Kampweg 5
Postbus 23
3769 ZG Soesterberg

www.tno.nl

T +31 346 356 211
F +31 346 353 977
info@tm.tno.nl

## TNO-rapport

**TNO-DV3 2005 098**

# Second opinion ten aanzien van de validatie van de spraaktechnologie gebruikt bij het inburgeringexamen

| | |
|---|---|
| Datum | 24 oktober 2005 |
| Auteur(s) | Judith Kessens, Sander van Wijngaarden & David van Leeuwen |
| Opdrachtgever | Ministerie van Justitie |
| Projectnummer | 013.75144 |

| | |
|---|---|
| Rubricering rapport | |
| Titel | Ongerubriceerd |
| Samenvatting | Ongerubriceerd |
| Rapporttekst | Ongerubriceerd |
| Bijlagen | Ongerubriceerd |

| | |
|---|---|
| Aantal pagina's | 53 (incl. bijlagen) |
| Aantal bijlagen | 8 |

# Voorwoord

**Inhoud rapport**

Volgens een nieuw wetsvoorstel zal het inburgeringsexamen voor vreemdelingen die zich in Nederland willen vestigen worden afgenomen met behulp van een spraakherkenner. Het gaat om twee toetsen:

- De “Toets Gesproken Nederlands (TGN)” die meet of een kandidaat over voldoende mondelinge taalvaardigheid beschikt,
- De toets “Kennis van de Nederlandse Samenleving (KNS)” die meet of een kandidaat over voldoende kennis over de Nederlandse samenleving en cultuur beschikt

In dit rapport beschrijft TNO een second opinion ten aanzien van de validatie van de gebruikte spraaktechnologie in beide toetsen. De second opinion is gebaseerd op resultaten van validatiestudies (niet door TNO uitgevoerd) die aan TNO ter inzage zijn gegeven. Het examen wordt in deze onderzoeken vanuit diverse invalshoeken (onderwijskundig, maar ook spraaktechnologisch) gevalideerd. Daarnaast heeft TNO op basis van de beschikbare data een aantal aanvullende validatiestudies uitgevoerd.

**Opzet rapport**

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 conclusies en aanbevelingen (blz. 3+4) en de toelichting op de conclusies en aanbevelingen (blz. 5 t/m 8).
- Door het lezen van het rapport zonder appendices wordt een completer beeld van de uitgevoerde second opinion verkregen. Hierbij is het mogelijk om een verkorte versie van hoofdstuk 4 te lezen door paragraaf 4.3 t/m 4.5 te vervangen door paragraaf 4.7.

Door het lezen van het rapport inclusief appendices wordt het meest complete beeld van de second opinion verkregen.

# Conclusies en Aanbevelingen

**Conclusies met betrekking tot Toets Gesproken Nederlands (TGN)**

Op basis van de validatiestudies worden door TNO de volgende conclusies getrokken:

- Er is evidentie gevonden dat de toets voldoende betrouwbare (=consistente) oordelen geeft
- Kandidaten met een sterk buitenlands accent worden door de spraakherkenner niet benadeeld met lagere deelscores
- Het trainingsmateriaal waar de spraakherkenner mee getraind is, is voor de meest frequente taalachtergronden representatief voor de praktijk

MAAR

- Er is geen bewijs gevonden dat de kwaliteit[1] van de toets voldoende is, noch dat de kwaliteit onvoldoende is. Er zijn aanwijzingen dat de toets op een zinnige manier taalvaardigheid meet, maar hiermee is niet gegarandeerd dat de kwaliteit van de toets voldoende is voor alle taalvaardigheidsniveaus.
- TNO heeft aanwijzingen gevonden dat de kwaliteit van de machinale oordelen minder goed is rond de A1min cesuur. Doordat gegevens omtrent de kwaliteit van vergelijkbare menselijke taalvaardigheidsbeoordelingen ontbreken ('human benchmark') is het niet mogelijk om te concluderen of de fouten die rond de A1min cesuur gemaakt worden acceptabel zijn.

**Conclusies met betrekking tot de toets Kennis Nederlands Samenleving (KNS)**

Op basis van statistische analyse van een kleine testset worden door TNO de volgende conclusies getrokken:

- Op basis van een statistische analyse van de kleine set test gegevens (N=59) wordt geschat dat voor 10-15% van de kandidaten de toetsuitslag onterecht is.
- Voor een nauwkeurigere schatting van de foutenpercentages is een grotere dataset nodig.
- Omdat gegevens van vergelijkbare menselijk afgenomen examens ontbreken en geen kwaliteitsnormen zijn gesteld kan niet worden geconcludeerd dat 10-15% (te) veel of (te) weinig is.

**Gebruik en reikwijdte van de conclusies**

De bovenstaande conclusies hebben betrekking op de prestatie van de spraaktechnologie in het testsysteem en de validatie van de cesuurinstelling van de toets. De kwaliteit van de spraaktechnologie wordt als voldoende beschouwd als de fouten die de geautomatiseerde toets maakt vergelijkbaar zijn met de fouten die menselijke examinators maken; zowel de oordelen van een geautomatiseerde toets als de oordelen van menselijke beoordelaars zullen nooit 100% foutloos zijn. Echter, omdat onbekend is welk percentage fouten in deze situatie door mensen gemaakt worden, kan niet worden geconcludeerd dat de spraaktechnologie beter, even goed of slechter presteert dan een menselijke examinator. De geschiktheid van het systeem als alternatief voor menselijke examinators kan daarom niet worden aangetoond, de ongeschiktheid ook niet.

De conclusies betreffen het te verwachten percentage ten onrechte zakken of ten onrechte slagen. Met dit percentage alleen is nog niet te zeggen of de toets acceptabel is. De uitgangspunten die beleidsmatig gaan gelden betreffende mogelijkheden om in

[1] Een ander terminologie die in dit rapport gebruikt wordt is *validiteit* (zie appendix A voor een definitie)

beroep te gaan tegen de uitslag, mogelijke second opinion van twijfelgevallen of het kiezen van de cesuur niveaus bepalen uiteindelijk of het te verwachten percentage acceptabel zal zijn.

**Eindconclusie**

- Op basis van de uitgevoerde testen bestaat de indruk dat de TGN toets een consistente uitslag geeft.
- Op basis van de uitgevoerde testen kan niet worden geoordeeld of de TGN toets wel of niet geschikt is voor gebruik, omdat kwaliteitsnormen niet zijn gesteld.
- Op basis van een statistische analyse van een kleine set testgegevens wordt geschat dat voor 10-15% van de kandidaten de KNS toetsuitslag onterecht is. Of dit foutenpercentage acceptabel is hangt af van de beleidsmatige uitgangspunten.

Er zijn mogelijkheden om zowel de toets, de cesuurinstelling, als de validatie te verbeteren. De noodzaak hiertoe hangt af van de beleidsmatige uitgangspunten en de richting is aangegeven in de aanbevelingen

**Aanbevelingen**

Op basis van de uitgevoerde evaluatie beveelt TNO aan over te gaan tot invoering van de toetsen en daarbij de volgende stappen te ondernemen:

1) Verzamel voor de TGN en KNS onafhankelijke data die voldoende representatief is voor de praktijk
2) Formuleer succescriteria (normen) voor de TGN en de toets KNS. Daartoe is het nodig om meerdere menselijke beoordelingen te verzamelen:
   - Bepaal een ‘human benchmark’: Dit zijn menselijke beoordelingen die beschouwd kunnen worden als alternatief voor de automatische toets
   - Bepaal een referentieoordeel ten opzichte waarvan de foutpercentage van de toets en de ‘human benchmark’ berekend kunnen worden
3) Valideer de toetsen door de foutpercentages (onterecht zakken/slagen) op de nieuwe datasets te bepalen, en deze volgens het vastgestelde succescriterium te vergelijken met de ‘human benchmark’ (aanbeveling 2).
4) Verbeter de TGN rond de A1min cesuur

# Toelichting op conclusies en aanbevelingen

**Verschillen in de gebruikte TGN validatieprocedures tussen TNO en CINOP**

*Algemeen*
Het doel van de toepassing van spraaktechnologie in de TGN is de automatisering van de beoordeling van het taalvaardigheidniveau van een kandidaat (ten opzichte van een vastgesteld minimum). De kwaliteit[2] van de toets wordt als voldoende beschouwd als de gemaakte fouten te vergelijken zijn met de fouten van het niet-automatische alternatief: deskundige menselijk beoordelaars. Zowel de oordelen van een geautomatiseerde toets als de oordelen van menselijke beoordelaars zullen nooit 100% foutloos zijn.

Uit bestudering van de rapportage van CINOP wordt duidelijk dat de door CINOP gehanteerde validatie-procedures op enkele punten wezenlijk verschillen van de standaard-werkwijze voor spraaktechnologie. In toelichting hierop stelt CINOP dat de gevolgde aanpak gangbaar en geaccepteerd is voor toetsen en examens. Er is een duidelijk verschil in perspectief: TNO bekijkt de set toetsen als een product dat op spraaktechnologie gebaseerd is, en waarvan bewezen moet worden (op de algemeen geaccepteerde manier) dat de spraaktechnologie naar behoren werkt. CINOP benadert de validatie van de toetsen op dezelfde wijze als elke willekeurige toets zonder spraakherkenning. Daarnaast is er een verschil in de keuze van het evaluatiemateriaal. Zoals eerder uiteengezet wordt voor spraaktechnologisch onderzoek geëvalueerd op materiaal dat onafhankelijk is van training en ontwikkeling/optimalisatie van het systeem. In de door CINOP gehanteerde werkwijze is het trainingsmateriaal over het algemeen apart strikt gehouden, maar is het ontwikkel- en validatiemateriaal niet altijd duidelijk gescheiden. De reden dat dit niet altijd is gebeurd heeft te maken met de historische ontwikkeling van de toets. In de loop van de tijd zijn namelijk een aantal aanvullende studies uitgevoerd waarvoor drie extra datasets zijn (Den Haag, Amsterdam, MFA-FIT) verzameld.

