervangen door automatische beoordeling.

4 Onderzoeksvraag 2

Onderzoeksvraag 2
Is de zak-/slaaggrens van de TGN op het goede niveau ingesteld?

Voor het vaststellen van de zak-/slaaggrens gebruikt TNO dezelfde methode die door het consortium gebruikt is. Hierdoor zijn de nieuwe resultaten direct te vergelijken met de eerder verkregen resultaten. In een vooronderzoek heeft TNO de door het consortium gebruikte methode nauwkeurig bestudeerd. Aangezien in het TGN-eindrapport [refB] de uitgevoerde methodologie niet volledig is beschreven was hiervoor overleg met het consortium noodzakelijk. In paragrafen 4.1 wordt de schalingsmethode uitvoering beschreven.

Uit het vooronderzoek blijkt dat er twee verbeteringen mogelijk zijn in de methode die gebruikt is om de verschillende datasets gezamenlijk te analyseren. Onderzoeksvraag 2 is daarom in twee deelvragen opgesplitst:

Onderzoeksvraag 2a

Vinden we dezelfde zak-/slaaggrens als we de schalingsmethode herhalen met de twee verbeteringen die uit het vooronderzoek volgen?

In een vooronderzoek constateert TNO dat een tweetal verbeteringen mogelijk is t.a.v. het gezamenlijk analyseren van de twee gebruikte datasets. Onderzocht zal worden of dezelfde zak-/slaaggrens wordt gevonden als deze verbeteringen worden toegepast.

Onderzoeksvraag 2b

Wordt dezelfde schaal gevonden voor de data verzameld in het binnen- vs. buitenland?

Voor de hogere taalvaardigheidniveau's zijn de aantallen groot genoeg om de zak-/slaaggrens met voldoende nauwkeurigheid te bepalen. Voor deze hogere taalvaardigheidniveau's zal daarom met de praktijkdata de schaal opnieuw bepaald worden om zo vast te stellen of er verschillen zijn met de schaal die is geschat op grond van data verzameld in Nederland.

In paragraaf 4.2 en 4.3 worden onderzoeksvraag 2a en 2b beschreven.

4.1 Schalingsmethode

In paragrafen 4.1.1 t/m 4.1.4 wordt de schalingsmethode uitgebreid beschreven. Daarnaast worden in paragraaf 4.1.5 de verbeteringen gegeven die TNO heeft gevonden. Deze verbeteringen betreffen het gezamenlijk analyseren van de gebruikte datasets.

4.1.1 Benodigde data voor het vaststellen van de schaal

De schalingsmethode bestaat uit drie stappen:

1. Schatten van de CEF-grenswaarden

De continue taalvaardigheidschaal is opgedeeld in 8 discrete CEF-niveaus: <A1min, A1min, A1, A2, B1, B2, C1 en C2. Op basis van een groot aantal menselijke beoordelingen van deze CEF-niveaus kan geschat worden waar op de continue taalvaardigheidschaal de CEF-grenswaarden liggen. Dit is gedaan met een multi-facets Rasch-analyse (FACETS). De CEF-grenswaarden op de FACETS-schaal worden aangeduid met CEF-grenswaarden$_{FACETS}$.

2. Bepalen van de regressiefuncties

De FACETS-analyse geeft ook een schatting van de kandidaatvaardigheden op de FACETS-schaal. De TGN doet dit ook, echter op een andere schaal. De vaardigheidschattingen op resp. de TGN- en FACETS-schaal worden theta$_{TGN}$ en theta$_{FACETS}$ genoemd. Door voor een groot aantal kandidaten zowel theta$_{FACETS}$ als theta$_{TGN}$ te schatten, kan de relatie tussen theta$_{FACETS}$ en theta$_{TGN}$ bepaald worden.

Door het uitvoeren van stap 1 en 2 ligt de plaats van de CEF-grenswaarden vast. Op basis van de antwoorden op de toets genereert de TGN een schatting van de taalvaardigheid (theta$_{TGN}$). Deze taalvaardigheidschatting wordt - volgens de relatie gevonden in stap 2 - vertaald naar de FACETS-schaal. Aangezien de CEF-grenswaarden op de FACETS-schaal bekend zijn uit stap 1, kan het CEF-niveau van de kandidaat bepaald worden.

3. Transformatie naar de rapportageschaal

De zak-/slaaggrenzen van de TGN zijn echter gedefinieerd op de rapportageschaal van de TGN, deze schaal loopt van 10 t/m 80. In de derde stap wordt de transformatiefunctie van theta$_{TGN}$ naar een score op de rapportageschaal vastgelegd (score$_{TGN}$).

In Tabel 10 staan de benodigde data vermeld per stap van de schalingsmethode.

Tabel 10 Benodigde data per stap van de schalingsmethode

	Benodigde data	Schattingen van
stap 1: Schatten van de CEF-grenswaarden	menselijke CEF-oordelen voor N kandidaten	8 CEF-grenswaarden$_{FACETS}$
stap 2: Bepalen van de regressiefuncties	yheta$_{FACETS}$ + theta$_{TGN}$ per deelvaardigheid voor N kandidaten	relatie (theta$_{FACETS}$, theta$_{TGN}$)
stap 3: Transformatie naar de rapportageschaal	grensscores van de Spaanse en Engelse versie van de toets + CEF-grenswaarden$_{FACETS}$	relatie (theta$_{FACETS}$, score$_{TGN}$)

4.1.2 Stap 1: Schatten van de CEF-grenswaarden met FACETS-analyse

De CEF-grenswaarden worden geschat op basis van analyse van menselijke beoordelingen van taalvaardigheid. De CEF-oordelen zijn gegeven voor drie typen taaluitingen:

- 'verhaaltjes navertellen': Kandidaten moeten een kort verhaal navertellen. Het CEF-oordeel wordt gegeven op basis van het beluisteren van een audio-opname van het navertelde verhaal.
- 'interview': In een adaptief interview wordt aan de hand van vast interviewprotocol (zie bijlage 11 van [ref G]) het CEF-niveau van een kandidaat bepaald.
- 'open vragen': Kandidaten moeten antwoord geven op een aantal open vragen die aan het einde van de toets gesteld worden. Het CEF-oordeel wordt gegeven door audio-opnamen van deze antwoorden te beoordelen.

De CEF-grenswaarden zijn geschat met een multi-facets Rasch-analyse, hiervoor is het programma FACETS gebruikt [refA]. Bij deze analyse is het uitgangspunt dat het oordeel (de toegekende CEF categorie) voor een taaluiting afhankelijk is van drie facetten:

1. de vaardigheid van de kandidaat,
2. de moeilijkheid van het item en
3. de strengheid van de beoordelaar.

Het model is als volgt gedefinieerd:

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k \qquad (4)$$

waarin:

P_{nijk} = de kans dat kandidaat n op item i van beoordelaar j een oordeel k krijgt
P_{nijk-1} = de kans dat kandidaat n op item i van beoordelaar j een oordeel k-1 krijgt
B_n = de vaardigheid van kandidaat n
D_i = de moeilijkheid van item i
C_j = de strengheid van beoordelaar j
F_k = de moeilijkheid van de stap van categorie k-1 naar categorie k

Om het model te identificeren wordt in FACETS de gemiddelde kandidaatvaardigheid (B_n) en de gemiddelde beoordelaarstrengheid (C_j) gelijk gesteld aan 0. De CEF-grenswaarden (afgeleid van F_k) zijn gecentreerd rond de item moeilijkheid en daarom ook gemiddeld 0.

Op alle items wordt als resultaat een menselijk CEF oordeel gegeven, m.a.w. alle items hebben dezelfde antwoordschaal. Als model is dan ook gekozen voor het 'Rating Scale' model waarin de afstanden tussen de CEF-grenswaarden onderling per definitie voor alle items gelijk zijn.

Bij menselijke oordelen over taaluitingen is het de bedoeling dat alle beoordelaars het eens zijn in hun oordeel over dezelfde kandidaat. Verschillen in strengheid tussen beoordelaars zijn in feite ongewenst. In een ideale situatie zouden alle C_j gelijk zijn aan 0, want de bedoeling is dat CEF-oordelen die door verschillende beoordelaars gegeven zijn, onderling vergelijkbaar zijn.

Eveneens is het de bedoeling om op basis van elk van de items een oordeel te geven dat vergelijkbaar is met het oordeel over dezelfde kandidaat bij andere items. De menselijke beoordelaar doet zijn best om in de beoordeling rekening te houden met de moeilijkheid van de taak. Verschillen in moeilijkheid tussen items hebben daarmee een zelfde interpretatie als verschillen in beoordelaarstrengheid. In een ideale situatie zouden alle D_i gelijk zijn aan 0, want de bedoeling is dat CEF-oordelen die op basis van verschillende items gegeven worden, onderling vergelijkbaar zijn.

Het doel van de FACETS-analyse is om CEF-grenswaarden te vinden die een plaats hebben op de taalvaardigheidschaal van de kandidaten, onafhankelijk van beoordelaarstrengheid of itemmoeilijkheid. De CEF-grenswaarden die uit de FACETS-analyse volgen zijn juist wel relatief aan beoordelaarstrengheid en itemmoeilijkheid. Er

moet een keuze gemaakt worden om de relatieve CEF-grenswaarden uit de FACETS-analyse uit te drukken op een vaste plaats op de kandidaatvaardigheidsschaal. Een logisch verdedigbare keuze is om de CEF-grenswaarden uit te drukken ten opzichte van de gemiddelde beoordelaarstrengheid en itemmoeilijkheid. De schaal is al verankerd ten opzichte van een gemiddelde beoordelaarstrengheid van 0. De gecentraliseerde CEF-grenswaarden moeten echter nog gecorrigeerd worden voor de gemiddelde itemmoeilijkheid.

In het TGN-eindrapport [refB] staat vermeld dat in totaal 5.984 CEF-oordelen zijn gegeven voor 1.008 kandidaten (facet 1), voor 59 typen CEF-beoordelingen (facet 2), en door 13 beoordelaars (facet 3). Niet alle data zijn echter in de analyse meegenomen: 193 kandidaten waren niet gespecificeerd en 17 extreme datapunten werden verwijderd (datapunten met groot residu), resulterend in een totaal van 4401 geldige datapunten.

Uit de FACETS-analyse volgen schattingen van de CEF-grenswaarden op de FACETS-schaal. De eerder gevonden CEF-grenswaarden$_{FACETS}$ staan weergegeven in Tabel 11. Deze zijn overgenomen uit tabel 5.3 van het TGN-eindrapport [refB].

Tabel 11 Eerder gevonden CEF-grenswaarden$_{FACETS}$ bij overgang naar het aangegeven CEF-niveau

	Grenswaarde
A1min	-8,13
A1	-5,96
A2	-2,69
B1	-0,28
B2	2,47
C1	5,7
C2	8,88

4.1.3 Stap 2: Bepalen regressiefuncties

Uit de multi-facets Rasch-analyse volgt voor iedere kandidaat een schatting van de vaardigheid (facet 1), waarbij rekening wordt gehouden met de moeilijkheid van het type CEF-beoordeling (facet 2) en de strengheid van de beoordelaars (facet 3).

In stap 2 wordt de TGN-vaardigheidsschaal aan de FACETS-vaardigheidsschaal gekoppeld. Deze koppeling wordt gelegd voor iedere deelvaardigheid. Er wordt aangenomen dat het om een lineair verband gaat:

$$theta_{FACETS,i} = a_i * theta_{TGN,i} + b_i \quad (4)$$

Hierbij is 'i' een index voor de deelvaardigheden: woordenschat, zinsbouw, uitspraak en vloeiendheid. Met een lineaire regressieanalyse worden de parameters (a_i en b_i) geschat op basis van $theta_{FACETS}$ en $theta_{TGN,i}$ van een groot aantal kandidaten[14].

Nadat de parameters (a_i en b_i) geschat zijn kunnen de schattingen van de vier deelvaardigheden op ieder van de TGN-schalen ($theta_{TGN,i}$) door het toepassen van de regressiefuncties in (4) vertaald worden naar een taalvaardigheidsschatting op de FACETS-schaal ($theta_{FACETS}$)..

[14] De R-kwadraten behorende bij de regressiefuncties zijn: zinsbouw (0.46) en woordenschat (0.47), uitspraak (0.44) en vloeiendheid (0.42)

4.1.4 Stap 3: Transformatie naar de rapportageschaal

Er is gekozen voor een rapportageschaal tussen 10 en 80. Naast de TGN bestaat er ook Engelse (SET-10, zie [refK]) en Spaanse (SST, zie [refL]) versie van de toets. Er is voor gekozen om scores die worden verkregen voor de verschillende toetsen eenzelfde betekenis te geven. Om deze reden wordt de transformatiefunctie van vaardigheid_{CEF} naar de rapportageschaal bepaald op basis van de grensscores[15] van de SET-10 en de SST en de $\text{CEF-grenswaarden}_{FACETS}$ uit Tabel 11:

$$\text{CEF-grensscores}_{SET10/SST} = c * \text{CEF-grenswaarden}_{FACETS,TGN} + d \quad (5)$$

Dit resulteert in de volgende transformatiefunctie:

$$\text{score}_{TGN,i} = c^* \text{theta}_{FACETS,i} + d \quad (6)$$

De TGN totaalscore wordt vervolgens als volgt verkregen:

- Met formule (4) wordt de TGN deelvaardigheidschatting ($\text{theta}_{TGN,i}$) omgezet naar een taalvaardigheidschatting op de FACETS-schaal ($\text{theta}_{FACETS,i}$),
- Met formule (6) wordt $\text{theta}_{FACETS,i}$ omgezet naar een deelscore ($\text{score}_{TGN,i}$),
- De deelscores ($\text{score}_{TGN,i}$) worden in geperkt tussen 10 en 90,
- De TGN totaalscore is de gemiddelde deelscore en wordt gerapporteerd tussen 10 en 80.

De CEF-grenswaarden worden met formule (5) omgezet naar TGN grensscores.

4.1.5 Verbeteringen ten aanzien van het gezamenlijk analyseren van ongekoppelde datasets

Om te komen tot een betrouwbare schatting van de CEF-grenswaarden heeft het consortium gebruik gemaakt van twee datasets die apart van elkaar verzameld zijn. Voor het schatten van de CEF-grenswaarden (stap 1) moest daarom deze twee datasets in één FACETS-analyse geanalyseerd worden. TNO vindt dat het gezamenlijk analyseren van deze datasets op twee punten te verbeteren is. De eerste verbetering heeft ermee te maken dat er voor de analyse twee datasets zijn gebruikt, zie Tabel 12. In de FACETS-analyse is het mogelijk om de twee datasets samen te analyseren, mits de gegevens voor twee van de drie facetten (kandidaat, item en beoordelaar) overlap vertonen. De datasets overlappen niet via facet 1 (kandidaten), aangezien er geen gemeenschappelijke kandidaten zijn. Er is ook geen overlap op facet 2 (items), aangezien in dataset 1 andere itemtypes (verhaaltjes) werden gebruikt dan in dataset 2 (open vragen, interviews).