Om de verschillen in aanpak te concretiseren worden beide aanpakken hieronder (enigszins vereenvoudigd) samengevat. De beschrijving van deze aanpak van CINOP is mede geformuleerd aan de hand van mondelinge uitleg door CINOP en LTS.

*TNO aanpak TGN validatie*
De aanpak van TNO is de standaard aanpak voor validatie van spraaktechnologie. Dit houdt in dat je een gefundeerde eis stelt aan de prestaties, en vervolgens bewijst dat aan die eis voldaan wordt. In dit geval kan zo'n eis bijvoorbeeld zijn: de spraakherkenner mag niet meer fouten maken dan een alternatief, conventioneel examen. Door bij dezelfde steekproef van proefpersonen scores te bepalen op de nieuwe toetsen en een conventioneel examen is een degelijk (statistisch gefundeerd) bewijs te leveren. Als zo'n bewijs ontbreekt is de validatie niet toereikend, ook al zijn er andere aanwijzingen dat de technologie wel degelijk goed functioneert. Er ligt een duidelijke en harde bewijslast bij de partij die de spraaktechnologie introduceert. TNO stelt dat bewijs volgens de hardere normen van spraaktechnologie in principe wel degelijk te leveren is, en de vastgestelde problemen (met name de genoemde vergelijking met resultaten van conventionele toetsen) op te lossen zijn. Het gebruik van spraaktechnologie maakt de toetsen wel degelijk principieel anders dan andere toetsen; er mag dus niet zonder meer worden aangenomen dat het validatie-recept van CINOP dezelfde garanties geeft voor de prestaties van de toetsen als bij conventionele toetsen.

[2] Een ander terminologie die in dit rapport gebruikt wordt is *validiteit* (zie appendix A voor een definitie)

*CINOP aanpak TGN validatie*
Ook CINOP stelt concrete eisen aan de prestaties (onder andere in de vorm van betrouwbaarheidsscores). Echter, de stelling is dat hard bewijs dat beter wordt gepresteerd dan een referentie-toets (met menselijke examinatoren) nauwelijks te geven is: strikt genomen is onbekend wat *daadwerkelijk* het taalvaardigheidsniveau van de proefpersonen is geweest. Als een kandidaat voor een conventionele toets zakt maar voor de TGN slaagt, is het alsnog mogelijk dat de TGN gelijk had en de menselijke examinator ongelijk; ook bij conventionele toetsen worden fouten gemaakt. In plaats van rigide bewijs, zoals in de spraakherkenning wordt vereist, wordt gezocht naar evidentie voor kwaliteit. Allerlei relevante aspecten van de toets worden tegen het licht gehouden, en aan tests onderworpen. Als hierbij geen problemen worden geconstateerd (prestaties beneden vastgestelde criteria) dan wordt gesteld dat de toets valide is.

**Keuze van maten voor vaststelling van kwaliteit van spraaktechnologie**

TNO stelt dat het bewijs omtrent kwaliteit van de spraaktechnologie in de TGN niet zou moeten zijn gebaseerd op basis van een correlatiematen, maar op analyse van zak/slaagpercentages van de kandidaten. Een hoge correlatiecoëfficiënt kan immers niet garanderen worden dat het aantal kandidaten dat onterecht zakt of slaagt laag is. Zowel TNO als CINOP hebben een analyse uitgevoerd waarbij het percentage ten onrechte gezakte en geslaagde kandidaten is geschat. Er is echter één belangrijk verschil; CINOP schat het percentage onterechte beslissingen op basis van een model, terwijl TNO het percentage onterechte beslissingen schat op basis van data. TNO noemt een beslissing terecht als deze overeenkomt met het meerderheids- of unanieme oordeel van drie menselijke beoordelaars. Zowel TNO als CINOP stellen dat de instelling van de cesuur een beleidskwestie is; de keuze hangt af van het belang dat door de opdrachtgever wordt gehecht aan de gemaakte typen fouten.

Op basis van de validatiestudies worden door TNO de volgende conclusies getrokken:

- Er is evidentie gevonden dat de toets voldoende betrouwbare (=consistente) oordelen geeft
- Kandidaten met een sterk buitenlands accent worden door de spraakherkenner niet benadeeld met lagere deelscores
- Het trainingsmateriaal waar de spraakherkenner mee getraind is, is voor de meest frequente taalachtergronden representatief voor de praktijk

MAAR

- Er is geen bewijs gevonden dat de kwaliteit van de toets voldoende is, noch dat de kwaliteit onvoldoende is. Er zijn aanwijzingen dat de toets op een zinnige manier taalvaardigheid meet, maar hiermee is niet gegarandeerd dat de kwaliteit van de toets voldoende voor alle taalvaardigheidsniveaus.
- TNO heeft aanwijzingen gevonden dat de kwaliteit van de machinale oordelen minder goed is rond de A1min cesuur. Doordat gegevens omtrent de kwaliteit van vergelijkbare menselijke taalvaardigheidsbeoordelingen (‘human benchmark’) ontbreken het niet mogelijk om te concluderen of de fouten die rond de A1min cesuur gemaakt worden acceptabel zijn.

De uitgevoerde analyses en de gebruikte data die zijn beoordeeld in deze ‘second opinion’ hebben een aantal nadelen. Ten eerste ontbreekt een ‘human benchmark’. Hiervoor zijn referentie-oordelen nodig: een oordeel dat het werkelijke taalvaardigheidsniveau van de kandidaat benadert. Deze referentiewaarden kunnen niet bestaan uit de bevindingen van slechts één enkele examinator per kandidaat; deze kan

zich immers ook vergissen. Echter, door gemiddeldes van meerdere oordelen te hanteren kan een nauwkeurige benadering van de daadwerkelijke taalvaardigheid van de kandidaat worden verkregen. Door nu zowel het machinaal verkregen oordeel als een menselijke oordeel dat op een conventionele manier is verkregen te vergelijken met exact hetzelfde referentieoordeel kan een eenduidig succescriterium gesteld worden. De mogelijke tegenwerping dat zelfs een gemiddelde van vele examinatoren nog niet noodzakelijk de waarheid oplevert is irrelevant: als het referentieoordeel maar geaccepteerd wordt als een goede benadering van het werkelijke taalvaardigheidniveau, dan mag elke afwijking van deze referentie als onterechte uitslag worden aangemerkt. Ten tweede is bijna al het materiaal in de studie gebruikt voor ontwikkeling van de toets (zoals het trainen van de modellen van het automatische scoringscomponent, het bepalen van de schalings- en normeringsparameters, het bepalen van de itembank enz.). De enige dataset die niet betrokken is in ontwikkeling van de toets zijn de Amsterdam data. Het nadeel van deze data is:

- Het gaat om slechts twee soorten menselijke beoordelingen, het is moeilijk om op basis van deze beoordelingen zowel een ‘human benchmark’ als een referentieoordeel te bepalen
- Het gaat om relatief weinig data (voor 94 kandidaten zijn drie menselijke en machinale beoordelingen aanwezig)
- Het is onbekend in hoeverre de data representatief zijn voor de praktijk (het is bijvoorbeeld aannemelijk dat rond de A2 cesuur in de praktijk meer kandidaten voorkomen met A2 niveau)

**Validatie aanpak toets KNS en het belang van onafhankelijke data**

Het doel van de toepassing van spraaktechnologie in de toets KNS is het automatisch bepalen of een antwoord correct is of niet. De toepassing van spraaktechnologie kan succesvol worden genoemd als de fouten die de spraakherkenner maakt voldoende klein zijn ten opzichte van (conventionele) alternatieve methoden om antwoorden te registreren.

De validatiestudie die is uitgevoerd voor de toets KNS is in eerste instantie niet correct uitgevoerd, omdat de validatie-set gebruikt is voor het trainen van het taalmodel van de spraakherkenner. Om deze reden heeft Ordinate de validatie nogmaals uitgevoerd voor een onafhankelijke validatie-set van 59 sprekers. Op basis van de door Ordinate geleverde data heeft TNO een validatie uitgevoerd waarbij het percentage onterecht zakken en slagen is berekend. Hierbij is aangenomen dat een kandidaat over voldoende kennis over de Nederlandse samenleving bezit als er een score van 80% of meer wordt gehaald op de toets KNS (op basis van de antwoorden die gescoord zijn op basis van menselijke beoordelingen). Echter, de voorgestelde cesuur is in tweede instantie gewijzigd in 70%. In de (kleine) testset komen weinig kandidaten voor die een toetsscore hebben rond de 70%. Op basis van statistische analyse van deze kleine, onafhankelijke testset worden door TNO de volgende conclusies getrokken:

- Op basis van een statistische analyse van de kleine set test gegevens (N=59) wordt geschat dat 10-15% van de kandidaten de toetsuitslag onterecht is.
- Voor een nauwkeurigere schatting van de foutenpercentages is een grotere dataset nodig
- Omdat gegevens van vergelijkbare menselijk afgenomen examens ontbreken en en geen kwaliteitsnormen zijn gesteld kan niet worden geconcludeerd dat 10-15% (te) veel of (te) weinig is.