Het probleem dat beide datasets niet samen te analyseren zijn, is theoretisch gezien onoplosbaar. Alleen door aannames te maken over de data kunnen de datasets samen geanalyseerd worden. Het consortium heeft voor het koppelen van de data gebruik gemaakt van een kunstmatige beoordelaar die 20 CEF-beoordelingen heeft gemaakt (deze CEF-beoordelingen worden op basis van de overige data geschat). Deze methode van het kunstmatig introduceren van data wordt ‘imputatie’ genoemd. De koppelingsmethodiek bleek echter niet correct te zijn uitgevoerd, aangezien er na het toevoegen van de imputaties nog steeds ongekoppelde datasets waren.

[15] Deze scores zijn getransformeerd naar het scorebereik van de TGN

Het is ook mogelijk om de data te koppelen door de aanname te doen dat bepaalde kandidaten, items of beoordelaars in beide datasets identiek zijn. Een aantal van de beoordelaars komt in beide datasets voor. Aangezien deze beoordelaars voor beide datasets de taak hadden om een kandidaat een score te geven in de vorm van een niveau op de CEF-schaal, mag aangenomen worden dat de strengheid van een beoordelaar in beide datasets overeenkomt. Om de datasets te koppelen, is dan nog een aanname nodig in de vorm van een gelijkheidsassumptie voor items of voor kandidaten. TNO kiest voor een gelijkheidsassumptie voor items omdat alle items bedoeld zijn om kandidaten in dezelfde CEF-categorieën in te delen. De assumptie stelt tijdelijk één item uit de dataset 1 gelijk aan één item uit datset 2. Hierdoor ontstaat een link, die het mogelijk maakt om de datasets samen te analyseren. De analyse wordt een aantal keren herhaald, steeds met een andere gelijkheidsassumptie. In iedere analyse zal de schatting van het A1min niveau iets anders zijn. Doel van deze methode is om na te gaan of de schattingen van de zak-/slaaggrenzen gevoelig zijn voor de manier van koppelen.

De tweede verbetering betreft het gebruik van twee verschillende definities van beoordelaar binnen dataset 2. Dataset 2 bestaat uit twee subsets: a) ‘interviews’ en b) ‘open vragen’, zie Tabel 12.

Tabel 12 Overzicht van de gebruikte datasets.

dataset	subset	type taaluiting
1	geen	verhaaltjes navertellen
2	a	interviews
	b	open vragen

Voor subset 2a wordt een groep van beoordelaars die hetzelfde type CEF-beoordeling hebben gegeven als één beoordelaar opgenomen, terwijl het in de rest van de data om individuele beoordelaars gaat. Hierdoor valt de dataset 2 uiteen in twee ongekoppelde subsets. De beste oplossing is om alsnog de individuele beoordelaars te achterhalen. Dit bleek echter praktisch gezien niet mogelijk te zijn. Om deze reden kiest TNO ervoor om de ongekoppelde ‘interview’- data (dataset 2a) uit de analyse te verwijderen en te onderzoeken wat het effect hiervan is op de hoogte van de zak-/slaaggrens.

Samenvatting schalingsmethode

De schalingsmethode is in dit hoofdstuk uitgebreid beschreven.

TNO heeft twee verbeteringen gevonden om de datasets die in de analyse gebruikt zijn samen te analyseren:

1. Voer de koppelingsmethode op een correcte manier uit en maak andere aannames om de twee ongekoppelde datasets (1+2) samen te analyseren
2. Verwijder de ongekoppelde interviewdata, zodat er geen ongekoppelde subsets zijn binnen dataset 2.

4.2 Onderzoeksvraag 2a

Doel van onderzoeksvraag 2a is om te onderzoeken of het maken van andere aannames om de ongekoppelde data samen te analyseren invloed heeft op de schattingen van de van de zak-/slaaggrenzen.

4.2.1 Methode onderzoeksvraag 2a

De FACETS-analyse zal herhaald worden met de oorspronkelijke dataset waaruit subset 2a (de interviewdata) is verwijderd. Daarnaast zullen de gebruikte datasets (1+2) gekoppeld worden door items uit beide datasets tijdelijk aan elkaar gelijk te stellen. In dataset 1 zitten vier verschillende items (open vragen) en in dataset 2 zitten 58 verschillende items (verhaaltjes). Aangezien er geen inhoudelijke reden is om specifieke items uit de verschillende datasets gelijk aan elkaar te stellen, wordt er voor gekozen om ieder van de vier items uit de dataset 1 negen keer te koppelen met een willekeurig gekozen item uit dataset 2. In totaal worden dus 36 analyses uitgevoerd.

Het gebruikte 'Rating Scale' model veronderstelt dat de afstanden tussen de CEF-grenswaarden$_{FACETS}$ onafhankelijk zijn van het specifieke item. Voor alle items gezamenlijk worden de grenzen dan ook maar één keer geschat. Daarom mag verwacht worden dat de afstanden tussen CEF-grenswaarden$_{FACETS}$ in de verschillende analyses gelijk zijn. Als resultaat van de herhaalde analyses zullen een minimum, een maximum, een gemiddelde en een standaarddeviatie van de A1min zak-/slaaggrens geschat worden. Hierdoor ontstaat inzicht in de variatie in de A1min zak-/slaaggrens als gevolg van verschillende gelijkheidsassumpties voor de koppeling.

Vervolgens zullen de nieuwe schattingen van de CEF-grenswaarden$_{FACETS}$ vergeleken worden met de eerder gevonden CEF-grenswaarden$_{FACETS}$. Voordat deze vergelijking gedaan kan worden moeten de schalen geëquivaleerd worden. Hiertoe wordt de verschuiving van de oorsprong geschat voor gemeenschappelijke kandidaten en beoordelaars. De equivalering houdt in dat de oorsprong van de vaardigheidsschaal in alle analyses gelijk wordt gemaakt. Daarnaast drukken we de CEF-grenswaarden$_{FACETS}$ uit op vaste plaatsen op de kandidaatsvaardigheidsschaal. Zoals vermeld in paragraaf 4.1.2 dient er dan gecorrigeerd te worden voor de gemiddelde itemmoeilijkheid. Na equivalering en correctie voor de gemiddelde itemmoeilijkheid, kan bepaald worden:

- of de eerder gevonden A1min-grens zich bevindt in de range van nieuwe schattingen volgens de koppeling obv gelijkheidsassumpties;
- of het absolute verschil tussen de schattingen in de originele analyse en de schattingen volgens de koppeling o.b.v. gelijkheidsassumpties praktisch relevant is (minimaal 2 punten op de TGN rapportageschaal ≅ 0.5 punten op de FACETS-schaal).

Ook is gekeken of de vaardigheidsschattingen voor kandidaten, items en beoordelaars in originele en nieuwe analyses met elkaar overeenkomen en of er sprake is van substantiële Differential Item Functioning (DIF, zie [refJ]) bij verhaaltjes die zowel in het binnenland als in het buitenland gebruikt zijn. Wegens kleine aantallen responsen zijn de DIF resultaten per item slechts voor een klein aantal items beschikbaar. Ook bij deze items hebben de DIF toetsen weinig statistische power. Om deze reden worden de resultaten hier niet gepresenteerd. Tenslotte is onderzocht of het onderscheidend vermogen van de items op de lage niveaus groot genoeg blijft als de interviews uit de dataset worden verwijderd.

In de hierboven beschreven analyse is stap 1 van de schalingsmethode nauwkeuriger bekeken. Voor de instelling van de zak-/slaaggrens is het echter ook van belang op welke manier de CEF-grenswaarden$_{FACETS}$ een vaste plaats hebben gekregen op de kandidaatsvaardigheidsschaal (stap 2+3 van de schalingsmethode). Om deze reden is – in overleg met het consortium - de transformatie van de CEF-grenswaarden$_{FACETS}$ naar de TGN rapportageschaal nader bekeken.

4.2.2 Data onderzoeksvraag 2a

In Tabel 13 staan de statistieken van de datasets die gebruikt zijn voor onderzoeksvraag 2a. De dataset die gebruikt is voor de nieuwe analyses is een subset van de originele dataset.

Tabel 13 Statistieken van de datasets gebruikt voor onderzoeksvraag 2a.

	Kandidaten	Items	beoordelaars	data
Originele analyse	815	59	13	4401
Nieuwe analyse	538	57	9	3375

De nieuwe analyse kent de volgende verschillen met de originele analyse:
- 277 kandidaten minder: Dit zijn de kandidaten die alleen interviewbeoordelingen hebben gekregen.
- 2 items minder: Het interview-item is verwijderd en voor de koppeling worden in iedere analyse telkens twee items aan elkaar gelijk gesteld.
- 4 beoordelaars minder: De kunstmatige beoordelaar en de drie interviewbeoordelaars zijn verwijderd.

4.2.3 Resultaten onderzoeksvraag 2a: afstanden tussen de CEF-grenswaarden$_{FACETS}$

De originele datasets bestond uit twee ongekoppelde datasets:

dataset 1 met alleen items van het type ‘verhaaltjes navertellen’;
dataset 2 met alleen items van het type ‘open vragen’.

Om de datasets toch samen te kunnen analyseren zijn de datasets gekoppeld door items aan elkaar gelijk te stellen. In Tabel 14 staan de items vermeld die in de analyses aan elkaar gelijk gesteld zijn. Alle gekozen ‘verhaaltjes’ zijn gekoppeld met alle ‘open vragen’. In totaal zijn de vier ‘open vragen’ gekoppeld met negen willekeurig gekozen ‘verhaaltjes’, dus in totaal zijn 36 analyses uitgevoerd. De multi-facets Rasch-analyses zijn uitgevoerd met het programma FACETS [refA].

Tabel 14 Items die in de analyses aan elkaar gelijk zijn gesteld.

Itemtype	Open vragen				Verhaaltjes navertellen								
Itemnummer	2	3	5	6	1	21	37	46	66	69	83	92	102
Itemmoeilijkheid[16]	5,0	4,6	3,9	4,6	3,8	4,4	1,7	4,7	3,4	4,2	6,5	6,0	3,2

De gemiddelde, minimale, maximale en standaarddeviatie van de CEF-grenswaarden$_{FACETS}$ voor de herhaalde analyses staan vermeld in Tabel 15. Hierbij betekent ‘A1min’ de grens tussen de niveaus ‘< A1min’ en ‘A1min’.

Tabel 15 Gemiddelde (gem.), minimum (min.), maximum (max.) en standaarddeviatie (s.d.) voor de CEF-grenswaarden$_{FACETS}$ in de herhaalde analyses.

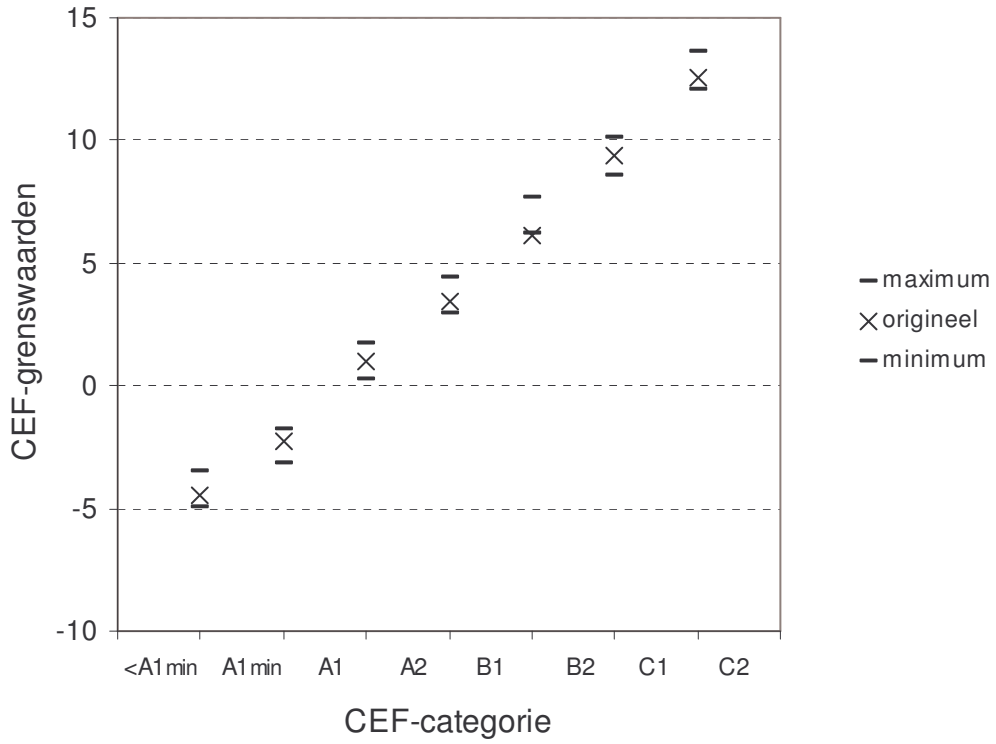
	Gem.	Min.	Max.	S.d.
A1min	-7.98	-8.12	-7.82	0.07
A1	-6.25	-6.37	-6.11	0.07
A2	-2.78	-2.89	-2.68	0.06
B1	-0.12	-0.20	-0.06	0.04

[16] Koppeling via gelijkstellen van ‘verhaaltje 1’ en ‘open vraag 2’

B2	3.07	3.00	3.12	0.03
C1	5.33	5.16	5.49	0.08
C2	8.74	8.51	9.01	0.12

In Figuur 2 staan de afstanden tussen de CEF-grenswaarden$_{FACETS}$ grafisch weergegeven. Voor de nieuwe analyses zijn de minimale en maximale waarden weergegeven. Voor de originele analyse zijn de waarden gebruikt uit tabel 5.3 van het TGN-eindrapport [refB]. De schalen zijn geëquivaleerd en alle CEF-grenswaarden zijn uitgedrukt op de kandidaatvaardigheidsschaal door te corrigeren voor gemiddelde itemmoeilijkheid (zie paragraaf 4.1.2).

In figuur 2 is te zien dat de originele waarde (X) voor de A1min-grens ligt tussen de minimale en maximale waarde van de nieuwe analyses. Voor het B2-niveau (categorie=5) geldt dat de originele afstand afwijkt van de afstanden die in de nieuwe analyses worden gevonden.



Figuur 2 CEF-grenswaarden$_{FACETS}$ gevonden in originele en herhaalde analyses.

Het is ook mogelijk om de afstanden tussen CEF-grenswaarden$_{FACETS}$ te vergelijken zonder te corrigeren voor de gemiddelde itemmoeilijkheid en zonder de schalen te equivaleren. Minimum en maximum liggen in dat geval dichter bij elkaar omdat er geen ruis is door equivalering en door correctie voor de gemiddelde itemmoeilijkheid. Deze vergelijking is weergegeven in Bijlage G, Figuur 6. Ook daar geldt dat de originele A1min-grens niet afwijkt van de A1min-grens die we vinden in de nieuwe analyses.

In appendix G laten we zien dat het onderscheidend vermogen op lage niveaus groot genoeg is. Verder laten we zien dat de vaardigheidsschattingen tussen de originele en nieuwe analyses zeer goed overeenkomen voor de gemeenschappelijke kandidaten, items en beoordelaars.

De resultaten laten dus zien dat de een andere manier van datakoppeling geen gevolgen heeft voor de afstanden tussen de CEF-grenswaarden$_{FACETS}$ die worden gevonden.

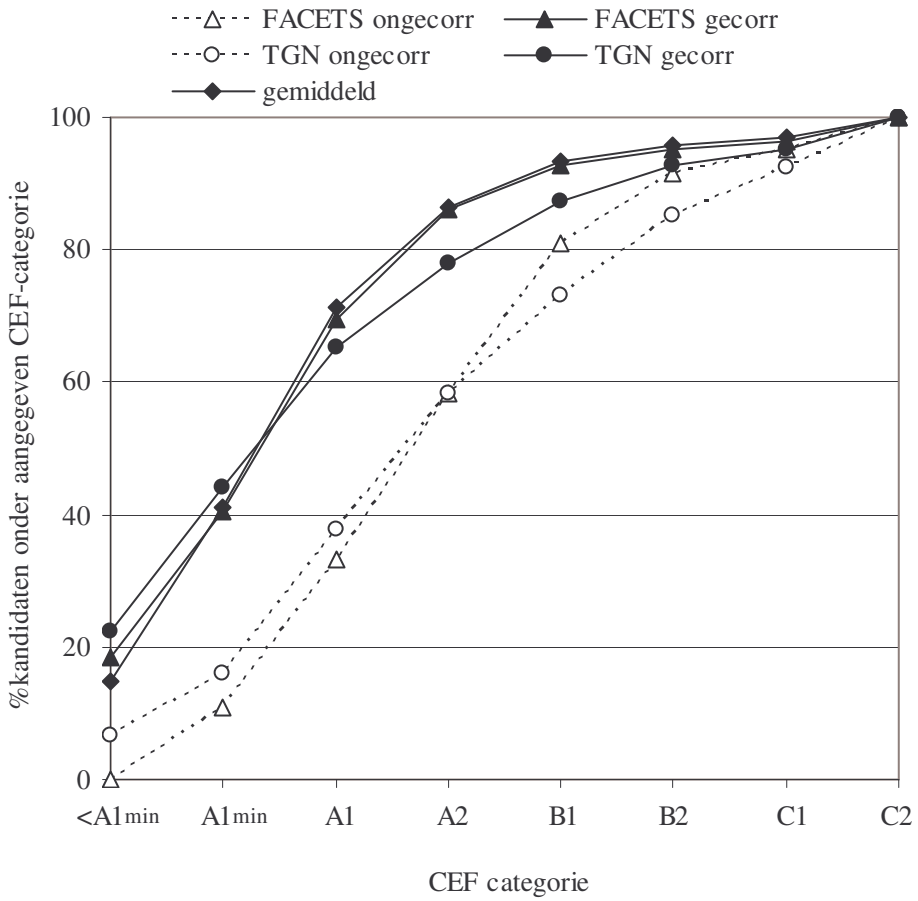
4.2.4 Resultaten onderzoeksvraag 2a: plaatsing van de CEF-grenswaarden$_{FACETS}$

TNO heeft niet alleen gekeken naar de afstanden tussen de CEF-grenswaarden, maar ook naar de absolute plaatsing van de grenswaarden op de kandidaatvaardigheidschaal. Hierbij is geconstateerd dat de CEF-grenswaarden absoluut gezien soepeler zijn ingesteld dan beoogd. In paragraaf 4.1.2 is uitgelegd dat er een keuze gemaakt moet worden ten opzichte van welk criterium de schaal verankerd wordt. Een logisch verdedigbare keuze is om de CEF-grenswaarden uit te drukken ten opzichte van de gemiddelde beoordelaarstrengheid en itemmoeilijkheid. Logisch gevolg is dat de CEF-grenswaarden dan gecorrigeerd moeten worden met de gemiddelde itemmoeilijkheid. In de originele schalingsmethode zijn echter de CEF-grenswaarden uit de FACETS-outputfile direct gebruikt, zonder rekening te houden met de gemiddelde itemmoeilijkheid[17]. Dit betekent dat de CEF-grenzen aanzienlijk soepeler zijn ingesteld dan volgens het beoogde criterium (het gemiddelde item). Een verschil van 4 punten op de FACETS-schaal is groot, het komt overeen met een verschil van ruim een vijfde van de schaal. Dit komt overeen met ruim één CEF-niveau.

Figuur 3 onderbouwt de constatering van TNO. Hierin staan de cumulatieve percentages kandidaten per CEF-categorie, die op vijf manieren zijn verkregen:

1. **Gemiddeld:** Door het gemiddelde te nemen van de originele CEF-oordelen over een kandidaat (afgerond). Het gemiddelde oordeel lijkt op een vaardigheidsschatting op de FACETS-schaal, maar het is niet hetzelfde aangezien geen rekening wordt gehouden met beoordelaarstrengheid en itemmoeilijkheid.
2. **FACETS ongecorrigeerd:** Door het CEF-niveau te bepalen op basis van de kandidaatvaardigheidschattingen op de FACETS-schaal en de CEF-grenswaarden$_{FACETS}$ uit de originele analyse.
3. **FACETS gecorrigeerd:** Door de CEF-grenswaarden$_{FACETS}$ uit de oorspronkelijke analyse te corrigeren met de gemiddelde itemmoeilijkheid (erbij optellen).
4. **TGN ongecorrigeerd**: Door het CEF-niveau te bepalen op basis van de originele TGN toetsscores en CEF-grensscores.
5. **TGN gecorrigeerd**: Door de toetsscores uit de originele analyse te corrigeren. Hiertoe worden de parameters van formule (5) opnieuw geschat op basis van de CEF-grenswaarden$_{FACETS}$ die gecorrigeerd zijn met de gemiddelde itemmoeilijkheid (erbij optellen). (WL: idem) Vervolgens worden de totaalscores opnieuw berekend met de gecorrigeerde formule (5).

[17] Naast gemiddelde itemmoeilijkheid speelt ook beoordelaarsstrengheid een rol. De gemiddelde beoordelaarsstrengheid (en kandidaatvaardigheid) is in de Facets-analyse echter op nul gezet, terwijl de itemmoeilijkheid wèl kan varieren.



Figuur 3: Cumulatieve percentage kandidaten per taalvaardigheidsniveau op basis van verschillende methoden om het CEF-niveau te bepalen

In Figuur 3 worden de cumulatieve percentages[18] kandidaten per CEF-categorie gegeven. Het cumulatieve percentage geeft aan hoeveel procent van de kandidaten een taalvaardigheid heeft onder het aangegeven niveau. Het is te zien dat de cumulatieve percentages overeenkomen met het gemiddelde CEF-oordeel als gecorrigeerde grenswaarden/scores worden gebruikt (doorgetrokken lijnen), terwijl de percentages te laag zijn als ongecorrigeerde grenswaarden/scores worden gebruikt (gestippelde lijnen).

In Figuur 3 valt verder op, dat de cumulatieve percentages op basis van de gecorrigeerde FACETS-grenswaarden (gevulde driehoekjes) beter overeenkomen met het gemiddelde CEF-oordeel dan de percentages op basis van de gecorrigeerde TGN-grensscores (gevulde rondjes). De oorzaak moet gezocht worden in het feit dat de grenswaarden op de FACETS-schaal rechtstreeks aan de menselijke oordelen ontleend zijn, terwijl de grensscores op de TGN-schaal vervolgens via een lineaire regressieanalyse aan de grenswaarden op de FACETS-schaal gerelateerd zijn. De schattingsfout die gemaakt wordt in de lineaire regressieanalyse komt dus alleen tot uiting in de percentages die gebaseerd zijn op de TGN-grensscores.

Bij het bekijken van de regressieanalyse (die door het consortium is uitgevoerd) is opgevallen dat de relatie tussen $theta_{FACETS}$ en $theta_{TGN}$ (stap 2 van de

[18] In Bijlage G, Tabel 18 zijn de absolute percentages kandidaten weergegeven per CEF-categorie

schalingsmethode) niet helemaal lineair is, omdat de machine bij het laagste taalvaardigheidsniveau meer onderscheid maakt dan de menselijke beoordelaars. Wanneer de grensscores opnieuw bepaald worden, bevelen wij aan om:

- De gevolgen van de nieuwe instelling van de grensscores te evalueren met data zoals vermeld in Figuur 3 (of Tabel 18, Bijlage G).
- Na te gaan of de transformatie met een lineaire functie wel optimaal is.

4.2.5 Discussie onderzoeksvraag 2a

Hoewel de A1min-grens soepeler is ingesteld dan beoogd betekent dit nog niet dat deze naar boven hoeft worden bijgesteld. Bij de instelling van een zak-/slaaggrens dient een afweging gemaakt te worden tussen de twee typen fouten die een toets maakt, namelijk het onterecht zakken en onterecht slagen. Deze besliskundige afweging zal gemaakt moeten worden op basis van inhoudelijke en beleidsmatige argumenten.

De verankering van de schaal nu gedaan is t.o.v. een item met een moeilijkheid die 4 punten lager ligt dan de gemiddelde moeilijkheid. Het ligt volgens TNO meer voor de hand om te kiezen voor verankering t.o.v. het beoogde criterium, namelijk de gemiddelde itemmoeilijkheid. Deze keuze is echter niet hard, er kan op basis van goede argumenten ook best gekozen worden voor verankering ten opzichte van een makkelijker of moeilijker item. Hetzelfde geldt voor beoordelaars: nu is impliciet gekozen voor verankering ten opzichte van een beoordelaar met strengheid 0, omdat dit in de FACETS-analyse per definitie gebeurt.

4.2.6 Conclusies onderzoeksvraag 2a

De gevonden afstanden tussen CEF-grenswaarden komen goed overeen met de eerder gevonden afstanden. De gehele schaal is echter verankerd volgens een criterium dat ruim één vijfde van de schaal soepeler is dan beoogd.

4.2.7 Samenvatting onderzoeksvraag 2a

Vraag 2a

- Met de verbeteringen vinden we dezelfde afstanden tussen de zak-/slaaggrenzen.
- De gehele schaal is echter verankerd volgens een criterium dat ruim één vijfde van de schaal soepeler is dan beoogd.

4.3 Onderzoeksvraag 2b

Doel van onderzoeksvraag 2b is om te onderzoeken of dezelfde schaal wordt gevonden wanneer deze gebaseerd wordt op data die verzameld zijn in het buitenland in plaats van in het binnenland.

4.3.1 Methode onderzoeksvraag 2b

Om na te gaan of de taalvaardigheidsschaal in het buitenland hetzelfde functioneert als in Nederland, zal de schaal op basis van de in het buitenland verzamelde menselijke CEF oordelen vergeleken worden met de bestaande schaal.

De praktijkdata zijn op ambassades in het buitenland verzameld en daarbij was het praktisch niet haalbaar om ‘interviews’ af te nemen. Verder bleek het niet haalbaar om extra ‘open vragen’ toe te voegen aan de toets. Een onderdeel van de toets dat niet wordt gebruikt bij de automatische scoring is het onderdeel ‘verhaaltjes navertellen’. De schatting van de CEF-grenswaarden$_{FACETS}$ in het buitenland zal daarom gebaseerd

worden op menselijke oordelen over ‘verhalen navertellen’. Om zeker te zijn dat de resultaten niet afhankelijk zijn van het item type, zullen ook de in Nederland verzamelde ‘verhaaltjes navertellen’ scores apart geanalyseerd worden. Bij de interpretatie zal rekening worden gehouden met het aantal kandidaten per scorebereik.

Om te beoordelen of de taalvaardigheidschaal hetzelfde functioneert, zal eerst gekeken worden naar de geschatte CEF-grenswaarden$_{FACETS}$ in binnen- en buitenland. Om toch zo goed mogelijk in te kunnen schatten wat de gevolgen zijn van eventuele verschillen in schaling, zullen vervolgens de CEF-grensscores van de TGN opnieuw bepaald worden (zoals in paragraaf 4.1) op basis van de CEF-grenswaarden$_{FACETS}$ die in de buitenlanddata geschat zijn. Daarmee wordt concreet duidelijk hoeveel de CEF-grensscores verschillen als deze op basis van de binnenland- versus de buitenlanddata vastgesteld worden.

4.3.2 Data onderzoeksvraag 2b

Om de zak-/slaaggrens in het buitenland opnieuw vast te kunnen stellen zijn menselijke oordelen verzameld van de taalvaardigheid van kandidaten uit het buitenland op basis van navertelde verhaaltjes. De verhaaltjes vormen geen onderdeel van de toets en zijn alleen voor validatiedoeleinden toegevoegd.

De CEF-oordelen over de ‘navertelde verhaaltjes’ zijn verzameld voor de eerste 500 kandidaten die de TGN aflegden, met inachtneming van dezelfde uitsluitingscriteria als voor onderzoeksvraag 1 (zie paragraaf 3.1.1). Per kandidaat zijn twee verhaaltjes naverteld en beoordeeld. Ieder verhaaltje werd beoordeeld door twee beoordelaars die willekeurig werden getrokken uit een pool van 5 beoordelaars. Enkele statistieken voor de data die in de analyse zijn gebruikt staan vermeld in Tabel 16.

Voordat beoordelaars konden deelnemen aan het onderzoek dienden zij aan een training deel te nemen. Zowel de dataverzameling als de training zijn uitgevoerd door het consortium. TNO is aanwezig geweest bij de training, maar teneinde onafhankelijkheid te waarborgen zijn de door TNO gemaakte CEF-oordelen niet in het onderzoek betrokken. Bij de beoordeling van de verhaaltjes werd gebruik gemaakt van CEF-descriptoren voor gespreksvaardigheid, zie bijlage G. De beoordelaars moesten aan het einde van de training een test af leggen die op dezelfde manier was opgezet als de test voor de kwalitatieve beoordelingen van onderzoeksvraag 1 (zie paragraaf 2.1.4).

Tabel 16 Statistieken van de datasets gebruikt voor onderzoeksvraag 2b.

		kandidaten	items	beoordelaars	data
binnenland	origineel	815	59	13	4401
	nieuw	538	57	9	3375
buitenland		467	99	5	1570

In Tabel 16 staan de statistieken van de datasets die gebruikt zijn voor onderzoeksvraag 2b. Logischerwijs zitten er geen gemeenschappelijke kandidaten in de ‘binnenland’ en ‘buitenland’ datasets, want in het buitenland gaat het om nieuwe examenkandidaten. Op de ander facetten (beoordelaar en items) is er wél overlap: Er is één beoordelaar die zowel ‘binnenland’ als ‘buitenland’ data heeft beoordeeld en er zijn 53 gemeenschappelijke items.

Voor het bepalen van de CEF-grenswaarden is weer gebruik gemaakt van de FACETS-software [refA].