**Aanbevelingen**

Zoals eerder gesteld wordt het door TNO wel degelijk mogelijk geacht om te bewijzen dat de kwaliteit van een toets voldoende is. Hierin bestaat in taalvaardigheidstoetsen nog geen traditie, simpelweg omdat de op spraaktechnologie gebaseerde taalvaardigheidstoets een nieuw fenomeen is. TNO stelt vast dat de garanties waarop toepassers van spraaktechnologie normaal gesproken kunnen rekenen in dit geval niet onmiddellijk te geven zijn. Op grond van de bevindingen beveelt TNO aan over te gaan tot invoering van de toetsen en daarbij de volgende stappen te ondernemen:

1) Verzamel onafhankelijke data die volledig representatief zijn voor de praktijk
   - Geluidsfiles van een ruime hoeveelheid toetsafnames die ontstaan gedurende het gebruik van de TGN
   - Geluidsfiles van een ruime hoeveelheid toetsafnames die ontstaan gedurende het gebruik van de toets KNS
   - Neem bij de dataverzameling van de TGN twee verschillende steekproeven (datasets). Voor dataset 1 wordt een willekeurige steekproef van de kandidaten genomen, voor dataset 2 worden alleen kandidaten met een taalvaardigheidsniveau rond de cesuur geselecteerd. Dataset 1 is volledig representatief is voor de praktijk, terwijl dataset 2 gebruikt kan worden om de toets rond de cesuur te valideren en optimaliseren.
   - De vereiste hoeveelheid geluidsopnamen bestaat voor zowel de toets KNS als de TGN uit 200-300 toetsafnamen voor dataset 1 en 100-200 voor dataset 2, dus een totaal van 600-1000 toetsafnamen.

2) Formuleer succescriteria (normen) voor de TGN en de toets KNS
   - Laat alle opnamen beoordelen door (minimaal) vier menselijke beoordelaars die in staat zijn om een betrouwbaar oordeel over de taalvaardigheid van de kandidaat te vormen.
   - Gebruik per kandidaat steeds één menselijk oordeel als “human benchmark,” waarmee de geautomatiseerde toetsen kunnen worden vergeleken
   - Bepaal op basis van de overige oordelen (minimaal 3) een referentieoordeel. Omdat dit het gezamenlijk oordeel van meerdere deskundigen is, wordt geaccepteerd dat dit oordeel het ware taalvaardigheidsniveau van de kandidaat voldoende benadert
   - Bepaal een criterium voor succes van de geautomatiseerde toetsen relatief ten opzichte van de human benchmark.

3) Valideer de toetsen door de foutpercentages (onterecht zakken/slagen) op de nieuwe datasets te bepalen, en deze volgens het vastgestelde succescriterium te vergelijken met de human benchmark (aanbeveling 2).

4) Verbeter de TGN rond de A1min cesuur
   - Hertrain de spraakherkenner met meer materiaal van lage taalvaardigheidsniveaus.
   - Voor hertraining van de spraakherkenner kan gebruik gemaakt worden van de data die beschikbaar zijn gekomen in de aanvullende experimenten die na de pretest zijn uitgevoerd (Den Haag, Amsterdam, MFA-FIT), vooropgesteld dat er nieuwe (onafhankelijke) validatie-datasets beschikbaar komen (aanbeveling 1)
   - Hertrain/herschaal andere componenten van het automatische scoringssysteem op vergelijkbare manier als de spraakherkenner.
   - Pas eventueel meer complexe classificatie-algorithmen dan een eenvoudige cesuur toe.

# Inhoudsopgave

H DET-curve toets KNS voor een cesuur van 80%

# 1 Inleiding

Op 5 april 2005 heeft de Tweede Kamer ingestemd met een wijziging van de Vreemdelingenwet 2000. De wijziging betreft het stellen van een inburgeringvereiste bij het toelaten van bepaalde categorieën vreemdelingen. Het nieuwe inburgeringexamen Buitenland, dat met behulp van een geautomatiseerd systeem zal worden afgenomen, bestaat uit twee toetsen:

- De "Toets Gesproken Nederlands (TGN)" die meet of een kandidaat over voldoende mondelinge taalvaardigheid beschikt,
- De toets "Kennis van de Nederlandse Samenleving (KNS)" die meet of een kandidaat over voldoende kennis over de Nederlandse samenleving en cultuur beschikt.

Voor het inburgeringsexamen Buitenland is door de Commissie Franssen een nieuw mondeling taalvaardigheidniveau van A1min voorgesteld [ref1]. Dit taalvaardigheidniveau bevindt zich onder het laagste niveau (A1) dat in het "Common European Framework" (CEF) wordt beschreven. De TGN zal ook ingezet worden voor het inburgeringsexamen Binnenland. Voor de mondelinge taalvaardigheid van oudkomers en nieuwkomers in Nederland is een minimale taalvaardigheidseis van A2 (CEF) voorgesteld. De TGN 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 ontwikkelde spraakherkenner, die bekend is onder de naam Phonepass, verwerkt. Deze spraakherkenner bepaalt of de mondelinge taalvaardigheid voldoet aan de gestelde minimumeisen. De toets KNS bestaat uit vragen die aan de kandidaat gesteld worden. Op basis van de automatisch herkende antwoorden beoordeelt het systeem of een examenkandidaat voldoende kennis heeft over de Nederlandse samenleving. De twee toetsen zijn in opdracht van het Ministerie van Justitie ontwikkeld door het CINOP[3] in samenwerking met LTS[4] en Ordinate[5].

Het doel van de inzet van spraaktechnologie bij beide toetsen is automatisering. Voor de TGN wordt de spraakherkenner ingezet om de inhoudelijke correctheid van de antwoorden te controleren en om kwalitatieve aspecten van de spraak (zoals uitspraak en vloeiendheid) te beoordelen. Op basis van zowel deze inhoudelijke als kwalitatieve maten wordt bepaald of een kandidaat over voldoende mondelinge vaardigheid beschikt. Voor de toets KNS is de taak van de spraakherkenner principieel anders: de spraakherkenner wordt alleen gebruikt om automatisch te bepalen of een antwoord correct is of niet. Door de toetsen te automatiseren worden voordelen behaald (o.a. tijdwinst, flexibiliteit, schaalbaarheid) ten opzichte van de conventionele aanpak. Hoe succesvol de automatisering door toepassing van spraaktechnologie is, is derhalve af te meten door de resultaten van de toets te vergelijken met de resultaten van menselijke toetsers ("human benchmark"). In hoeverre de fouten die de toets maakt ten opzichte van de 'human benchmark' acceptabel zijn hangt af van de beleidsmatige uitgangspunten.

Het Ministerie van Justitie heeft in het recente verleden een aantal onderzoeken laten uitvoeren. De resultaten van deze studies zijn (nog) niet openbaar, maar wel aan TNO ter inzage gegeven. Het examen wordt in deze onderzoeken vanuit diverse invalshoeken (onderwijskundig, maar ook spraaktechnologisch) gevalideerd. In dit rapport wordt

[3] Onderdeel van CINOP Advies BV, 's Hertogenbosch, Nederland, www.cinop.nl
[4] Language Testing Services, Velp, Nederland
[5] Ordinate Corporation, Menlo Park, Califonië, USA, www.ordinate.com

verslag gedaan van een ‘second opinion’ door TNO over de validiteit van de spraaktechnologische aspecten van het inburgeringexamen. Doel van deze ‘second opinion’ is om op basis van onderzoeksgegevens die beschikbaar zijn (tot stand gekomen zonder medeweten of medewerking van TNO Defensie en Veiligheid) te beoordelen of op dit moment zonder voorbehouden geconcludeerd kan worden dat de geselecteerde spraakherkenningtechnologie voldoet voor het doel waarvoor deze zal worden ingezet.

## 1.1 Terminologie gebruikt in dit rapport

In dit rapport worden een aantal specifieke termen gebruikt die in appendix A worden samengevat.

## 1.2 Opzet rapport

De opzet van het rapport is als volgt: In hoofdstuk 2 zal de gehanteerde werkwijze bij de totstandkoming van deze ‘second opinion’ worden toegelicht. In hoofdstuk 3 zal de werking van beide toetsen kort worden beschreven en zal uitgelegd worden op welke manier spraaktechnologie is ingezet in beide toetsen. In hoofdstuk 4 zullen de resultaten van de ‘second opinion’ worden gegeven. Tenslotte in hoofdstuk 5 de conclusies en aanbevelingen kort en bondig geformuleerd worden. In hoofdstuk 6 wordt een toelichting gegeven op de conclusies en aanbevelingen.

# 2 Gehanteerde werkwijze

In paragraaf 2.1 wordt gedefinieerd wat TNO verstaat onder validatie van een spraaktechnologisch systeem. Vervolgens wordt in paragraaf 2.2 de algemene werkwijze uitgelegd voor validatie van spraaktechnologische systemen. Aangezien deze werkwijze verschilt van de algemene werkwijze voor validatie van toetsen en examens, wordt de werkwijze van toetsen en examens in paragraaf 2.3 uitgelegd. Tenslotte wordt in paragraaf 2.4 stapsgewijs uitgelegd wat de werkwijze is die in deze 'second opinion' gevolgd is.

## 2.1 Algemene werkwijze voor de ontwikkeling van een spraaktechnologisch systeem

Voor het ontwikkelen van een spraaktechnologisch systeem bestaat een algemeen geaccepteerde aanpak, die bestaat uit een drietal fases:

1) Training
2) Ontwikkeling/optimalisatie
3) Validatie

Tijdens training wordt het systeem geschikt gemaakt voor de specifieke taak waarvoor deze ingezet zal worden. Hiertoe worden de modellen die gebruikt worden door de spraakherkenner en/of het systeem waarvan de spraakherkenner onderdeel is, getraind met een grote hoeveelheid materiaal dat representatief is voor de toepassing. Tijdens ontwikkeling of optimalisatie worden de parameters van het systeem geoptimaliseerd. De laatste fase is de validatie van het getrainde en geoptimaliseerde systeem. De second opinion van TNO heeft alleen betrekking op de derde fase, de validatie.