4.3.3 Resultaten onderzoeksvraag 2b

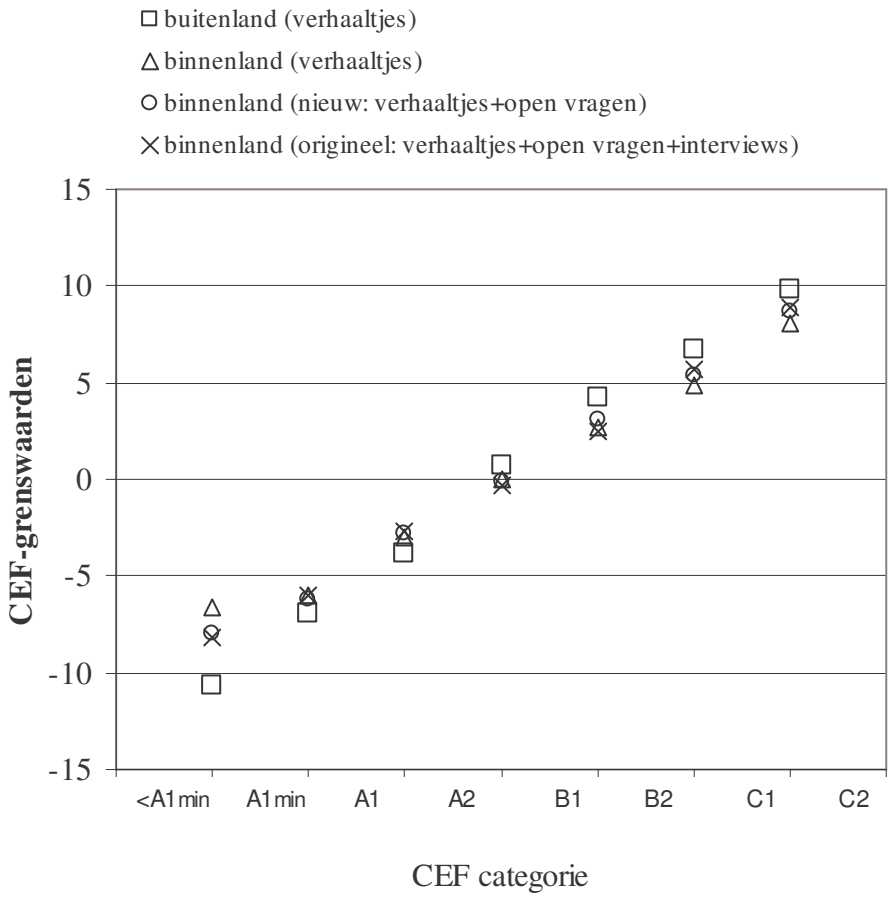
Analyse van de nieuwe buitenlanddata levert nieuwe schattingen van de afstanden tussen de CEF-grenswaarden$_{FACETS}$. Bij het opstellen van het onderzoeksprotocol (zie appendix A) verwachtten wij op basis van de reeds verzamelde toetsuitslagen dat het aantal kandidaten rond de A1min-grens klein zou zijn. Uit de FACETS-analyse blijkt echter dat het aantal kandidaten rond de A1min-grens aanzienlijk is. Het verschil kan verklaard worden aangezien de grensscore van de toets is ingesteld t.o.v. een soepel criterium (zie 4.2.4).

Om te onderzoeken of eventuele verschillen in afstanden verklaard kunnen worden door het gebruikte itemtype (in het buitenland zijn alleen 'verhaaltjes' beoordeeld) zijn de CEF-grenswaarden$_{FACETS}$ ook apart bepaald voor alleen de verhaaltjes uit het binnenland. De schattingen op basis van de buitenlanddata zijn weergegeven in Figuur 4, samen met drie varianten van de schattingen in het binnenland: de originele schattingen, de nieuwe schattingen, en schattingen op basis van alleen 'verhaaltjes navertellen'. Het meest opvallende is dat de range van de geschatte CEF-grenswaarden$_{FACETS}$ in het buitenland aanzienlijk groter is dan in alledrie de binnenland analyses. Verder is te zien dat de A1min-grens voor alleen de verhaaltjes in het binnenland minder ver van de A1-grens afligt dan voor alle data in het binnenland. De verdelingen van kandidaatvaardigheid in binnen- en buitenland zijn vrijwel gelijk[19], dus dit kan geen verklaring zijn voor de gevonden verschillen in range.

De verklaring voor de verschillen in range ligt waarschijnlijk in de schattingsmethode van de FACETS-software. Onzekerheid over de maximum-likelihood parameterschattingen resulteert in schattingen die bias (vertekening) naar de extremen hebben. Door bias naar de extremen wordt de range groter. De resultaten suggereren verder dat de A1min-grenswaarde in het buitenland (volgens de menselijke beoordelaars) verder van de andere grenswaarden af ligt. Het is niet aannemelijk dat dit verschil geheel het gevolg is van bias, omdat juist op het '<A1min-' en A1min-niveau relatief veel data zijn.

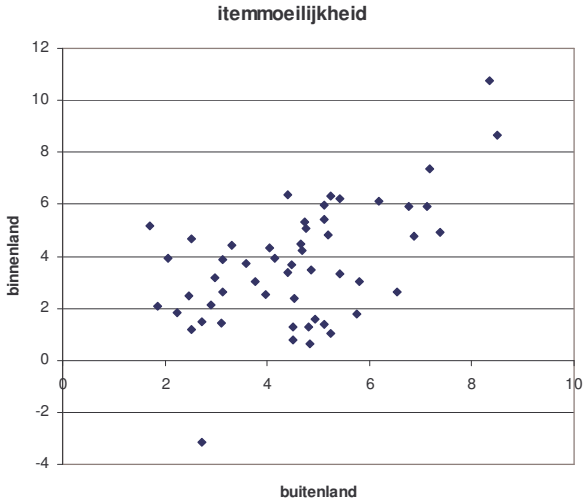
Een ander verschil tussen de binnen- en buitenlanddata is dat het soort telefoonverbinding (MFA-netwerk) in het buitenland anders is dan in een deel[20] van de binnenlanddata (PSTN-netwerk). De kwaliteit van het MFA-netwerk is minder goed (o.a.'package-loss') dan die van het PSTN-netwerk. Het TGN-eindrapport vermeldt dat met het MFA-netwerk minder hoge scores worden behaald dan met het PSTN-netwerk. Als dit effect ook optreedt bij menselijke beoordelaars en dan met name bij hoogtaalvaardige kandidaten, zou de A1min-grens daardoor verder van de A1-grens af kunnen liggen.

[19] zie Figuur 8 in appendix H
[20] van dataset 1



Figuur 4 Afstanden CEF-grenswaarden$_{FACETS}$ voor binnen- en buitenland data.

Naast het vergelijken van de afstanden tussen de CEF-grenswaarden in binnen- en buitenland, hebben we de geschatte itemmoeilijkheid in binnen- en buitenland vergeleken. In Figuur 5 staat de geschatte itemmoeilijkheid van de verhaaltjes die zowel in binnen- als buitenland gescoord zijn. Er is wel een duidelijk verband tussen de parameters, vooral de moeilijke items (rechts boven in de figuur) zijn zowel in binnen- als buitenland moeilijk. De standaardfouten voor moeilijkheidsschattingen zijn echter groot, dit verklaart waarom de verschillen in moeilijkheid voor veel items erg groot zijn.



Figuur 5 Geschatte itemmoeilijkheid van gemeenschappelijke items in binnen- en buitenland.

Om de gevolgen van de gevonden verschillen schaling concreet te maken zijn de CEF-grensscores van de TGN opnieuw bepaald op basis van de CEF-grenswaarden$_{FACETS}$ die in de buitenlanddata geschat zijn. De resultaten staan vermeld in Tabel 17. De A1min-grens op basis van een analyse met de items 'verhaaltjes navertellen' verzameld in het buitenland komt acht punten lager te liggen dan in de huidige situatie.

De grenswaarden worden via een lineaire regressie analyse getransformeerd naar de grenswaarden in de Engelse en Spaanse tests (zoals in paragraaf 4.1.4). In deze tests is het A1min-niveau echter nog niet gedefinieerd. Op basis van de gevonden regressiefunktie wordt de A1min-grenswaarde op de TGN-schaal geprojecteerd. Het verschil tussen de bestaande grenswaarden en de grenswaarden op basis van de buitenlanddata wordt daardoor met name zichtbaar in de ligging van de A1min-grens.

Tabel 17 CEF-grensscores op basis van buitenland- en de binnenlanddata

	grensscores op basis van buitenlanddata	grensscores op basis van binnenlanddata
A1min	8	16
A1	25	26
A2	35	37
B1	49	47
B2	60	57
C1	68	68
C2	78	80

4.3.4 Discussie onderzoeksvraag 2b

Voor zover de gevonden verschillen niet te verklaren zijn op basis van steekproefgrootte, of verschillen in telefoonnetwerk, zijn ze in strijd met het FACETS-model. De verschillende ligging van de A1min-grenswaarde in binnen- en buitenland impliceert dat het objectief meten van taalvaardigheid met de items 'verhaaltjes navertellen' niet mogelijk is. Dit komt overeen met de bevinding in het TGN-eindrapport [refB, paragraaf 4.1.3.2] dat de opgaven ‘verhaaltjes navertellen’ voor laagtaalvaardigen te moeilijk zijn. Vanwege de complicatie van bias, het verschil in telefoonnetwerk en het feit dat conclusies over verhaaltjes navertellen niet één op één

naar de machine te vertalen zijn, kunnen geen harde conclusies getrokken worden over de schaling in het buitenland.

De aanwijzing dat de A1min-grensscore op basis van 'verhaaltjes' in het buitenland lager geschat wordt, suggereert dat de menselijke beoordelaars in het buitenland een andere grens hanteren voor het toekennen van het A1min-niveau. Dit is een overweging die bij het instellen van de A1min CEF-grensscore meegenomen moet worden.

4.3.5 Conclusies onderzoeksvraag 2b

Het A1min-niveau in het buitenland lijkt (volgens menselijke beoordelaars van de items 'verhaaltjes navertellen') lager te liggen dan in het binnenland. Er kunnen echter geen harde conclusies getrokken worden over de schaling in het buitenland.

4.3.6 Samenvatting onderzoeksvraag 2b

Vraag 2b

- De A1min zak-/slaaggrens geschat op basis van navertelde verhaaltjes lijkt in het buitenland lager te liggen dan in het binnenland

4.4 Overwegingen bij de conclusie

Aangezien de schaal is verankerd volgens een criterium dat soepeler is dan beoogd, dient bijstelling van de huidige instelling van de zak-/slaaggrens overwogen te worden. Dit kan bereikt worden door de schaal te verankeren t.o.v. een strenger criterium (bijvoorbeeld gemiddelde itemmoeilijkheid). De omvang van de bijstelling dient op grond van inhoudelijke en beleidsmatige argumenten gekozen te worden.

Als de schaal verankerd wordt t.o.v. de gemiddelde itemmoeilijkheid, dan betekent dit dat een kandidaat met een vaardigheid die op de A1min-grens ligt voor een item met een gemiddelde moeilijkheid evenveel kans heeft om ten onrechte te zakken als ten onrechte te slagen. Er zijn voor de A1min-grens echter methodologische argumenten om de schaal minder streng te verankeren:

- Er zijn veel meer kandidaten die in werkelijkheid boven A1min zitten[21], dan onder A1min. Deze verdeling maakt dat bij een neutrale instelling van zak-/slaaggrens absoluut gezien meer kandidaten onterecht zakken dan onterecht slagen (want, er zijn veel meer potentiële onterecht zakkers).
- Op basis van menselijke oordelen over ‘verhaaltjes navertellen’ in het buitenland, wordt een grotere afstand tussen A1min en A1 gevonden, wat impliceert dat het A1min-niveau in het buitenland mogelijk (volgens menselijke beoordelaars) lager ligt. Een soepele instelling van zak-/slaaggrens komt hieraan tegemoet.
- De relatie tussen menselijke en machine scores lijkt juist op het laagste niveau niet helemaal lineair te zijn, terwijl bij stap 2 van de schalingsmethode wel een lineair verband wordt aangenomen. Hierdoor komt de A1min-grens mogelijk juist iets hoger te liggen.

[21] volgens de FACETS-analyse heeft 6.5% van de kandidaten een taalvaardigheid onder de A1min zak-/slaaggrens, hierbij is gecorrigeerd voor gemiddelde itemmoeilijkheid

Daarnaast is er een inhoudelijk argument om voorzichtig om te gaan met het verhogen dan de A1min zak-/slaaggrens. De definitie van het A1min-niveau is nieuw en gecompliceerd. Aangezien kandidaten met een taalvaardigheidniveau van A1min of lager slechts weinig Nederlands produceren is het moeilijk om hierover een oordeel te geven. Vanwege deze complicatie van de definitie van A1min-niveau lijkt een soepele toets redelijker dan een strenge.

4.5 Conclusie

- De zak-/slaaggrens is ingesteld volgens een criterium dat ruim een vijfde van de schaal soepeler is dan beoogd.

Dit betekent dat bijstelling van de huidige instelling van de zak-/slaaggrens overwogen dient te worden. Bijstelling kan bereikt worden door de schaal te verankeren t.o.v. een strenger criterium (bijvoorbeeld gemiddelde itemmoeilijkheid). De omvang van de bijstelling dient op grond van inhoudelijke en beleidsmatige argumenten gekozen te worden. Een andere verankering van de schaal heeft gevolgen voor de grenswaarden voor alle niveaus van de toets.

4.6 Samenvatting

Onderzoeksvraag 2

Resultaten schalingsmethode

TNO voerde twee verbetering door t.a.v. het gezamenlijk analyseren van de datasets:

1. Koppel de data op basis van andere aannames over de data
2. Verwijder de ongekoppelde interviewdata

Resultaten vraag 2a

- Met de verbeteringen vinden we dezelfde afstanden tussen de zak-/slaaggrenzen.
- De gehele schaal is echter verankerd volgens een criterium dat ruim een vijfde van de schaal soepeler is dan beoogd.

Resultaten vraag 2b

- De zak-/slaaggrens geschat op basis van navertelde verhaaltjes lijkt in het buitenland lager te liggen dan in het binnenland.

Conclusie

- De zak-/slaaggrens is ingesteld volgens een criterium dat ruim een vijfde van de schaal soepeler is dan beoogd.

Consequentie

- Bijstelling van de huidige instelling van de zak-/slaaggrens dient overwogen te worden.

5 Referenties

[refA] Linacre, J.M. (1988),
A Computer Program for the Analysis of Multi-Faceted Data,
Chicago, IL: Mesa Press.

[refB] Kerkhoff, A., Poelmans, P., de Jong, J., Lennig, M.,
Verantwoording Toets Gesproken Nederlands, CINOP in samenwerking met Language Testing Services en Ordinate,
19 september 2005

[refC] commissie Franssen/Franssen, J. et al. (2004),
Inburgering getoetst. Advies over het niveau van het nieuwe inburgeringexamen in het buitenland,
Den Haag.

[refD] M. Tijssen, A. Berntsen, J. de Jong, M. Lennig (2005)
Verantwoording Inburgeringsexamen Kennis van de Nederlandse Samenleving
CINOP in samenwerking met Language Testing Services en Ordinate,

[refE] Kessens, J.M., van Wijngaarden, S., van Leeuwen, D. (2005),
Second opinion ten aanzien van de validatie van de spraaktechnologie gebruikt bij het inburgeringexamen,
rapport nr.TNO-DV3 2005 098.

[refF] Minister voor Vreemdelingenzaken en Integratie
brief 5409878/06, 28 april 2006

[refG] prof. W. J.Van der Linden
Professor of Measurement and Data Analysis, Faculty of Behavioral Sciences
University of Twente.

[refH] Heuvelmans, A.P.J.M. & Sanders, P.F. (1993),
Beoordelaarsovereenstemming,
In: Eggen, T.J.H.M. & P.F. Sanders. Psychometrie in de praktijk,
Arnhem, Cito.

[refI] Fleiss, J. L & P. E. Shrout (1978),
Approximate interval estimation for a certain intraclass correlation coefficient,
Psychometrika, 43, 259 – 262.

[refJ] Shrout, P.E. & Fleiss, J.L. (1979),
Intraclass Correlations: Uses in Assessing Rater Reliability,
Psychological Bulletin, 2, 420-428.

[refK] Ordinate Corporation (2004)
Set-10 Test Description & Validation Summary

[refL] Ordinate Corporation (2004)
Spoken Spanish Test (SST) Test Description & Validation Summary

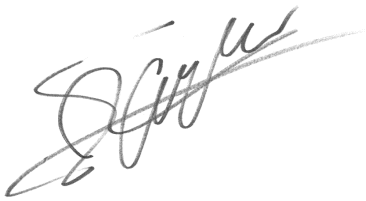
[refM] Jodoin and Gierl (2001),
Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. Applied Measurement in Education,
Vol. 14, No. 4, Pages 329-349.

6 Ondertekening

Soesterberg, februari 2007
TNO Defensie en Veiligheid

dr. ir. J.M. Kessens
Auteur 1

Leiden, februari 2007
TNO Kwaliteit van Leven



drs. Gert Jacobusse
Auteur 2

A Gedetailleerd onderzoeksvoorstel

In dit document worden de twee onderzoeksvragen gedetailleerd beschreven.
Dit onderzoeksvoorstel is opgesteld door TNO en is tot stand gekomen in overleg met PMI, CINOP/LTS/Ordinate. Voor een aantal onderdelen van het onderzoeksvoorstel is TNO geadviseerd door prof. W. J.Van der Linden[22].

A.1 Onderzoeksvraag 1

De volgende onderzoeksvraag zal beantwoord worden:

Zijn er substantiële verschillen in de beoordeling tussen het systeem dat automatisch uitslagen genereert voor de Toets Gesproken Nederlands (TGN) en voor de Toets Kennis van de Nederlandse Samenleving (KNS) en menselijke beoordelingen?

Het automatische scoringssysteem van de TGN en de toets KNS rapporteert een totaalscore die voor de TGN ligt tussen 10 en 80 en voor de KNS tussen 0% en 100%. Deze totaalscores vormen respectievelijk een beoordeling van de mondelinge taalvaardigheid van een kandidaat en van diens kennis over een gedefinieerde verzameling aspecten van de Nederlandse samenleving. Op basis van deze totaalscore en de zak-/slaaggrens wordt de toetsuitslag van een kandidaat bepaald: een kandidaat is gezakt of geslaagd. Het automatische scoringssysteem is zó getraind dat de automatische scoring zo goed mogelijk de menselijke scoring benadert. Dit impliceert dat het menselijke oordeel het criterium is waartegen een computerbeoordeling gemeten wordt. Om dit criterium met hoge statistische betrouwbaarheid te schatten moet een score verkregen worden die gebaseerd is op meerdere menselijke beoordelingen. Het doel van dit onderzoek is om te achterhalen of er verschillen bestaan tussen de beoordeling door de computer en door de menselijke beoordelaars. Indien die verschillen klein zijn (kleiner dan de vooraf vastgestelde norm), betekent dit dat de automatische scoring de kandidaten niet bevoor- of benadeelt ten opzichte van het menselijke oordeel, en behoeft het scoringssysteem niet te worden aangepast.

Als die verschillen wél substantiëel zijn (groter dan de norm), kan door PMI besloten worden tot een apart onderzoek, waarin de oorzaken van de mogelijke verschillen achterhaald worden en zal bekeken worden hoe die verschillen geminimaliseerd kunnen worden. Tevens zullen dan aanbevelingen gedaan worden op welke wijze het automatische systeem het beste hertraind kan worden en op welke manier de effectiviteit van de hertraining gevalideerd kan worden.

A.1.1 Criterium

Validiteit van het automatische scoringssysteem wordt bepaald ten opzichte van een criterium dat gebaseerd is op het viermaal beoordelen van de opgenomen responsen. Hierbij wordt gecorrigeerd voor eventuele verschillen in strengheid die kunnen optreden tussen de verschillende beoordelaars. Als maat voor validiteit wordt genomen de voor attenuatie gecorrigeerde correlatie tussen de automatische toetsscores en het criterium. Indien nodig en mogelijk zal gecorrigeerd worden voor ‘restriction of range’. Als de validiteitcoëfficiënt voldoet aan een vooraf gestelde kwaliteitsnorm, betekent dit dat aangetoond is dat de kandidaten niet worden bevoor- of benadeeld ten opzichte van

[22] Professor of Measurement and Data Analysis, Faculty of Behavioral Sciences University of Twente.

het menselijke oordeel. De kwaliteitsnorm die gesteld wordt is $r \geq 0.90$.
Om een statistisch betrouwbaar criterium te garanderen worden de beoordelaars van tevoren getraind. Alleen beoordelaars die aan een vooraf gestelde kwaliteitseis voldoen worden in het onderzoek betrokken.

A.1.2 Training beoordelaars en transcribenten

Er worden twee typen beoordelingen gemaakt: kwalitatieve oordelen en woordelijke transcripties. Voor de opzet, uitvoer en inhoudelijke kwaliteit van de training van de beoordelaars en transcribenten is CINOP/LTS verantwoordelijk. Bij de uitvoering van de beoordelingsprocedure wordt gebruik gemaakt van de automatische procedures voor dataverzameling van Ordinate.

Bij de training zal TNO aanwezig zijn. Tevens zal TNO één beoordelaar/transcribent afvaardigen. Teneinde de onafhankelijke positie van TNO te waarborgen zullen oordelen van TNO echter niet in de data worden opgenomen. Waarborging van de kwaliteit van de beoordelingen/transcripties zal worden gerealiseerd door aan het einde van de training beoordelaars een proef te laten afleggen. De proef bestaat uit een aparte set van 50 willekeurig uit de buitenlanddata getrokken responsen (afkomstig van verschillende kandidaten) die verder niet in het hoofdonderzoek worden betrokken. Tijdens de proefperiode krijgen alle beoordelaars dezelfde set van 50 responsen voorgelegd.

Voor de kwalitatieve beoordelaars wordt de beoordelaarsovereenstemming geschat met een intraklasse-correlatiecoëfficiënt [ref 2]. Als norm wordt gesteld dat de ondergrens van het 90% betrouwbaarheidsinterval [ref 3, ref 4] voor de geschatte beoordelaarsovereenstemming bij gebruikmaking van vier willekeurige beoordelaars hoger is dan 0.90. De transscribenten worden geëvalueerd op grond van vergelijking met modeltransscripties. De modeltransscripties worden opgesteld door een expert. Als norm wordt gesteld dat 90% van het aantal getransscribeerde woorden over de totale set van 50 responsen moet overeenstemmen met het model. Uitsluitend beoordelaars en transcribenten die aan de norm voldoen worden in het hoofdonderzoek betrokken.

A.1.3 Data

De praktijkdata bestaan uit de set van toetsafnames waarvan zijn uitgesloten:

- toetsafnames van ambassade-personeel (gebruikt voor training van het ambassadepersoneel);
- toetsafnames waarbij het automatisch systeem geen toetsuitslag heeft gegeven wegens technische problemen of de aanwezigheid van ruis/achtergrondlawaai;
- onvolledige toetsafnames.

Na uitsluiting van bovengenoemde toetsafnames ontstaat een dataset van 500 toetsuitslagen die in de praktijk zijn verkregen. De eerste helft van deze praktijkdata (kandidaten 1-250) zal gebruikt worden om de validiteit van het scoringssysteem vast te stellen. Als het nodig mocht zijn om het automatische scoringssysteem te hertrainen worden hiervoor de bestaande trainingsdata gebruikt, aangevuld met de eerste helft van de praktijkdata en/of eerder verzamelde data. Het effect van eventuele hertraining kan dan gevalideerd worden met de tweede (onafhankelijke) helft van de praktijkdata (kandidaten 251-500).

A.1.4 Opzet onderzoek

Om het automatische scoringsysteem te valideren voeren zowel mens als machine exact dezelfde taak uit met exact hetzelfde materiaal. Er worden twee typen beoordelingen gemaakt.

- Woordelijke transcripties.
  - Voor de 'kort-antwoord items en de 'tegenstelling'-items van de TGN (22) en voor alle items van de KNS (30 items) wordt op basis van de woordelijke transcriptie en het antwoordmodel een dichotome score verkregen; namelijk 0=fout, of 1=goed.
  - Voor de 'zinnen herhalen'-items van de TGN (23) wordt op basis van de woordelijke transcriptie een polytome score verkregen; variërend van 0 tot het maximaal aantal correct herhaalde woorden in de herhaalde zin.
- Kwalitatieve beoordelingen.
  - Voor de 'zinnen herhalen'-items van de TGN (23) worden CEF-beoordelingen gemaakt voor uitspraak en vloeiendheid. Dit zijn polytome scores lopend van 0 (≤ A1min) t/m 7 (= C2).

In tabel A.1 zijn de typen beoordelingen en de daaruit afgeleide typen scores samengevat. Er wordt gewerkt met een pool van beoordelaars (meer dan vier). Om halo-effecten te voorkomen wordt iedere kandidaatrespons aan 4 willekeurig uit de pool getrokken beoordelaars voorgelegd.

Tabel A.1 Typen beoordelingen benodigd voor het onderzoek.

Toets	Aspect	Type score	menselijk oordeel per item	aantal beoordeelde items per kandidaat[23]
TGN	woordenschat	dichotoom: 0/1[24]	woordelijke transcriptie	22
	zinsbouw	polytoom: 0-max.	woordelijke transcriptie	23
	vloeiendheid	polytoom: 0-7	score tussen 0-7	23
	uitspraak	polytoom: 0-7	score tussen 0-7	23
KNS	kennis	dichotoom: 0/1[2]	woordelijke transcriptie	30

Op basis van de verschillende typen itemscores, genereert het automatische scoringssysteem een totaalscore. Met de menselijke beoordelingen wordt voor ieder aspect een aparte analyse uitgevoerd waarbij de vier menselijke oordelen worden gecombineerd tot één menselijke totaalscore. De menselijke totaalscore is de gemiddelde subscore. De machinale en de menselijke totaalscores zullen in één figuur weergegeven worden (één figuur voor de TGN en één figuur voor de KNS). Indien de voor attenuatie gecorrigeerde correlatie groter of gelijk is aan 0,9 worden de verschillen tussen menselijke en machinale scores niet als substantiëel beschouwd. Als de gecorrigeerde correlatie kleiner is dan 0,9, zal in een apart onderzoek achterhaald worden wat mogelijke oorzaken zijn van verschillen. Voor de TGN zullen dan aparte regressieanalyses uitgevoerd worden voor de afzonderlijke deelaspecten.

---

[23] Het eerste item van iedere nieuwe opgavensoort wordt niet in de score meegenomen.

[24] Deze scores wordt automatisch afgeleid op basis van de door mensen vervaardigde woordelijke transcripties.

A.2 Onderzoeksvraag 2

De volgende onderzoeksvraag zal beantwoord worden:

Is de zak-/slaaggrens van de Toets Gesproken Nederlands (TGN) op het goede niveau ingesteld?

In het onderzoek wordt tevens de zak-/slaaggrens zoals die thans is vastgesteld, met data die in het praktijkonderzoek worden verzameld, geconfronteerd. Het doel hiervan is om vast te stellen of met de praktijkdata eenzelfde zak-/slaaggrens wordt gevonden als met de data die in de eerdere analyse zijn gebruikt. Om te bepalen of de gevolgde methodologie voor de instelling van de zak-/slaaggrens herhaald kon worden heeft TNO in een vooronderzoek de methodologie eerst nauwkeurig bestudeerd. Ondertussen kon op basis van de al binnengekomen toetsuitslagen ingeschat worden dat slechts een klein gedeelte van de kandidaten (< 5%) een taalvaardigheidniveau heeft van A1min of lager. Dit betekent dat het niet mogelijk is om met de praktijkdata met voldoende nauwkeurigheid vast te kunnen stellen of de cesuur goed is ingesteld. Om deze reden is de onderzoeksvraag in twee delen gesplitst.

De zak-/slaaggrens wordt opnieuw vastgesteld met de oorspronkelijke data. De koppeling van de data zal echter op een andere manier gebeuren, aangezien onzeker is in hoeverre de manier van datakoppeling de hoogte van de zak-/slaag-grens beïnvloedt.

Voor de hogere taalvaardigheidniveau's zijn de aantallen wèl groot genoeg om de zak-/slaaggrens met voldoende nauwkeurigheid te bepalen. Voor deze hogere taalvaardigheidniveau's zal daarom met de praktijkdata de schaling opnieuw bepaald worden om zo vast te stellen of er verschillen zijn met de schaling op grond van data verzameld in Nederland.

Mochten de resultaten van het onderzoek hiertoe aanleiding geven, dan zal bezien worden òf en op welke wijze tot een aanpassing van de zak-/slaaggrens kan worden besloten.

A.2.1 Vooronderzoek

In een vooronderzoek is de schalingsmethodologie voor de bepaling van de zak-/slaaggrens nader bekeken, aangezien in het TGN-eindrapport [ref1] details over de uitgevoerde methodologie ontbraken. Door bestudering van het rapport en veelvuldig overleg met CINOP/LTS/Ordinate zijn de details van de methodologie duidelijk geworden.

Van belang is dat voor de bepaling van de zak-/slaaggrens is gebruik gemaakt van twee datasets. Deze twee datasets konden niet aan elkaar gekoppeld worden via het facet 'type beoordeling' aangezien CEF-beoordelingen van een andere type zijn gebruikt:

- 'Verhaaltjes navertellen': Kandidaten moeten een kort verhaal navertellen. Het CEF-oordeel wordt gegeven op basis van het beluisteren van een audio-opname van het navertelde verhaal.
- 'Interviews': In een adaptief interview wordt aan de hand van vast interviewprotocol (zie bijlage 11 van [ref 1]) het CEF-niveau van een kandidaat bepaald.
- 'Open vragen': Kandidaten moeten antwoord geven op een aantal open vragen die aan het einde van de toets gesteld worden. Het CEF-oordeel wordt gegeven door audio-opnamen van deze antwoorden te beoordelen.