## 2.2 Algemene werkwijze voor de validatie van een spraaktechnologisch systeem

Voor het valideren van spraakherkenningstechnologie bestaat een algemeen geaccepteerde aanpak, die is te herleiden tot de algemene principes van weten-schappelijk onderzoek. Vooraf dient een criterium voor succes te worden geformuleerd: aan welke eisen moet de technologie voldoen? Deze eisen moeten helder aangeven tot op welke hoogte de fouten, die altijd door de techniek kunnen worden gemaakt, acceptabel worden geacht. Aangezien in een spraaktechnologisch systeem taken geautomatiseerd worden, worden de prestaties van het spraaktechnologisch systeem vaak vergeleken met een 'human benchmark'. De 'human benchmark' een systeem dat vergelijkbaar is met het automatische systeem, maar waarbij alle taken door een mens worden uitgevoerd. Als succescriterium wordt daarom vaak gekozen de fouten die het spraaktechnologisch systeem maakt ten opzichte van een 'human benchmark'.

Vervolgens worden experimenten uitgevoerd om te beoordelen in hoeverre aan het succescriterium wordt voldaan. Voor dergelijk evaluatie-experimenten bestaan zekere internationale "standaards" (NIST, Eagles Handbook, internationale literatuur). Deze geven richtlijnen voor de meest cruciale aspecten van de validatie, waaronder:
- De gekozen maat en methodologie (geschiktheid van de gekozen maat voor de specifieke doelstelling)
- Samenstelling van het gekozen validatiemateriaal (onder andere taalachtergrond, type spraak, aantal proefpersonen, onafhankelijkheid van test- en trainmateriaal)
- Statistische betrouwbaarheid van de uitkomsten van de experimenten

## 2.3 Algemene werkwijze voor validatie van toetsen en examens

De algemene werkwijze voor validatie van spraaktechnologische systemen verschilt van de werkwijze die gangbaar en geaccepteerd is voor validatie van toetsen en examens[6]. Ook voor toetsen en examens worden concrete eisen aan de prestaties (onder andere in de vorm van betrouwbaarheidsscores) gesteld. Echter, de stelling is dat hard bewijs dat beter wordt gepresteerd dan een 'human benchmark' (conventionele toets met menselijke examinatoren) nauwelijks te geven is: strikt genomen is onbekend wat *daadwerkelijk* het taalvaardigheidsniveau van de proefpersonen is geweest. Als een kandidaat voor een conventionele toets zakt maar voor de automatische toets slaagt, is het alsnog mogelijk dat de automatische toets gelijk had en de menselijke examinator ongelijk; ook bij conventionele toetsen worden fouten gemaakt. In plaats van rigide bewijs, zoals in de spraakherkenning wordt vereist, wordt gezocht naar evidentie voor validiteit. Allerlei relevante aspecten van de toets worden tegen het licht gehouden, en aan tests onderworpen. Als hierbij geen problemen worden geconstateerd (prestaties beneden vastgestelde criteria) dan wordt gesteld dat de toets valide is.

## 2.4 Gehanteerde werkwijze gevolgd in deze second opinion

In deze 'second opinion' heeft TNO heeft de algemene werkwijze voor validatie van spraaktechnologische systemen gevolgd. Op grond van informatiebronnen die door het Ministerie van Justitie en door derden (Ordinate, CINOP, LTS) beschikbaar zijn gesteld is beoordeeld in hoeverre de volgende vragen te beantwoorden zijn:

1. Wat zijn de gekozen succescriteria?
2. Volgen deze criteria logisch uit de doelstelling voor de technologie?
3. Zijn de experimenten methodologisch correct uitgevoerd?
4. Voldoet het gekozen validatiemateriaal aan de aanvaarde praktijk voor validatie van spraaktechnologie? In spraaktechnologisch onderzoek zijn twee aspecten van het validatiemateriaal met name belangrijk:

   *- Onafhankelijkheid*
   Het is van belang dat de validatie-data niet gebruikt zijn bij de totstandkoming van de toets (training van de spraakherkenners en automatische scoringssystemen). Als de validatie-data niet onafhankelijk zijn zal een te rooskleurig resultaat verkregen worden.

   *- Representativiteit.*
   De uitgevoerde experimenten moeten voldoende representatief zijn voor de werking van de toets in de praktijk. Representativiteit van de uitgevoerde experimenten zal getoetst worden op basis van de achtergrondgegevens van de sprekers die van belang zijn voor de prestaties van de spraakherkenner (zoals taalachtergrond, taalvaardigheidniveau). Aan de representativiteit van de akoestische condities is in het kader van deze second opinion in beperkte mate aandacht geschonken.

Door systematisch de informatie te inventariseren waarmee bovenstaande vragen voor beide typen toetsen (TGN en KNS) te beantwoorden zijn, is een oordeel over de toereikendheid van eerder uitgevoerde validatiestudies tot stand gekomen. Deze 'second opinion' is gebaseerd op de literatuur (zoals vermeld in de referentielijst), op gesprekken

[6] In een toelichting stelt CINOP/LTS dat de door hen gevolgde aanpak gangbaar en geaccepteerd is voor toetsen en examens

en emailconversatie met de partijen die de eerdere validatiestudies hebben uitgevoerd, op data die beschikbaar zijn gesteld door de betrokken partijen en op bij TNO aanwezige kennis en ervaring op het gebied van spraakherkenningtechnologie. Literatuur gerelateerd aan de Phonepass-technologie in andere talen (Spaans/Engels) is eveneens in het onderzoek betrokken.

# 3 Totstandkoming toetsscores door spraakherkenning

In dit hoofdstuk wordt de spraaktechnologische basis van de toetsen (beknopt) uiteengezet. Deze informatie is ondersteunend bij het verkrijgen van een dieper inzicht in de toetsen, waardoor de uitgevoerde validatiestudies beter op waarde kunnen worden geschat.

De technologie van het inburgeringexamen komt voort uit de Phonepass-toets, ontwikkeld door het bedrijf Ordinate, waarmee de mondelinge taalvaardigheid in onder andere het Engels en het Spaans kan worden bepaald. In opdracht van het Ministerie van Justitie heeft CINOP in samenwerking met LTS en Ordinate een Nederlandse versie van de toets (TGN) ontwikkeld. Het gemeten taalvaardigheidniveau wordt in dit geval gebruikt om te bepalen of een vreemdeling aan het vereiste mondelinge taalvaardigheidniveau voldoet. Daarnaast is ook een toets ontwikkeld (KNS) die de kennis van de Nederlandse samenleving meet. Hiertoe is de spraakherkenner uit de TGN aangepast voor toepassing in de toets KNS. Ook de toets KNS is door CINOP in samenwerking met LTS en Ordinate ontwikkeld.

## 3.1 Totstandkoming toetsscores in de TGN

### *3.1.1 Opzet toets*

De TGN verschilt van de Engelse en Spaanse toets omdat niet alle kandidaten die de toets afleggen kunnen lezen en schrijven. Om deze reden zijn de onderdelen verwijderd waarbij tekst voorgelezen moet worden (onderdelen “voorlezen” en “zinnen bouwen”). De TGN bestaat uit een drietal onderdelen en 45 items:

a) Zinnen herhalen (23 items): De gesproken uiting dient correct te worden herhaald.
b) Kort-antwoord: (13 items): Een vraag moet (kort) beantwoord worden
c) Tegenstellingen: (9 items): Het tegengestelde woord moet gezegd worden.

### *3.1.2 Opbouw totaalscore*

In Tabel 1 is weergegeven hoe de totale toetsscore is opgebouwd. De *inhoudelijke* correctheid wordt bepaald uit de deelscores “zinsbouw” en “woordenschat”. De *kwalitatieve* correctheid wordt bepaald op basis van de deelscores “vloeiendheid” en “uitspraak”. Iedere deelscore heeft een gelijk gewicht in de totale score.

| **Totale score** | **Deelscore** | **Toetsonderdeel** |
|---|---|---|
| inhoudelijk | zinsbouw | zinnen herhalen |
| | woordenschat | kort-antwoord + tegenstellingen |
| kwalitatief | vloeiendheid | zinnen herhalen |
| | uitspraak | zinnen herhalen |

Tabel 1: Opbouw van de totale score

### *3.1.3 Automatische scoring*

De spraakherkenner is onderdeel van het automatische scoringscomponent die gebruikt wordt voor zowel inhoudelijke als kwalitatieve scoring. Invoer zijn de geregistreerde

antwoorden van een kandidaat (geluidsfiles), uitvoer is de totaalscore. De automatische scoringscomponent bestaat uit een groot aantal modellen. De opbouw en training van de scoringscomponent is verkort beschreven in appendix B. Voor de ontwikkeling van de toets en voor de training/schaling van de modellen van de automatische scoringscomponent is gebruik gemaakt van een dataset van 821 moedertaal- en 1522 niet-moedertaalsprekers. Een deel van deze pretest-data is apart gehouden voor validatie van het automatische scoringssysteem.

## 3.2 Totstandkoming toetsscores in de toets KNS

### *3.2.1 Opzet toets*

Alvorens de toets wordt afgenomen dient de kandidaat de film “Naar Nederland” te bekijken. In deze film wordt Nederlands gesproken. Daarnaast zijn er een aantal versies van de films in de moedertaal van de kandidaat beschikbaar. Er is gekozen voor een toetsvorm naar anologie van het theoretische deel van het rijexamen. Kandidaten krijgen tijdens het examen een boekje met 30 vragen voorgelegd. De vragen worden gesteld aan de hand van een foto uit de film (still). Ter voorbereiding kunnen kandidaten oefenen met de film en een opgavenboekje dat alle 100 vragen en antwoorden bevat met de daarbij horende foto’s. Daarnaast kan er ook geoefend worden met een cassette waarop de gesproken vragen zijn te horen. Een kandidaat kan het juiste antwoord geven en vervolgens beluisteren of het antwoord correct is. De toets bestaat uit een zevental onderdelen waaronder bijvoorbeeld; geografie, vervoer en wonen, geschiedenis en staatsinrichting. De items zijn zo samengesteld dat deze te beantwoorden zijn door kandidaten met een A1min taalvaardigheidniveau. Alle items bestaan uit drie soorten vragen: 1) gesloten vragen met een ja/nee antwoord, 2) gesloten vragen met twee antwoordmogelijkheden, 3) open vragen met een eenduidig antwoord.