Voor dataset 1 zijn CEF-beoordelingen van het type ‘verhaaltjes navertellen’ gebruikt, terwijl voor dataset 2 ‘interviews’ en ‘open vragen’ werden gebruikt. De datasets zijn gekoppeld via imputatie van 35 datapunten (schattingen).

A.2.2 Hoofdonderzoek

Het hoofdonderzoek bestaat uit twee delen:

- opnieuw vaststellen van de A1min zak-/slaaggrens;
- vaststellen of er verschillen zijn in schaling zijn op grond van data verzameld in Nederland vs. in het buitenland.

De opzet van de twee deelonderzoeken zal hieronder beschreven worden.

A.2.3 Opnieuw vaststellen van de A1min zak-/slaaggrens

Op basis van de uitslagen van de eerste 250 examens (inclusief uitsluitingen, zie paragraaf 1.3) is in te schatten dat slechts een klein gedeelte van de kandidaten (< 5%) een taalvaardigheidniveau heeft van A1min of lager. Om deze reden is het niet mogelijk om op basis van de praktijkdata met voldoende nauwkeurigheid te bepalen of A1min zak/slaaggrens op het goede niveau is ingesteld.

Uit het vooronderzoek is gebleken dat de gevolgde methodologie correct is. Het is echter onzeker hoe gevoelig de instelling van de zak-/slaaggrens is voor de manier van datakoppeling. Aangezien de nieuwe data verzameld in het buitenland niet gebruikt kunnen worden voor het vaststellen van de A1min zak-/slaaggrens, zal de A1min zak-/slaaggrens opnieuw moeten vastgesteld worden op basis van de eerder verzamelde datasets die door het consortium zijn gebruikt (pretest en MFA-fit). TNO zal de datasets op een andere manier koppelen, om na te gaan hoe gevoelig de schatting van het A1min niveau is voor de manier van koppelen. De koppeling van de datasets die TNO voorstelt is gebaseerd op de gelijkheidsassumptie dat alle items bedoeld zijn om kandidaten in dezelfde CEF-categorieën in te delen. De assumptie stelt tijdelijk één item uit de pretest data gelijk aan één item uit de MFA-fit data. Hierdoor ontstaat een link, die het mogelijk maakt om de datasets samen te analyseren. De analyse wordt een aantal keren herhaald, steeds met een andere gelijkheids assumptie. In iedere analyse zal de schatting van het A1min niveau iets anders zijn.

Als resultaat zullen een minimum, een maximum, een gemiddelde en een standaarddeviatie (uitgedrukt op de rapportageschaal) van de A1min zak-/slaaggrens geschat worden. Hierdoor ontstaat inzicht in de onzekerheid over de A1min zak-/slaaggrens als gevolg van de koppeling.

Over de exacte uitvoering van de datakoppeling zal TNO nog een extern advies inwinnen. Dit externe advies zal worden meegenomen in de uiteindelijke uitvoering van onderzoeksvraag 2a.

A.2.4 Vaststellen of er verschillen zijn in schaling op grond van data verzameld in Nederland vs. in het buitenland

Zoals eerder vermeld is het niet mogelijk om met de praktijkdata de A1min zak-/slaaggrens voldoende nauwkeurig te bepalen. Om toch een uitspraak te kunnen doen over eventuele verschillen in schaling, zullen grenzen tussen de CEF-niveaus waarvoor wel genoeg data zijn, geschat worden met de praktijkdata.

De praktijkdata zijn op ambassades in het buitenland verzameld, en daarbij was het praktisch niet haalbaar om ‘interviews’ af te nemen. Verder bleek het niet haalbaar om extra ‘open vragen’ toe te voegen aan de toets. Een onderdeel van de toets dat niet wordt gebruikt bij de automatische scoring is het onderdeel ‘verhaaltjes navertellen’. Kandidaten dienen twee korte verhalen na te vertellen. De schatting van de grenzen tussen CEF-niveaus zal daarom gebaseerd worden op menselijke oordelen over ‘verhalen navertellen’ in de praktijkdata. Bij analyse van de pretest gegevens is gebleken dat met ‘verhalen navertellen’ weinig onderscheid gemaakt kan worden binnen de laagste CEF niveaus. Aangezien de vergelijking betrekking heeft op de hogere CEF niveaus, zal dit echter geen probleem zijn.

Voor de opzet, uitvoering en kwaliteit van de training van de beoordelaars is CINOP/LTS verantwoordelijk. De gevolgde procedure voor training van de beoordelaars en de kwaliteitsnorm die aan de beoordelaars wordt opgelegd is hetzelfde als beschreven in paragraaf ‘training beoordelaars en transcribenten’ van onderzoeksvraag 1.

De grenzen tussen CEF-niveaus zullen geschat worden op basis van menselijke beoordelingen van ‘verhalen navertellen’ die voor de praktijkdata verzameld zijn. Voor 500 kandidaten zal een CEF-oordeel worden gegeven op basis van het beluisteren van een audio-opname van de twee navertelde verhalen. Hierbij zal ervoor gezorgd worden dat tenminste één van de beoordelaars ook de pretest gegevens beoordeeld heeft. Door deze koppeling te maken, is er de mogelijkheid om de grenswaarden uit te drukken op dezelfde $theta_{CEF}$ schaal als de grenswaarden in tabel 5.3 in het TGN-eindrapport [ref 1].

Het criterium dat gesteld wordt is dat geen van de nieuw geschatte grenswaarden significant ($p < 0,05$) afwijkt van de in tabel 5.3 genoemde afkappunten [ref1]. Op basis van de standaard errors van de schattingen wordt een 95% betrouwbaarheidsinterval voor het verschil tussen de oude en de nieuwe schatting gemaakt. Dit betrouwbaarheids-interval is niet alleen de toets of het verschil significant is, maar geeft ook aan hoe groot het verschil maximaal kan zijn. Met andere woorden: hoe zeker we ervan zijn dat in de praktijkdata dezelfde grenswaarden gelden.

A.3 Referenties

[ref 1] Kerkhoff, A., Poelmans, P., de Jong, J., Lennig, M.,
Verantwoording Toets Gesproken Nederlands,
CINOP in samenwerking met Language Testing Services en Ordinate, 19 september 2005

[ref 2] Heuvelmans, A.P.J.M. & Sanders, P.F. (1993),
Beoordelaarsovereenstemming,
In: Eggen, T.J.H.M. & P.F. Sanders,
Psychometrie in de praktijk. Arnhem, Cito.

[ref 3] Fleiss, J. L & P. E. Shrout (1978),
Approximate interval estimation for a certain intraclass correlation coefficient,
Psychometrika, 43, 259 – 262.

[ref 4] Shrout, P.E. & Fleiss, J.L. (1979),
Intraclass Correlations: Uses in Assessing Rater Reliability,
Psychological Bulletin, 2, 420-428.

B Transcriptieprotocol woordelijke transcripties

INSTRUCTIONS FOR DUTCH TRANSCRIBERS

B.1 < OVERVIEW >

Goal of Transcriptions:
The goal of the transcription effort is to provide an accurate word-level transcription of what the speaker said, with unambiguous markings for extraneous events and disfluencies. An accurate word-level transcription of these responses is the minimal requirement in order to make the data useful for training the automatic speech recognizer and score development. ACCURACY of transcription is PARAMOUNT and accuracy should not be sacrificed for speed, but speed is good too!

Trancribing Tools:
Transcribers use a web-based tool that brings up one sound file at a time. This sound file will be played automatically the first time. The sound file can be replayed by clicking on the arrow at the beginning of the player sliding bar. When clicked,the arrow changes into two vertical lines and remains thus while the sound file is playing.
To pause the sound file, click on the two vertical lines and they will change into the arrow again. Click on the arrow to resume playing the sound file. It is also possible to select particular sections of a sound file by dragging the rectangular button along the player sliding bar. When these features are used, the sound file does not always play completely to the end. Please ensure that the complete sound file is being replayed by verifying that the rectangular button has reached the end of the player sliding bar.
The volume can be adjusted by left-clicking on the speaker icon and adjusting the slider bar.

Above the player sliding bar a text box appears. It is in this text box, henceforth called the Transcription box, that the transcription of the sound file should be entered. In order to make transcription easier and quicker, the text of the expected response is provided in the transcription box.

To help you determine the content of the sound file, the original prompt (i.e., the text that the speaker saw or the speech that the speaker heard) corresponding to the current sound file is given at the top of the screen in the Prompt box. Depending upon the task, the text in the Prompt box and the Transcription box may or may not be the same.
For example, if the task is to repeat a sentence, they will be the same, but if the task is to answer a question, they will be different. Use the text in the Transcription box as the starting point for your transcription, making changes to the text to reflect what is actually said.

The transcription tool does NOT allow you to save partial transcriptions. If you have only partially transcribed a sound file and need to quit the file before completing it, please copy the transcription you have done to a temporary working file and save the file. The next time you come across the sound file in your Transcription Set, you can cut the transcription from your working file and paste it into the Transcription box.

To submit your transcription:
When you are satisfied that your transcription is accurate, it is a good idea to listen to the sound file one final time before submitting your transcription. To submit the transcription, click on the Enter button. Once you have clicked on the Enter button in the Transcription tool, the transcription is saved as is. Please ensure that you have fully transcribed the sound file BEFORE clicking on the Enter button. If you have submitted your transcription and realize that you made a mistake, you can amend your transcription by clicking on the << button, amending your transcription and clicking on the Enter button again.

To skip a particular sound file:
If you don't know how to transcribe a particular sound file, you may skip this sound file and go to the next sound file by clicking on the >> button. If you choose to move on to the next sound file by using the >> button, any partial transcription of the skipped sound file will NOT be saved.

To choose a new set:
You can select a different set of sound files by clicking the Choose New Set button.
To end your transcription session, click the End Grading button.

To end the transcription session:
To end your transcription session, click the End Grading button.

Time out:
For security reasons, the system will automatically log you off if the Transcription tool has been idle for more than 15 minutes, i.e. you have not submitted a transcription within 15 minutes of listening to a sound file. Please be aware that if you are in the middle of transcribing a sound file and are interrupted, your partial transcription may be lost if you are away for more than 15 minutes.

Any problems or questions:
If you come across a problem in a file and need help, or if you are not sure about how to transcribe a file, please contact your designated contact person. Please specify the sound file I.D. and give a brief description of the problem. To obtain the sound file I.D., left click on the Clip button at the bottom below the Transcription box. Make sure you include the file I.D. when you contact your assigned contact person. This will enable us to listen to the response with which you are having a problem.

B.2 < INSTRUCTIONS >

The transcriber's task is to amend the text which is present in the Transcription box according to the rules given below, in order to obtain an accurate word-level transcription of the response, including extraneous events and disfluencies.
If there is **no response** at all, leave the Transcription box blank, i.e., delete whatever text is there. However, if any audible breath or noise is heard, this should be indicated by using one of the extraneous or non-speech event symbols below. If the utterance is spoken over an intrusive and continuous background noise (someone else speaking, music or traffic in the background...) or if the recording is very poor, mark the fact by checking a box below the Transcription box that has the appropriate description of what you hear.
Transcriptions should specify exactly what was said by the speaker using **EXISTING** Dutch words or one of the **SYMBOLS** described below.

B.3 Case

All transcription is done using lowercase.

B.4 Punctuation

No punctuation is used, except for the apostrophe. *'s avonds*, *zo'n*, *'t is* and so on are all legitimate transcriptions.

Compound words are either transcribed as a single compound word, if that is how they are usually written or split into two separate words. No hyphenation is used, even though in normal written Dutch a hyphen (-) is used. EXAMPLE: *vijfhonderduizend, snelweg*, but *west europese, zee engte*.

B.5 Reduced forms, Abbreviations, Acronyms, Spelled Words

Reduced forms are written with apostrophes. If the Text prompt shows ‘ 'k moet hier weg’ and the speaker says ‘'k moet hier weg’, you write
If ‘kinderen’ is heard as ‘kienderen’, write *kinderen*.
If ‘kleuter’ is heard as ‘klooter’, write *kleuter*.

- Consonants may sound ‘foreign’. Some Dutch consonants and consonant clusters are difficult to pronounce if they don’t occur in the speaker’s own language.
  For example, the Dutch consonant cluster ‘schr’ is very difficult to pronounce for many non-native speakers Thus, this pronunciation of *schrijven* should be considered a variant and not a mispronunciation.
  Similarly, some non-natives (and natives) may pronounce final ‘-ng’ as ‘-ngk’. Thus, the pronunciation of *vertraging* should be considered a variant and not a mispronunciation.
  However if the speaker clearly mispronounces the consonants in the word, it should be considered mispronunciation. If ‘chronic’ is mispronounced as ‘shronic’, write **chronic*
- Non-native speakers (as well as many native speakers) may not pronounce the final -n in plural forms of nouns or in verbs. Such omissions of final -n will not be marked as a mispronunciation.
  Thus, even though the speaker does not pronounce the final -n in ‘bladeren’, write *bladeren*.

B.6 Unintelligible speech

Unintelligible speech is marked with a single instance of the marker @. The marker @ is entered in the transcription in place of whatever speech was present, regardless of the length of the unintelligible or untranscribable segment. If the speech is unintelligible because it is in a language other than Dutch, transcribe the unintelligible (non-Dutch) portion with the marker @.
EXAMPLE: The speaker was supposed to say ‘De bal kwam naast het doel terecht’, but part of what the speaker says is unintelligible. Therefore, the transcription of this particular example should be de bal kwam naast @ .

B.7 Non-Speech Events

We transcribe two kinds of non-speech events: mouth noise and hesitations.

Used to mark any mouth or nose noises on the part of the speaker. Place the # symbol in the transcription exactly at the place where the event occurs. Mouth noises are defined as any non-speech noises from the speaker, excluding verbalized hesitations. Included in this category are breath noises (inhale and exhale), tongue clicks, lip smacks, throat clearing, snorts, sniffs, sneezes, coughs, and laughs, and any combination of the above. However, breaths, tongue clicks and lip smacks that are significantly lower (quieter) than the speaker's speech do not need to be marked.
Example 1: *# nat*
Example 2: *vinger #*

Uh Used to mark all verbal hesitations. Again, place the uh symbol in the transcription exactly at the place where the event occurs. All verbalized hesitations, whether the speaker actually says 'uh', or says something different, such as 'um', 'mm', 'eh', etc., are marked as *uh*. **There is no attempt made to distinguish between the different possible verbalized hesitations, or to characterize them on a phonetic level.** Duration is also not considered: verbalized hesitations of any length are marked simply with one uh marker.
Example: *# uh hoe lang daar uh heb je last van*

Note: By definition, *#* and *uh* should **NEVER** co-occur with speech. They should simply be inserted at the place where they occur in the utterance.

B.7.1 Extraneous Events

Extraneous events are marked using square brackets enclosing a descriptor for the type of event. Four different event descriptors are used: [N], [S], [R] and [!]. Each of these is defined in detail below:

[N] Used to mark background noise. Background noise from any source is transcribed as [N]. Possible noise sources include but are not limited to: dogs barking, bird or other pet noises, finger tapping, and TV or radio noise (including TV or radio speech). [N] is a 'catch-all' marker that is used for any noise that is notable but does not fall into one of the other categories below.

[S] Used to mark background speech. Background speech is defined as audible speech from other talkers near the speaker. Audible speech from other speakers is to include any speech that is loud enough to be identified as speech, whether the words can be made out or not. It may also include laughter that is part of a normal conversation. Do not attempt to transcribe background speech, simply indicate that it is present with the marker [S]. Isolated laughter (from people in the background) should be marked as [N]. Note that laughter from the speaker is covered by the marker #, as described below.

[R] Used to mark line noise or artifacts of recording. This is any noise that sounds like it is due to the telephone connection or recording system such as clicks, pops, ring crosstalk, etcetera. This is NOT used for microphone noise.

[!] Used to mark any unusual or unresolvable recording, which can't be transcribed in any way by using all the other symbols.

Single Events or Continuous Events Over a Few Words

In the case of single events, [N], [R] and [S] may occur before the first word, after the last word or between two words without co-occurring with speech at all.

Example 1 (Background Noise): In this example, there is some background noise immediately after the last word: *it's raining [N]*
Example 2 (Background Speech): In this example, there is some background speech after the last word: *this is it is the desk [S] #*

In addition to the independent occurances as described above, [N], [R] and [S] may co-occur with speech. They may occur as single events that coincide with the speaker's production of a single word, or they may span several words. In this case the response should be transcribed by placing markers such that they indicate the words over which the noise or speech occurs.

Example 3 (Single Event): If the speaker was reading the sentence 'put the table outside' and a dog barked right in the middle of the word 'table', the following would be a correct transcription: *put the [N>] table [<N] outside*

Example 4 (Continuous Event Over Words): In the sentence shown above, if there was background speech starting during the word table and continuing throughout the rest of the utterance, the transcription would be: *put the [S>] table outside [<S]*

Example 5: In this example, noise can be heard starting at 'to' in the sentence 'If I want to see it I have to get up…' and stops after 'and' in the sentence 'It is red and blue…'. The transcription would be: *i have to [N>] get up ... ... it is red and [<N] blue*

Note: The two end markers [>] [<] should always be used in matched pairs.

Continuous, All-File-Long Events

If audible events occur over the course of the whole file (thus during the entire utterance), mark one or more of the check-boxes on the transcription window. The check boxes are:
Continuous background noise [N]
Continuous background speech [S]
Poor recording, or continuous line noise [R]
Unusual or unresolvable recording [!]

Table C.1 A summary table of the symbols.

Symbols	Meanings	Examples
-	A part of a word spoken is missing; it can be the beginning or the end of a spoken word. The use of the hyphen "-" is reserved to indicate this **incomplete word**. Mark the start or end of an incomplete word by enclosing the missing remainder of the word in parentheses, and terminating the missing part between the parentheses with the "-" symbol.	*innocent uh noninno(cent)- noninnocent* *-(an)swer*
=	The equals sign "=" is reserved to indicate a spoken **word truncated** at the end of a sound file. If the very last word of a sound file is cut off, add the "=" symbol to the end of the truncated word.	*... five=*
*	**Mispronunciations** should be represented as the intended word, with one star (*) immediately preceding the mispronounced word.	*when i go *grow up*
@	Any stretch of **unintelligible** or undecipherable speech is represented as @.	*twenty @ years old*
uh	Any verbalized **hesitations** from the speaker's side such as "um", "mm", "eh", etc. Do not attempt to distinguish different possible hesitation sounds. Just use *uh* for any hesitation sound.	*i want uh a cake*
#	Any mouth noises from the speaker's side such as breath noise (inhale and exhale), tongue clicks, lip smacks, throat clearing, sniff, sneezes, coughs, and laughs.	*# sixteen eighteen #*
N	Background noise; could be a single event or a continuous event. Check the appropriate box below the Transcription box if it is a continuous event.	*i like* [N] *pasta* (a single event)
S	Background speech; could be a single event or a continuous event. Check the appropriate box below the Transcription box if it is a continuous event.	*he stole* [S] *my new bag* (a single event)
R	Line noise/recording noise; could be a single event or a continuous event. Check the appropriate box below the Transcription box if it is a continuous event.	*i* [R] *like my new car* (a single event)

Note: To the extent that non-speech events and disfluencies are noted, the conventions for marking have been adapted from guidelines used to transcribe the MACROPHONE database.

C Beoordelingsprotocol uitspraak

C.1 Descriptoren voor Uitspraak

Definitie van Uitspraak.

- Vaardigheid klinkers, medeklinkers en klemtoon in zinsverband te produceren als een moedertaalspreker;
- beheersing van de fonologie (klanken en klemtoon) van alledaagse woorden.

Algemene scoringsregels.

- Bij TWIJFEL tussen twee scores: geef altijd de laagste;
- NUL wordt gebruikt voor STILTE of voor een IRRELEVANT of totaal ONBEGRIJPELIJK antwoord.

Toelichting

- C en V staan respectievelijk voor Consonant (medeklinker) en Vocaal (klinker).

C.2 Scoreschaal voor Uitspraak

6	'MOEDERTAAL' Uitspraak. Alle C's en V's worden uitgesproken zoals een moedertaalspreker dat doet. De spraak is onmiddellijk en met zekerheid verstaanbaar. De spreker gebruikt reductie, assimilatie en weglating van klanken zoals gebruikelijk is in vlot taalgebruik van moedertaalsprekers. Klemtonen zijn correct.
5	GEVORDERDE Uitspraak. Uitspraak van C's en V's is helder en ondubbelzinnig. Enkele afwijkingen in klank of klemtoon zonder gevolgen voor de begrijpelijkheid. Ieder woord kan gemakkelijk worden verstaan. Klemtonen in gangbare woorden zijn correct.
4	REDELIJK GOEDE Uitspraak. Uitspraak van de meeste C's en V's is correct. Sommige consistente uitspraakfouten leiden bij een aantal woorden mogelijk tot onduidelijkheid. Een paar C's of V's worden mogelijk in bepaalde contexten regelmatig verkeerd uitgesproken of weggelaten. Door het uitspreken van de stomme e of andere klinkers die in lopende spraak horen weg te vallen, wordt het klemtoonpatroon mogelijk verstoord.
3	MATIGE Uitspraak. Bepaalde C's en V's worden consistent verkeerd uitgesproken. Spraak is over het algemeen verstaanbaar, maar de toehoorder moet wel wennen aan het accent. Bovendien worden sommige C's regelmatig vervormd of weggelaten en worden consonant clusters vereenvoudigd. De klemtoon is bij sommige woorden verkeerd geplaatst, of onduidelijk.

Scoreschaal voor Uitspraak (vervolg)

2	STORENDE Uitspraak. Veel C's en V's worden verkeerd uitgesproken, waardoor een sterk buitenlands accent ontstaat dat het begrip hindert. De toehoorder kan mogelijk een belangrijk deel van de woorden (≥ 33%) niet verstaan. Bovendien worden veel C's vervormd of weggelaten en worden veel consonant clusters vereenvoudigd. De plaatsing van klemtonen is onduidelijk, onbeklemtoonde V's worden niet verkort of juist weggelaten, soms wordt een hele lettergreep toegevoegd of weggelaten.
1	SLECHTE Uitspraak. De uitspraak is eigenlijk volledig die van een andere taal. Veel C's en V's worden verkeerd uitgesproken, verhaspeld of weggelaten. De toehoorder zal in het begin nauwelijks iets kunnen verstaan. Er is weinig verschil tussen beklemtoonde en onbeklemtoonde lettergrepen. Meerdere woorden hebben niet het juiste aantal lettergrepen.
0	GEEN EVIDENTIE. Mogelijk wel Nederlands, maar door zeer slechte uitspraak niet verstaanbaar ook niet met moeite. Of: Stilte, geen Nederlands, of een enkele zucht of spraakklank.

D Beoordelingsprotocol vloeiendheid

D.1 Descriptoren voor Vloeiendheid

Definitie van Vloeiendheid.

- Vloeiend en snel kunnen spreken hetgeen blijkt uit passend ritme, frasering, pauzering en woordklemtoon in doorlopende spraak. 'Het loopt en klinkt zoals het hoort'.

Begripsdefinitie.

- Een langere uiting is een uiting met negen of meer woorden (≥ 9 woorden).

Algemene scoringsregels.

- Bij TWIJFEL tussen twee scores: geef altijd de laagste.
- NUL wordt gebruikt voor STILTE of voor een IRRELEVANT of totaal ONBEGRIJPELIJK antwoord.

D.2 Scoreschaal voor Vloeiendheid

6	'MOEDERTAAL' Vloeiendheid. De uiting van de kandidaat vertoont vloeiend ritme en frasering als van een moedertaalspreker, zonder aarzelingen, herhalingen, valse starts of onnatuurlijke reductie van fonemen.
5	GEVORDERDE Vloeiendheid. De uiting van de kandidaat heeft een aanvaardbaar ritme met de juiste frasering en woordklemtoon. Uitingen bevatten hooguit een enkele hapering, herhaling of valse start. Opvallend onnatuurlijke fonologische reductie komt niet voor.
4	GOEDE Vloeiendheid. De uiting van de kandidaat heeft een aanvaardbaar tempo, maar kan wat onevenwichtig klinken. Langere uitingen kunnen meer dan een enkele aarzeling bevatten, maar de meeste woorden worden in lopend zinsverband uitgesproken. Er zijn weinig herhalingen of valse starts. Er vallen geen lange pauzes en de spraak klinkt niet stoterig.
3	MATIGE Vloeiendheid. De uiting van de kandidaat kan wat onevenwichtig of stoterig klinken. Uitingen met 6 of meer woorden bevatten tenminste een opeenvolging van drie vloeiend lopende woorden en niet meer dan twee of drie aarzelingen of valse starts. Er kan een langere pauze voorkomen, maar niet twee of meer.
2	STORENDE Vloeiendheid. De uiting van de kandidaat vertoont onregelmatige frasering of zinsritme. Gebrekkige frasering, stoterig verloop - per lettergreep - en/of veelvuldige aarzelingen, herhalingen of valse starts maken de uiting 1duidelijk onevenwichtig en onderbroken. Langere u0itingen kunnen een of twee lange pauze bevatten en vertonen mogelijk onjuiste woord- of zinsklemtoon.
1	NIET VLOEIEND. De spraak van de kandidaat is langzaam en verloopt moeizaam, met nauwelijks enige frasering en veelvuldig aarzelingen, herhalingen, valse starts, en/of grove fonologische reducties. De meeste woorden worden los uitgesproken en er kunnen meer lange pauzes voorkomen.
0	GEEN EVIDENTIE. Mogelijk wel Nederlands, maar door het vele horten en stoten niet te volgen ook niet met moeite. Of: Stilte, geen Nederlands, of alleen gezucht en wellicht een enkele spraakklank.

E Procedure kwalitatieve beoordelingen

PROCEDURE:
Gebruik een normale telefoon (niet draadloos of mobiel) met toon (niet puls).

- Bel: 0800-0223558.
- Een stem verwelkomt je in het Nederlands vervolgens vraagt een tweede stem om je identificatiecode (TIN) in te toetsen. Het systeem gaat er in eerste instantie nog vanuit dat je een toets wilt maken.
- Toets je PIN in.
- Het systeem herkent de code van een beoordelaar en gaat over op Engels: 'Grading, for help enter pound (=hekje #) twice'.
  Vervolgens noemt het systeem het beoordelingsaspect: 'Pronunciation' of 'Fluency' of 'Council of Europe' of 'Story Telling'. En de mogelijke scores: 'Grades are 0 thru 6. etcetera'.
- Pak de betreffende descriptoren erbij.
- Je hoort meteen een stimulus, en vervolgens een aantal kandidaten die op dezelfde stimulus reageren.
- Beluister er eerst een paar zonder een score in te toetsen. Bedenk wel een score maar toets vervolgens op ** (tweemaal ster), dan wordt de kandidaat weer op de stapel gelegd en komt later weer terug.
- Als je denkt dat je vertrouwd bent geraakt, toets dan je score op de telefoon.
- Je kunt een respons onmiddellijk opnieuw horen door '7' te toetsen.
- Je kunt altijd stoppen met beoordelen door gewoon op te hangen, het systeem onthoudt waar je was.
- Wanneer je iets vreemds hoort, of je wilt over een bepaalde respons later een opmerking maken, toets dan 8 en noteer 'call' (kandidaat), section (toetsonderdeel) en item.

Beoordelingen: Niveaus en codes.

Niveaus Uitspraak & Vloeiendheid	Toets telefoon
C2	6
C1	5
B2	4
B1	3
A2	2
A1	*1
A1-min	1
< A1-min	*0
GEEN EVIDENTIE	0

Beoordelaarsmenu

1	*2*	*3*
4	*5*	*6*
herhaal *7*	identificeer *8*	aspect *9*
overslaan *	*0*	#

0 - 6	Scores 0 t/m 6
# of *	zijn voorvoegsels
##	Help
7	herhaal laatste response
#7	herhaal stimulus en response
8	identificeer: nummer kandidaat, onderdeel en item
#8	verwijder laatst ingetoetste score
9	identificeer: beoordelingsaspect
#9	identificeer: scoringscategorieën
**	sla deze response over
#**	sla dit item over
#*#	sla dit onderdeel over
#*#	sla dit onderdeel over