### *3.2.2 Automatische scoring*

De spraakherkenner in de toets KNS wordt gebruikt om automatisch de inhoudelijke correctheid van de antwoorden te bepalen. Invoer zijn de antwoorden van een kandidaat (geluidfiles), uitvoer is een score. De score is een schatting van het percentage correcte antwoorden dat op de gehele itembank (100 vragen) gehaald zal worden. Voor de ontwikkeling van de toets en de training/schaling van de modellen van de automatische scoringscomponent is gebruik gemaakt van een dataset van ruim 1000 niet-moedertaalsprekers. De spraakherkenner in de toets KNS is een aangepaste versie van de spraakherkenner uit de TGN (het taalmodel is aangepast). Voor meer informatie over de opbouw en training van de scoringscomponent van de toets KNS, zie appendix C.

## 3.3 Onafhankelijkheid trainingsmateriaal en validatiemateriaal

Zoals genoemd in paragraaf 2.4 is het in spraaktechnologisch onderzoek van belang dat het validatiemateriaal onafhankelijk is. Dit betekent dat het materiaal dat is gebruikt voor het trainen van de modellen van de spraakherkenner en de modellen van de automatische scoringscomponent, niet gebruikt mag worden voor validatie van het spraakherkenningssysteem. In het geval van de TGN is het materiaal dat gebruikt is voor de ontwikkeling van de toets ook gebruikt voor training van de spraakherkenner. Een deel van dit materiaal is apart gehouden voor validatie (bestaande uit 139 niet-moedertaalsprekers). De akoestische modellen van de spraakherkenner in de toets KNS zijn gelijk aan die van de TGN. Het taalmodel is getraind met het materiaal dat ook

gebruikt is voor ontwikkeling van de toets KNS. Een deel van dit materiaal is apart gehouden voor validatie (bestaande uit 59 niet-moedertaalsprekers).

# 4 Second Opinion Toets Gesproken Nederlands (TGN)

## 4.1 Werkwijze second opinion TGN

In dit hoofdstuk worden de beschikbare antwoorden gegeven op de vragen die in hoofdstuk 2 (werkwijze) zijn gesteld. Allereerst wordt beschreven wat het gekozen succescriterium is. Ten tweede wordt aangegeven in welke mate het geformuleerde succescriterium logisch te relateren is aan de doelstelling van de toets. Ten derde is beschreven of de experimenten methodologisch correct zijn uitgevoerd, alleen mogelijke tekortkomingen worden beschreven. Tenslotte is gekeken naar de onafhankelijkheid en representativiteit van het validatiemateriaal.

Een studie of analyse wordt een validatiestudie genoemd als het doel van deze studie is het leveren van evidentie of bewijs omtrent validiteit van de toets. Voor validiteit gebruiken wij in dit rapport dezelfde definitie als in de CINOP-rapportage. De toets is valide als deze daadwerkelijk hetgeen meet waar de toets voor ingezet wordt; namelijk mondelinge taalvaardigheid. Alle studies die betrekking hebben op ontwikkeling van de toets of onderdelen daarvan zijn niet in deze second opinion meegenomen.

Op verzoek van de opdrachtgever is ook een oordeel gegeven over de validatiestudies van de Engelse en Spaanse versie van de toets, beschreven in paragraaf 4.1. De validatiestudies die door CINOP[7] zijn uitgevoerd worden in chronologische volgorde beschreven. Toen TNO het onderzoek startte was een conceptversie van het eindrapport ‘Verantwoording TGN’ [ref2] beschikbaar. In paragraaf 4.3 worden de validatiestudies uit dit conceptrapport beschreven. Op 19 september 2005 kwam het eindrapport ‘Verantwoording TGN’ [ref3] uit. De validatiestudies uit het eindrapport zijn beschreven in paragraaf 4.4 en 4.5. In een apart hoofdstuk geeft TNO een oordeel over de representativiteit van de data wat betreft de taalachtergrond (paragraaf 4.6), aangezien dit aspect van de data gemakkelijker als geheel te beschrijven was dan apart per validatiestudie. Een samenvatting van de beoordeelde studies en van de conclusies die TNO trekt op basis van deze studies wordt gegeven in paragraaf 4.7. Daarnaast heeft TNO een aanvullende validatiestudie uitgevoerd volgens de werkwijze die gebruikelijk is in spraaktechnologisch onderzoek, deze studie wordt beschreven in hoofdstuk4.8. In Tabel 2 is de inhoud van de second opinion kort samengevat.

[7] De validatiestudies zijn uitgevoerd in samenwerking met Ordinate en LTS

| bron | paragraaf | validatiestudie |
|---|---|---|
| literatuur | 4.2 | Validatie van de Engelse en Spaanse versie van de toets |
| CINOP concept - rapport (4.3) | 4.3.1 | Validatie van het automatische scoringssysteem in de TGN |
| | 4.3.2 | Validatie benadeling kandidaten met een sterk buitenlands accent |
| | 4.3.3 | Correlaties CEF-oordelen en toetsscores |
| | 4.3.4 | Validatie van de gekozen A1min cesuur |
| CINOP eind-rapport (4.4+4.5) | 4.4.2 | Relatie toetsscores en beheersing van het Nederlands |
| | 4.4.3 | Relatie toetsscores en achtergrondvariabelen |
| | 4.4.4 | Correlatie TGN-scores en docentoordelen |
| | 4.4.5 | Overeenstemming mens-mens, mens-machine en machine-machine |
| | 4.5 | Second opinion over betrouwbaarheidsanalyses uit CINOP eindrapport |
| varia | 4.6 | Representativiteit van de data wat betreft de taalachtergrond |
| | 4.7 | Samenvatting beoordeling validatiestudies |
| | 4.8 | Aanvullende validatie TNO |

Tabel 2: Overzicht second opinion TGN

## 4.2 Validatie van de Engelse en Spaanse versie van de toets

*Succescriterium*
Voor de Engelse en Spaanse versie van de toets zijn verscheidene validatiestudies uitgevoerd [ref4] [ref5]. In deze validatiestudies worden de toetsscores gecorreleerd met menselijke oordelen verkregen met behulp van (gecertificeerde) methoden om taalvaardigheid te beoordelen. De “human benchmark” bestaat in dit geval uit globale oordelen over de taalvaardigheid. De validaties worden als succesvol beschouwd aangezien over het algemeen hoge correlaties worden gevonden. Voor Engelse luister- en spreekvaardigheidstesten worden bijvoorbeeld correlaties gevonden die variëren tussen 0,71 en 0,94.

*Geschiktheid t.b.v. doel*
In relatie tot het doel waarvoor de toetsen zijn ontwikkeld, namelijk het automatisch inschatten van het taalvaardigheidniveau op een schaal van zeer laag tot bijna-moedertaalbeheersing, is dit een toereikende aanpak. Er zijn echter twee redenen waarom het succes van de Engelse en Spaanse versies van de toets geen garantie is voor succes van de Nederlandse versie van de toets. Ten eerste zijn de resultaten toets- en taalafhankelijk. Ten tweede garandeert een hoge correlatie over de gehele taalvaardigheidsschaal niet dat met voldoende nauwkeurigheid bepaald kan worden of aan een zeker (laag) taalvaardigheidsniveau wordt voldaan.

*Validatiemateriaal*
Het validatatiemateriaal voldoet aan de normen voor onafhankelijkheid en representativiteit

## 4.3 Second opinion validatiestudies uit conceptrapport “Verantwoording van de TGN”

De validatiestudies die eerder zijn uitgevoerd (niet door TNO) en die zijn beschreven in een conceptversie van het rapport “Verantwoording Toets Gesproken Nederlands” [ref2] worden hieronder achtereenvolgens beschreven. Aangezien in de conceptversies van het

CINOP-rapport niet altijd duidelijk was aangegeven of een studie gericht is op het geven van evident/bewijs omtrent validateit van de toets, heeft TNO een keuze gemaakt. Hierbij is als uitgangspunt genomen dat alle studies die betrekking hebben op validatie van de spraaktechnologie of de toets, zijn beoordeeld.

### *4.3.1 Validatie van het automatische scoringssysteem in de TGN*

*Gekozen succescriterium*

In paragraaf 6.5 van [ref2] is een validatie beschreven waarbij het automatische scoringssysteem van de TGN apart is gevalideerd. In deze validatie is een vergelijking gemaakt tussen automatisch verkregen toetsscores en menselijke oordelen. De automatische score is opgebouwd uit afzonderlijke subscores. De menselijke beoordelaars voerden exact dezelfde taak uit als de spraakherkenner: zij beoordeelden vier verschillende aspecten van taalvaardigheid, gebaseerd op dezelfde (opgenomen) spraakfragmenten. Uit de menselijke subscores en de subscores van de spraakherkenner is op dezelfde manier de uiteindelijke totale toetsscore berekend. Er werd een hoge correlatiecoëfficiënt (0,94) gevonden tussen de menselijke en de machinale totale toetsscores.