F Beoordelingsprotocol gespreksvaardigheid

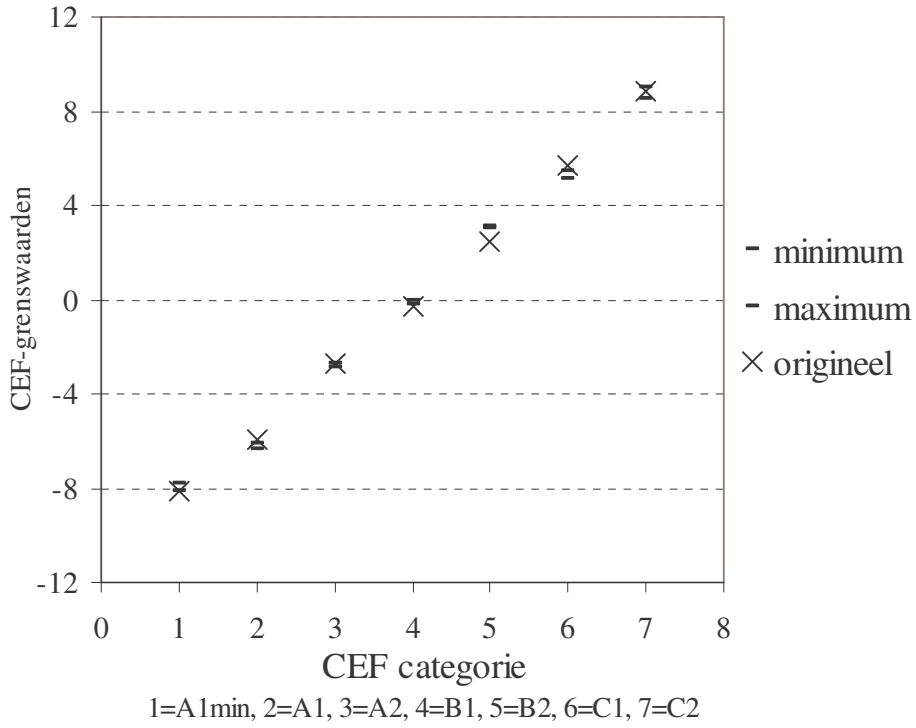
Beoordelingsschaal Gespreksvaardigheid op de CEF-Niveaus		
C	C2	***Brengt betekenisnuances nauwkeurig en op natuurlijke wijze over*** Kan spontaan en met een natuurlijke vloeiendheid ook langere interventies verrichten. Vertoont daarbij een consistente grammaticale en fonologische beheersing van gevarieerd en complex taalgebruik met inbegrip van een juist gebruik van verbindingswoorden en voegwoorden. Kan moedertaalsprekers moeiteloos verstaan.
	C1	***Drukt zich vloeiend en spontaan uit in duidelijke, goedgestructureerde spraak.*** Kan zich spontaan en vloeiend uitdrukken, bijna moeiteloos in een gelijkmatig lopend taalgebruik. Heeft een duidelijke en natuurlijke uitspraak. K*an* intonatie *var*iëren en gebruikt klemtoon om delen te benadrukken. Maakt zelden fouten. Vertoont beheersing van verbindingswoorden en voegwoorden. Verstaat praktisch iedere moedertaalspreker, maar moet wellicht soms om bevestiging vragen.
B	B2	***Brengt informatie en standpunten helder en zonder merkbare moeite over.*** Kan eenheden taal met een redelijk evenwichtig tempo produceren met weinig merkbare pauzes. Heldere uitspraak en intonatie. *Fouten leiden niet tot misverstande*n. *Helder, samenhangend betoog*, echter soms enigszins “springerig”. Kan in detail standaard moedertaalsprekers ook in een lawaaierige omgeving verstaan.
	B1	***Communiceert begrijpelijk de belangrijkste punten m.b.t. vertrouwde zaken.*** Kan op begrijpelijke wijze doorspreken, hoewel evident pauzerend voor het plannen en herstellen van grammaticale en lexicale elementen Uitspraak is begrijpelijk hoewel bij tijden gekleurd door een buitenlands accent en ook uitspraakfouten optreden. Redelijk correct gebruik van een algemeen repertoire in voorspelbare situaties. Kan eenvoudige losse elementen verbinden tot een samenhangend geheel. Kan duidelijk sprekende moedertaalsprekers volgen, maar moet soms om herhaling vragen.
A	A2	***Communiceert basisinformatie over werk, achtergrond, familie, vrije tijd, etc.*** Kan zichzelf in korte zinnen verstaanbaar maken, hoewel pauzes, valse starts, en herformuleringen evident aanwezig zijn Uitspraak is over het algemeen helder genoeg om te worden verstaan ondanks een duidelijk buitenlands accent. Gebruikt een beperkt aantal eenvoudige structuren correct, maar maakt systematisch elementaire fouten. Kan woordgroepen verbinden met eenvoudige voegwoordeen zoals “en”, “maar”, en “omdat”. Kan zich tot hem/haar richtende, duidelijk sprekende moedertaalsprekers verstaan, wanneer zonodig om herhaling gevraagd kan worden.
	A1	***Doet eenvoudige uitspraken over persoonlijke gegevens en bekende onderwerpen.*** Kan omgaan met zeer korte, geïsoleerde, voornamelijk standaard uitingen. Veel pauzes om te zoeken naar uitdrukkingen en om minder bekende woorden uit te spreken. Spreekt met sterk buitenlands accent. Begrijpt de strekking van direct tot hem/haar gerichte en duidelijk gesproken vragen.

Onder A	**A1-min**	***Kan met behulp van losse woorden zaken van direct persoonlijk belang communiceren.*** Gebruikt losse woorden, enkele standaarduitdrukkingen en elementaire beleefdheids-frases maar is vanwege uitspraak moeilijk te verstaan. Begrijpt eenvoudige direct tot hem/haar gerichte en met zorg gesproken vragen naar of mededelingen over personalia, en een beperkt aantal concrete alledaagse begrippen. Kan vragen over dergelijke zaken soms ook met een of meer losse woorden beantwoorden. Conversatie is echter niet mogelijk
	***Onder*A1-min**	***Spreekt op een niveau dat lager is dan het bij A1-min beschreven niveau. Zou als toerist niet zonder hulp kunnen 'overleven'.*** Beheerst zo weinig woorden en/of uitdrukkingen dat verbale communicatie niet mogelijk is. Kan wellicht met veel hulp en begrip van de gesprekspartner enkele vragen naar eigen naam en adres of andere persoonlijke gegevens begrijpen. Kan dergelijke vragen echter meestal niet beantwoorden. Ook basishandelingen, zoals het maken van een afspraak of het geven of begrijpen van eenvoudige routebeschrijvingen, zijn niet mogelijk, hoewel soms met veel gebaren enige communicatie kan worden bereikt.

N.B. Ken bij twijfel tussen twee niveaus altijd het laagste niveau toe.

G Extra resultaten onderzoeksvraag 2a

In Figuur 6 zijn CEF-grenswaarden weergegeven gevonden in de originele en herhaalde analyses zonder te corrigeren voor verschuiving van de thetaschaal en itemmoeilijkheid.



Figuur 6: CEF-grenswaarden$_{FACETS}$ gevonden in originele en herhaalde analyses zonder schaalcorrectie

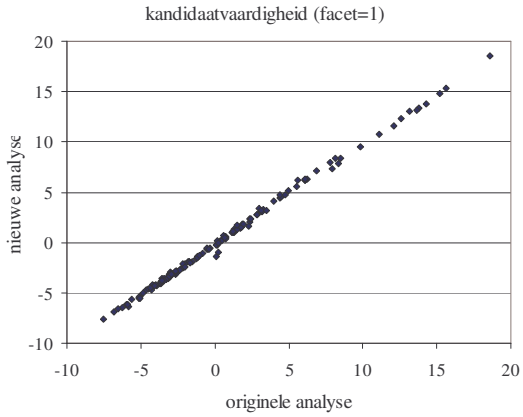
In Tabel 18 zijn de percentages kandidaten per CEF-categorie weergegeven.

Tabel 18 Percentages kandidaten per CEF-categorie

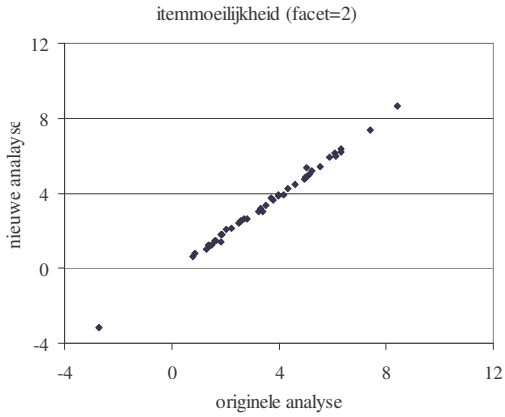
	gemiddeld	**FACETS ongecorrig**	**FACETS gecorrig**	**TGN ongecorrig**	**TGN gecorrig**
<A1min	15	0	19	7	22
A1min	26	11	22	9	22
A1	30	22	29	22	21
A2	15	25	17	21	13
B1	7	23	7	15	9
B2	2	10	2	12	5
C1	1	4	1	7	2
C2	3	5	4	7	5

In figuur 3 zijn de schattingen weergegeven van:
a kandidaatvaardigheid;
b itemmoeilijkheid;

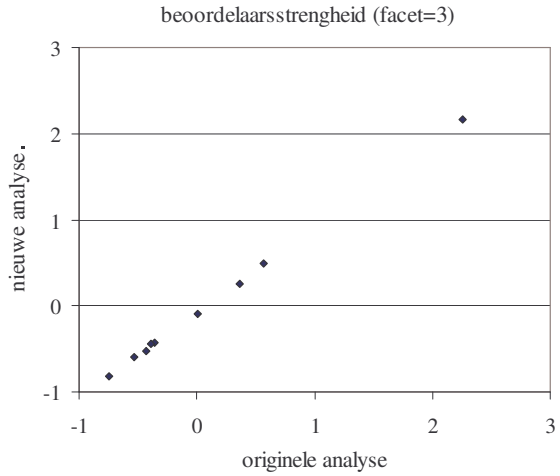
c beoordelaarstrengheid
voor de originele en voor één van de nieuwe analyses[25].



Figuur 3a Schattingen van kandidaatsvaardigheid in originele en nieuwe analyse.



Figuur 3b Schattingen van itemmoeilijkheid in originele en nieuwe analyse.



Figuur 3c Schattingen van beoordelaarsstrengheid in originele en nieuwe analyse.

De verschillen tussen de herhaalde analyses zijn zeer klein, daarom laten we hier slechts één analyse zien. Uit deze figuren blijkt dat de vaardigheidsschattingen zeer goed overeenkomen voor de gemeenschappelijke kandidaten, kandidaten en beoordelaars. Dus de foutieve uitvoering van de koppelingsmethode heeft geen gevolgen gehad voor de relevante uitkomsten.

Door het consortium is opgemerkt dat zonder de interviews het onderscheidend vermogen op de lage niveaus wellicht niet groot genoeg is. De ‘open vragen’ en ‘verhalen navertellen’ lijken namelijk minder geschikt te zijn voor lage taalvaardigheidniveaus.

Om na te gaan of het onderscheidend vermogen op de lage niveaus groot genoeg is, staan in tabel 13 de aantallen en de outfit per categorie vermeld voor de originele

[25] Koppeling via gelijkstellen van ‘verhaaltje 1’ en ‘open vraag 2’.

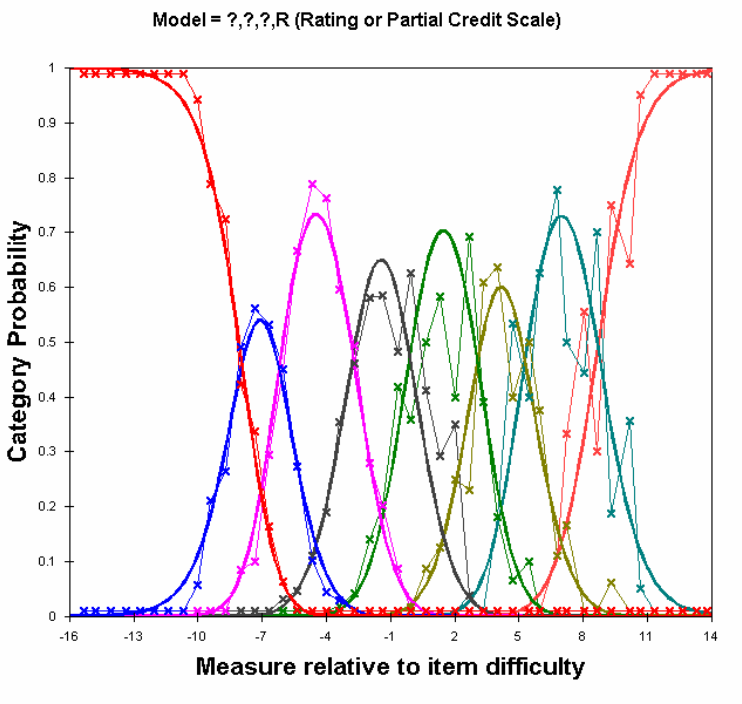
analyse en één van de nieuwe analyses[25]. Daarnaast wordt de standaard error (s.e.) gegeven voor de geschatte grenswaarde. De statistieken van de nieuwe data laten zien dat er geen reden is om aan te nemen dat het onderscheidend vermogen op de lage niveaus niet groot genoeg is.

- De fit op lage niveaus is vergelijkbaar met de fit op hoge niveaus.
- De standaard error is kleiner onderin de schaal.

Tabel 19 Aantallen en fit-statistieken voor originele en nieuwe analyse.

	originele analyse			nieuwe analyse		
	aantal	outfit	s.e.	aantal	outfit	s.e.
<A1min	476 (12%)	1,0		399 (14%)	1,0	
A1min	842 (21%)	0,8	0,07	564 (19%)	0,8	0,08
A1	1349 (34%)	0,9	0,05	1058 (36%)	0,9	0,06
A2	744 (19%)	1,1	0,06	555 (19%)	1,1	0,06
B1	301 (8%)	1,1	0,08	197 (7%)	1,2	0,10
B2	105 (3%)	1,3	0,15	60 (2%)	1,2	0,19
C1	53 (1%)	0,9	0,25	53 (2%)	0,9	0,26
C2	50 (1%)	1,1	0,27	50 (2%)	1,1	0,27

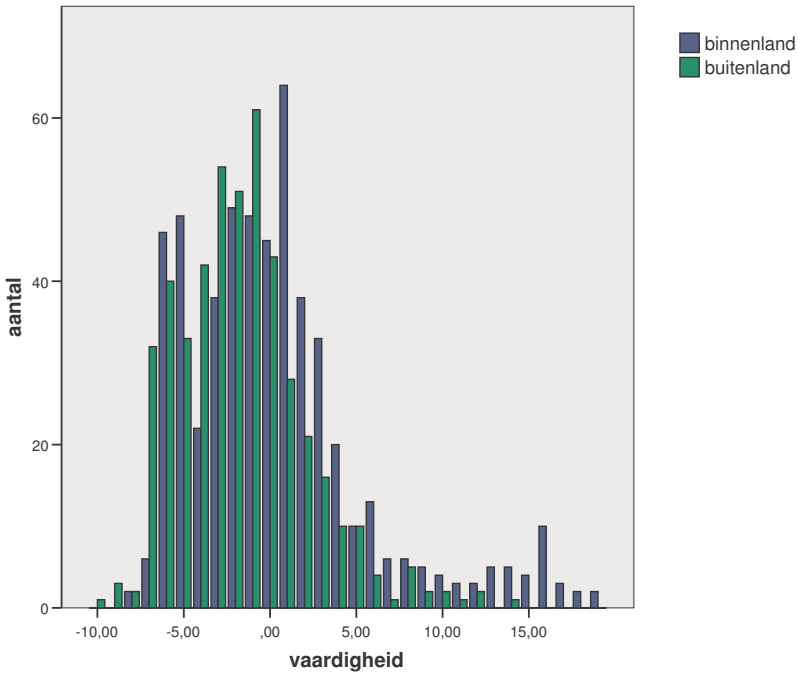
Ook grafische weergave van de modelfit (zie figuur 3) laat zien dat er geen reden is om het onderscheidend vermogen op de lage niveaus in twijfel te trekken.



Figuur 7 Modelcurves en geobserveerde data (x) per .65 interval.

H Extra resultaten onderzoeksvraag 2b

Figuur 8 laat zien dat er weinig verschil is tussen de verdeling van kandidaat-vaardigheden in binnen- en buitenland[26],



Figuur 8 verdeling vaardigheid kandidaten in binnen- en buitenland

[26] Gebaseerd op een gekoppelde analyse met alle data uit binnenland (koppeling 'open vraag 2' en 'verhaaltje 1' + zonder interviews) en de data verzameld in het buitenland.