*Geschiktheid t.b.v. doelstelling*

Er zijn drie redenen te noemen waarom een hoge correlatiecoefficient geen bewijs is dat de toets succesvol is. Ten eerste is de hoogte van een correlatiecoëfficiënt in dit geval geen garantie voor succes. Een toets is succesvol als het percentage onterechte toetsuitslagen laag genoeg is. Een hoge correlatiecoëfficiënt garandeert echter niet dat het percentages onterechte toetsuitslagen ook laag is: De correlatiecoëfficiënt wordt over de gehele schaal gemeten, terwijl de TGN meet of een kandidaat over één specifiek taalvaardigheidsniveau beschikt. Het is denkbaar dat de globale overeenstemming goed is, maar dat de verschillen rondom dat specifieke taalvaardigheidsniveau toch aanmerkelijk zijn. Dit percentage kan niet (alleen) worden voorspeld uit de hoogte van de correlatiecoëfficiënt. Met andere woorden: het succes van de toets (mate waarin ten opzichte van een menselijk oordeel fouten worden gemaakt) wordt niet direct gemeten door de gebruikte maat. Dit wordt geïllustreerd aan de hand van een rekenvoorbeeld in appendix D. Ten tweede zijn de beoordelingen niet gerelateerd aan de CEF-schaal; mens en machine kunnen op een consistente manier tot een oordeel komen, maar dit zegt niets over de relatie tussen de verkregen score en de CEF-schaal. Ten derde zijn de beoordelaars in de rol van de spraakherkenner gedrongen. Het is aannemelijk dat een “human benchmark,” gebaseerd op menselijke toetsoordelen die op een andere manier zijn verkregen dan de manier waarop de spraakherkenner dit doet, een minder goede correlatie oplevert.

*Validatiemateriaal*

De correlaties zijn berekend op een onafhankelijke subset van de zogenaamde pretestdata; dit is methodologisch correct. De gebruikte validatie-data zijn minder representatief rond de A1min cesuur.

### *4.3.2 Validatie benadeling kandidaten met een sterk buitenlands accent*

*Gekozen succescriterium*

In paragraaf 6.5.2 [ref2] is een contrahypothese geformuleerd en deze hypothese is getoetst. De contrahypothese houdt in dat kandidaten met een sterk buitenlands accent worden benadeeld doordat de spraakherkenner meer fouten maakt dan bij kandidaten die een minder sterk accent hebben. De gekozen maat is correlatie. Hierbij worden de menselijke oordelen over de uitspraak als maat genomen voor de mate van buitenlands

accent. De mate van benadeling door de spraakherkenner wordt onderzocht door het verschil tussen de automatische deelscores en de deelscores gebaseerd op menselijke oordelen te berekenen. Als de contrahypothese correct is wordt een positieve correlatie verwacht. Ordinate heeft voor TNO de correlatiecoefficienten berekend voor alle deelscores. Aangezien de correlatiecoefficient varieert van –0,10 tot –0,31 concludeert TNO dat de contrahypothese niet wordt bevestigd.

*Geschiktheid t.b.v. doelstelling*
De gekozen maat geschikt is voor het testen van de gestelde contrahypothese.

*Validatiemateriaal*
De correlaties zijn berekend op een onafhankelijke subset van de zogenaamde pretestdata; dit is methodologisch correct. De gebruikte validatie-data zijn minder representatief rond de A1min cesuur.

### *4.3.3 Correlaties CEF-oordelen en toetsscores*

*Gekozen succescriterium*
Evidentie voor validiteit wordt gevonden door de correlatie te berekenen tussen de toetsscores en menselijke beoordelingen. De menselijke oordelen werden gegeven door de eigen docent volgens de CEF-schaal. De correlatiecoefficient tussen CEF-oordelen en de toetsscore is 0,82.

*Geschiktheid t.b.v. doelstelling*
Zoals eerder genoemd is het nadeel van de gebruikte maat (correlatie) dat deze niet aansluit bij de doelstelling van de toets, namelijk het bepalen of een kandidaat over het vereiste taalvaardigheidniveau beschikt of niet.

*Methodologie*
De CEF-oordelen zijn gebaseerd op oordelen van de eigen docent. Deze oordelen zijn minder geschikt om te gebruiken aangezien de docenten niet allen getraind waren in het beoordelen met de CEF-schaal.

*Validatiemateriaal*
De correlaties zijn berekend op een onafhankelijke subset van de zogenaamde pretestdata; dit is methodologisch correct. De gebruikte validatie-data zijn minder representatief rond de A1min cesuur.

### *4.3.4 Validatie van de A1min cesuur*

*Gekozen succescriterium*
In het vervolg op paragraaf 7.7 (onder tabel 7.10 [ref2]), is de eerder ingestelde cesuur gevalideerd door te berekenen voor welk percentage van de kandidaten de TGN toetsuitslag overeenkomt met de menselijke oordelen. Het percentage toetsuitslagen bleek oorspronkelijk voor 73% van de gevallen overeen te komen met het unanieme oordeel van drie menselijke beoordelaars. Na aanpassing van de cesuur op grond van aanvullende experimenten stijgt dit percentage naar 78%.

*Geschiktheid t.b.v. doelstelling*
Het percentage toetsuitslagen dat overeenstemt met het unanieme oordeel van menselijke beoordelaars is in principe een goede maat, omdat deze direct aansluit bij de doelstelling. Nadeel van de gekozen maat is echter dat er geen onderscheid gemaakt wordt tussen het percentage terecht gezakte en terecht geslaagde kandidaten. De exacte percentages hangen van elkaar af en worden bepaald door de toetsscore die bepaalt of een kandidaat slaagt voor de toets (de cesuur). Door de keuze van de cesuur bestaat de mogelijkheid om het

éne type fout (onterecht zakken) uit te ruilen tegen het andere type fout (onterecht slagen). De gekozen maat geeft geen inzicht in de gemaakte afweging tussen beide typen fouten. Verder is het percentage terecht beslissingen afhankelijk van de verdeling van taalvaardigheidniveau's in de steekproef. De maat die gekozen is in de validatie geeft geen inzicht in het te verwachten percentage als de taalvaardigheidniveaus van de kandidaten in de praktijk afwijkend zijn van de niveaus die voorkomen in de validatieset.

*Validatiemateriaal*
Er is gebruik gemaakt van een onafhankelijke testset (MFA-FIT). Verder lijken de data die in deze validatie worden gebruikt representatief te zijn rond de A1min cesuur (15% van de kandidaten heeft een niveau kleiner dan A1min[8]), maar minder representatief rond de A2 cesuur (72% van de kandidaten heeft een niveau kleiner dan A2[9]).

## 4.4 Second opinion over validatiestudies uit eindrapport 'Verantwoording TGN'

Op 19 september 2005 is het eindrapport 'Verantwoording Toets Gesproken Nederlands' [ref3] verschenen. In hoofdstuk 6 van het CINOP-eindrapport zijn de analyses beschreven die zijn uitgevoerd om de betrouwbaarheid en validiteit van de TGN te onderbouwen. De toetsscores worden *valide* genoemd als zij gerelateerd zijn aan hetgeen de toets beoogt te meten, namelijk het mondelinge taalvaardigheidniveau. In deze paragraaf beschrijft TNO de 'second opinion' over deze validatiestudies.

### *4.4.1 Overeenstemming met menselijke beoordeling en effecten van verschillen in uitspraak*

De analyses beschreven in paragraaf 6.2.1 zijn al eerder beoordeeld in deze second opinion, zie paragraaf 4.3.1 en 4.3.2.

### *4.4.2 Relatie toetsscores en beheersing van het Nederlands*

*Succescriterium*
In paragraaf 6.2.2 worden de cumulatieve frequenties van de totaalscores voor niet-moedertaalsprekers (NMS) en moedertaalsprekers (MS) gepresenteerd. Aangezien de MS bijna uitsluitend hoge totaalscores halen wordt dit als evidentie gezien voor validiteit van de toets.

*Geschiktheid m.b.t. doel*
Met deze analyse kan aangetoond worden dat de TGN een zinnig onderscheid lijkt te maken tussen NMS en MS. De analyse sluit echter niet aan bij de doelstelling van de toets, aangezien het doel van de TGN niet is om een onderscheid te maken tussen MS en NMS, maar om onderscheid te maken tussen sprekers die wel of niet over een specifiek taalvaardigheidniveau (A1min of A2) beschikken.

*Validatiemateriaal*
Een methodologisch bezwaar is dat de analyses niet zijn uitgevoerd op alle pretestdata en deze dataset is niet onafhankelijk.

### *4.4.3 Relatie toetsscores en achtergrondvariabelen*

*Succescriterium*
In paragraaf 6.2.3 wordt een relatie gelegd tussen de toetsscores en een aantal eigenschappen en kenmerken van kandidaten waarvan verondersteld wordt dat ze niet

[8] Deze getallen zijn schattingen gebaseerd op het meerderheidsoordeel van 3 beoordelaars voor 249 kandidaten

samenhangen met mondelinge taalvaardigheid. Succescriterium is dat er geen verband wordt gevonden tussen de toetsscores en deze achtergrondvariabelen.

*Geschiktheid m.b.t. doel*
Met de gekozen methodologie kan aangetoond worden dat de scores niet ongewenst worden beinvloed door relevante achtergrondvariabelen, maar er wordt geen verband gelegd met de doelstelling van de toets. Een geschikter criterium zou zijn: Is het aantal onterechte toetsuitslagen onafhankelijk van de onderzochte achtergrondvariabelen?

*Validatiemateriaal*
Een methodologisch bezwaar is dat de analyses zijn uitgevoerd op de gehele pretest-dataset en deze dataset is niet onafhankelijk.

### *4.4.4 Correlatie TGN scores en docentoordelen*

*Succescriterium*
In paragraaf 6.2.4 (tabel 6.7) worden de correlatiecoëfficiënten gegeven tussen de TGN-scores en docentoordelen. De analyse tonen aan dat de correlatiecoëfficiënt hoger is ten opzichte van getrainde beoordelaars die hun oordeel geven op basis van een gestructureerd interview dan ten opzichte van beperkt/niet getrainde beoordelaars die een globaal oordeel geven. Aangezien de overeenstemming tussen de toetsscores en de oordelen die de CEF-niveau's beter representeren hoger is, is dit een gewenst effect.

*Geschiktheid m.b.t. doel*
Nadeel van de gebruikte methodologie is wederom dat een hoge correlatiecoëfficiënt geen garantie is voor succes. Verder is de gevonden correlatie niet heel hoog (0.70) en is correlatie slechts een kwantitatieve maat; een hoge mate van overeenstemming tussen twee oordelen garandeert niet dat de kwaliteit van beide oordelen ook hoog is.

*Validatiemateriaal*
Er is gebruik gemaakt van twee onafhankelijke datasets (Den Haag, en Amsterdam) en één dataset die niet onafhankelijk is (pretest).

### *4.4.5 Overeenstemming tussen mens-mens, mens-machine en machine-machine oordelen*

*Succescriterium*
In paragraaf 6.2.4 (tabel 6.8+6.9) worden de percentages overeenstemming tussen mens-mens, mens-machine en machine-machine oordelen gegeven. De validatie wordt als succesvol beschouwd als de mate van overeenstemming tussen mens-machine oordelen van dezelfde orde is als de mate van overeenstemming tussen menselijke beoordelaars.

*Geschiktheid m.b.t. doel*
De percentages overeenstemming zeggen alleen iets over de *mate* van overeenstemming, maar niet iets over de *kwaliteit*. Zoals CINOP ook al opmerkt is op basis van deze vergelijking niet te zeggen welke beoordelaar meer gelijk heeft, aangezien het ware taalvaardigheidniveau onbekend is. De kandidaten waarover de beoordelaars het eens zijn kunnen verschillen van de kandidaten waarover mens en machine hetzelfde oordeel geven. Een tweede nadeel van de gekozen maat is dat niet gecorrigeerd wordt voor overeenstemming op basis van kans. Dit is de overeenstemming die wordt bereikt met een machine die willekeurige beoordelingen produceert.

*Aanvullende analyse*
CINOP heeft op verzoek van TNO nieuwe berekeningen uitgevoerd waarbij overeenstemming is berekend met een correctie voor overeenstemming op basis van kans

(zie appendix E). Uit deze gegevens blijkt dat de mate van overeenstemming tussen mens en machine lager is voor de A1min cesuur dan voor de A2 cesuur. Verder blijkt dat door de correctie het verschil in overeenstemming tussen mens-mens en mens-machine oordelen groter wordt. Of de verschillen in mate van overeenstemming tussen mens-mens en mens-machine beoordelingen acceptabel is, is niet te zeggen aangezien er geen succescriterium is gedefinieerd.

*Validatiemateriaal*
Er is gebruik gemaakt van een onafhankelijke dataset (Amsterdam). De Amsterdam dataset is de enige set is die echt onafhankelijk is, aangezien deze dataset ook niet gebruikt is voor schaling en normering. De data zijn minder representatief rond de A2 cesuur (85% van de kandidaten heeft een niveau kleiner dan A2).

#### *4.4.6 Samenhang menselijke oordelen en aspecten van taalvaardigheid in de TGN*

In paragraaf 6.2.5 wordt de samenhang tussen menselijke oordelen en aspecten van taalvaardigheid in de TGN onderzocht. Aangezien deze analyse hypothesevormend is en er geen conclusies worden getrokken op basis van deze analyse, geeft TNO geen 'second opinion' over deze analyse.

### 4.5 Second opinion betrouwbaarheidsanalyses uit eindrapport 'Verantwoording TGN'

In paragraaf 6.1 [ref3] worden een aantal betrouwbaarheidsanalyses beschreven. Toetsscores worden *betrouwbaar* genoemd als de toetsresultaten bij verschillende afnamen consistent zijn.

*Gekozen succescriterium*
Er is gebruik gemaakt van twee betrouwbaarheidsmaten; de split-half betrouwbaarheid en de Cronbach alpha. Alle betrouwbaarheidsmaten hebben een waarde die groter is dan 0.80, het criterium dat aan de betrouwbaarheid is gesteld. Daarnaast wordt een aan betrouwbaarheid gerelateerde maat gegeven, de geschatte meetfout per CEF-niveau. De meetfout is kleiner voor de lage CEF-niveaus dan voor de hoge CEF-niveaus.

*Geschiktheid t.b.v. doel*
De betrouwbaarheidsmaten (split-half en cronbach alpha) die zijn gebruikt zijn methodologisch correct. Nadeel van de gekozen maat voor lokale betrouwbaarheid (meetfout) is dat het schattingen zijn op basis van een model en niet op basis van data.

*Validatiemateriaal*
Er is gebruik gemaakt van onafhankelijke testsets (onafhankelijk deel pretest, MFA-FIT). De pretest-data zijn minder representatief rond de A1min cesuur. De MFA-FIT data zijn minder representatief rond de A2 cesuur.

### 4.6 Representativiteit van de data wat betreft de taalachtergrond

In de voorgaande paragrafen is met name gesproken over de representativiteit van de data wat betreft het taalvaardigheidsniveau van de sprekers. Naast taalvaardigheidsniveau is de taalachtergrond van de kandidaten van belang. Om deze reden heeft TNO gekeken naar de representativiteit van de taalachtergrond van de pretest-data en de data van het MFA-FIT-experiment. Voor beide datasets was ruim 1/3 deel van de kandidaten afkomstig uit Afghanistan, Irak, Marokko of Turkije. In [ref2] (hoofdstuk 8.4, figuur 15) staan de verdeling van de toetsscores per land gegeven. Een vergelijking met de nationaliteit van MVV-aanvragers (bron: Ministerie van Justitie) geeft geen aanleiding om te veronderstellen dat de data die gebruikt zijn voor de training van de

spraakherkenningscomponent (pretest data) onvoldoende respresentatief zijn voor de praktijk: De meest voorkomende nationaliteiten van MVV-aanvragers zijn ook aanwezig in het pretestmateriaal.

## 4.7 Samenvatting en conclusies op basis van beoordeelde studies

*Validiteit van de TGN*
In Tabel 3 zijn de verschillende validatiestudies die uitgevoerd zijn door CINOP samengevat. De volgende conclusies kunnen getrokken worden omtrent de verschillende deelvragen die gesteld zijn:

*- Succescriterium*
Er is gekozen voor verschillende succescriteria, afhankelijk van de doelstelling van de validatiestudie.
*- Geschiktheid succescriterium t.b.v. doelstelling toets*
Aangezien de door CINOP toegepast werkwijze gericht is op het vinden van evidentie voor validiteit is het gekozen succescriterium over het algemeen niet gekoppeld aan de doelstelling van de toets. Hierdoor kan geen van de studies bewijs leveren omtrent validiteit van de toets.
*- Methodologie*
De toegepaste methodologie is over het algemeen correct.
*- Validatiemateriaal*
Er is niet altijd een strikt onderscheid gemaakt tussen trainings/ontwikkelings-materiaal enerzijds en valideringsmateriaal anderzijds.

De betekenis van deze conclusies voor de TGN kan als volgt worden samengevat:

- Er is aangetoond dat kandidaten met een sterk buitenlands accent niet door de spraakherkenner worden benadeeld met lagere deelscores
- Er is aangetoond dat het trainingsmateriaal waarmee de spraakherkenner getraind is, voor de meest frequente taalachtergronden representatief is voor de praktijk
- Er zijn aanwijzingen dat de toets op een zinnige manier taalvaardigheid meet;
  - Er wordt een redelijke tot hoge correlatie gevonden tussen de TGN toetsscores en verschillende menselijke beoordelingen.
  - De toets maakt onderscheid tussen moedertaal- en niet-moedertaalsprekers,
  - Er is geen verband gevonden met achtergrondvariabelen die niet van belang zijn voor mondelinge taalvaardigheid.
- De validatiestudies leveren geen bewijs omtrent de validiteit van de toets: Er is niet aangetoond dat de toets *voldoende* valide is, noch dat de toets *onvoldoende* valide is

| **para-graaf** | **validatiestudie** | **criterium** | **geschiktheid t.b.v doelstelling toets** | **methodologie** | **validatiemateriaal** |
|---|---|---|---|---|---|
| 4.3.1 | automatische scoring | correlatie | - maat niet gekoppeld aan zak/slaag toets<br>- mens beoordeelt als computer<br>- maat niet gekoppeld aan CEF-schaal | | *onafhankelijk deel pretest*<br>+ onafhankelijke testset<br>- minder representatief rond A1min |
| 4.3.2 | benadeling kandidaten met sterk accent door spraakherkenner | correlatie | - maat niet gekoppeld aan zak/slaag toets<br>+ maat geschikt voor testen van contra-hypothese | | *onafhankelijk deel pretest*<br>+ onafhankelijke testset<br>- minder representatief rond A1min |
| 4.3.3 | validatie Nederlandse toets | correlatie | - maat niet gekoppeld aan zak/slaag toets<br>+ menselijke oordelen op CEF-schaal | - menselijke beoordelingen minder betrouwbaar | *onafhankelijk deel pretest*<br>+ onafhankelijke testset<br>- minder representatief rond A1min |
| 4.3.4 | validatie A1min cesuur | percentage overeenstemming met unaniem menselijk oordeel | + maat gekoppeld aan zak/slaag toets<br>+ menselijke oordelen op CEF-schaal<br>- geen onderscheid terecht zakken/slagen | - A2 cesuur ontbreekt | *MFA-FIT*<br>+ onafhankelijke testset<br>- minder representatief rond A2 |
| 4.4.2 | Relatie toetsscores en beheersing van het Nederlands | verdeling totaalscores NMS en MS | - maat niet gekoppeld aan zak/slaag toets<br>+ toets onderscheidt NMS en MS | | *pretest*<br>- niet onafhankelijk<br>- minder representatief rond A1min |
| 4.4.3 | Relatie toetsscores en achtergrondvariabelen | verband totaalscores en achtergrondvariabele | - maat niet gekoppeld aan zak/slaag toets<br>+ geen ongewenste effecten toetsscores en achtergrondvariabelen | | *pretest*<br>- niet onafhankelijk<br>- minder representatief rond A1min |
| 4.4.4 | Correlatie TGN scores en docentoordelen | correlatie | - maat niet gekoppeld aan zak/slaag toets<br>+ correlatie met meer representatief CEF-oordeel is groter | | *Pretest, Den Haag, Amsterdam*<br>- niet alle data onafhankelijk<br>- niet alle data representatief |
| 4.4.5 | Overeenstemming mens-mens, mens-machine en machine-machine oordelen | verschil in overeen-stemming mens-mens en mens-machine oordelen | + maat gekoppeld aan doelstelling toets<br>- maat is kwantitatief niet kwalitatief<br>- succescriterium ontbreekt | - geen correctie voor overeenstemming op basis van kans | *Amsterdam*<br>+ onafhankelijk<br>- minder representatief rond A2 |

Tabel 3: Samenvatting second opinion validatiestudies beschreven in CINOP's concept- en eindrapport 'Verantwoording TGN'

*Betrouwbaarheid van de TGN*
In Tabel 4 is een samenvatting gegeven van de betrouwbaarheidsstudies die zijn uitgevoerd. Aangezien de gevonden waarden voor betrouwbaarheid groter zijn dan de betrouwbaarheid van 0.80 waaraan de toets volgens de opdracht moet voldoen concludeert TNO dat er evidentie is dat de toets voldoende betrouwbare (=consistente) oordelen geeft

| | **criterium** | **geschiktheid t.b.v doelstelling toets** | **validatiemateriaal** |
|---|---|---|---|
| 1 | split-half betrouwbaarheid | ja | *onafhankelijk deel pretest*<br>+ onafhankelijke testset<br>- minder representatief rond A1min |
| 2 | Cronbach's alpha | ja | *MFA-FIT*<br>+ onafhankelijke testset<br>- minder representatief rond A2 |
| 3 | geschatte meetfout | nee, indirecte maat; validatie op basis van model | geen validatiemateriaal gebruikt |

Tabel 4: Samenvatting second opinion betrouwbaarheidsstudies beschreven in CINOP-eindrapport 'Verantwoording TGN'

## 4.8 Aanvullende validatiestudie door TNO

Geen van de validaties heeft uitsluitsel kunnen geven of de toegepaste spraaktechnologie aan de eisen voldoet. Een belangrijke reden hiervoor is dat de algemene werkwijze voor validatie van toetsen en examens afwijkt van de methodologie die gebruikt wordt in spraaktechnologisch onderzoek. De doelstelling van de validatiestudies uitgevoerd door CINOP is om evidentie te vinden omtrent validiteit, aangezien het niet mogelijk is om het ware taalvaardigheidsniveau van een kandidaat te achterhalen.

### *4.8.1 Gevolgde werkwijze*

TNO heeft in deze studie de werkwijze gevolgd die gebruikelijk is in spraaktechnologisch onderzoek. Het ware taalvaardigheidsniveau van een kandidaat wordt hierbij zo goed mogelijk benaderd. Hiertoe is een referentieoordeel bepaald dat gebaseerd is op meerdere menselijke beoordelingen van het taalvaardigheidsniveau. Door de TGN toetsuitslagen te vergelijken met het referentieoordeel kan nu wel een kwalitatieve uitspraak gedaan worden: Hoe lager het totale percentage onterechte toetsuitslagen hoe beter de toets is. De totale fout is echter afhankelijk van de onderliggende taalvaardigheidsverdeling in het materiaal dat voor de analyse is gebruikt. Door onderscheid te maken tussen de twee typen fouten die gemaakt worden, namelijk het percentage ten onrechte geslaagde en ten onrechte gezakte kandidaten, kan ook de totale fout voorspeld worden als de taalvaardigheidsverdeling in de praktijk afwijkt van de dataset waarop de foutenanalyse is uitgevoerd. In Tabel 5 staan de mogelijke toetsuitslagen en de twee typen fouten weergegeven (fouten worden weergegeven door het symbool '↠')

| | **geslaagd** | | **gezakt** | |
|---|---|---|---|---|
| **taalvaardigheid >= cesuur** | terecht geslaagd | ↠ | onterecht gezakt | ↣ |
| **taalvaardigheid < cesuur** | onterecht geslaagd | ↣ | terecht gezakt | ↠ |

Tabel 5: Mogelijke uitslagen in termen van (on)terecht zakken/slagen

Het percentage onterecht gezakte en onterecht geslaagde kandidaten zijn aan elkaar gekoppeld en zijn uitwisselbaar door het instellen van de drempelwaarde van toets (cesuur). Daarbij werken de percentages fouten die gemaakt worden als communicerende vaten; Als de drempelwaarde laag wordt ingesteld, dan is het percentage onterechte geslaagde kandidaten hoog. Wordt de drempelwaarde hoog ingesteld, dan is juist het percentage onterecht gezakte kandidaten hoog. Aangezien onbekend is of de cesuur optimaal is ingesteld, worden in deze analyse voor alle mogelijke drempelwaarden van de toets de toetsuitslagen berekend. De relatie tussen de twee typen fouten die gemaakt worden kunnen in een curve uitgezet worden, de zogenaamde DET (Detection Error Trade-off)-curve.

### *4.8.2 Succescriterium*

Een kwaliteitsmaat voor de toets is de afstand tot de linkeronderhoek van de DET-curve; Als de toets foutloos is, dan reduceert de DET-curve tot één punt dat zich onder de linkeronderhoek bevindt. Hoe verder de lijn in een DET-curve van de foutloze toets (linkeronderhoek) is verwijderd, hoe meer fouten er worden gemaakt. Een andere kwaliteitsmaat is de afstand tot de toets die niet discrimineert. Een toets die niet discrimineert geeft willekeurige toetsscores, de totale fout die de toets maakt is 50%. Hoe groter de afstand tot de toevalstoets, des te beter discrimineert de toets.

### *4.8.3 Resultaten*

De DET-curve voor de A1min en A2 cesuur en voor gokkans (simulatie van toets die willekeurig scoort) en de theoretische achtergrond van de DET-curves zijn te vinden in appendix F. De DET-curves laten zien dat de TGN toets – zoals verwacht - voor zowel de A1min als de A2 cesuur beter discrimineert dan gokkans. Verder is in alle figuren te zien dat de A2-curve dichter bij de linkeronderhoek ligt dan de A1min-curve. Er lijken dus minder fouten gemaakt te worden rond de A2 cesuur dan rond de A1min cesuur

### *4.8.4 Human benchmark*

De DET-curves geven een beeld van de geschatte fouten die de automatische toets maakt. Echter, hiermee is nog niet vastgesteld welke, kwantificeerbare, minimum-eisen aan de toets worden gesteld: wanneer worden de prestaties goed genoeg gevonden? TNO stelt dat de toets voldoende goed presteert als de fouten die de automatische toets maakt vergelijkbaar zijn met de fouten die menselijke beoordelaars maken op een vergelijkbare toets die mondelinge taalvaardigheid meet (‘human benchmark’). Hoewel data van enkele referentietoetsen (het NT2 staatsexamen [ref7] en de NT2 profieltoets [ref8]) beschikbaar zijn, valt de bruikbaarheid van de referentiedata voor deze toepassing te betwisten. In elk geval heeft TNO ook met deze (eigen) aanvullende analyses geen helder bewijs voor validiteit kunnen vinden.

*4.8.5 Conclusies*

- De TGN scoort - zoals verwacht - aanzienlijk beter scoort dan de toevalsscore.
- Er worden meer fouten gemaakt rond de A1min cesuur dan rond de A2 cesuur.
- Of de fouten die de toets maakt vergelijkbaar zijn met de fouten die menselijke toetsers maken is niet te zeggen omdat een geschikte "human benchmark" ontbreekt.

## 4.9 Optimale instelling van toetscesuur

Een DET-curve geeft de relatie weer tussen de twee typen fouten. Het exacte percentage fouten hangt af van de drempelwaarde die gekozen wordt voor de toets. Zowel TNO als CINOP stellen dat de instellingen van deze drempelwaarde (de toetscesuur) een beleidsmatige kwestie is en derhalve een keuze die de opdrachtgever dient te maken. Er zijn verschillen in de methodologie die CINOP en TNO hanteren om een optimale instelling van de toetscesuur te bepalen. Deze verschillen worden in onderstaande paragrafen samengevat.

*4.9.1 Optimale cesuurinstelling volgens methodiek CINOP*

In bijlage 3 van het eindrapport 'Verantwoording TGN' [ref3] staat beschreven wat de effecten zijn van de keuze van de cesuur op de geschatte percentages terechte en onterechte beslissingen. De percentages terechte en onterechte beslissingen worden door CINOP geschat op basis van een model. De parameters van dit model zijn geschat met behulp van een grote hoeveelheid data. Het model voorspelt hoe groot de kans is dat de ware score geobserveerd wordt. Op basis van de kansen die het model voorspelt kunnen de percentage terechte en onterechte beslissingen geschat worden. De totale onterechte beslissingen zijn minimaal bij een A1min cesuur van 16 en een A2 cesuur van 37.

*4.9.2 Optimale cesuurinstelling volgens methodiek TNO*

TNO maakt schattingen van percentages terecht en onterechte beslissing op basis van data (MFA-FIT en Amsterdam-data). Het ware taalvaardigheidsniveau van een kandidaat wordt benaderd door het referentieoordeel. TNO stelt voor om de optimale cesuurinstelling te bepalen door de minimale detectiekosten[9] te berekenen. Deze kosten hangen af van het gewicht dat wordt gegeven aan de twee types fouten die gemaakt worden, namelijk onterecht zakken en slagen èn van de verdeling van de te verwachten taalvaardigheidsniveaus van de kandidaten. Om de optimale instelling van de cesuur te kunnen bepalen is het nodig om een inschatting te maken van de verdeling van de taalvaardigheidsniveaus van de kandidaten. Verder kan de cesuur op beleidsmatige gronden aangepast worden. Indien het bijvoorbeeld ontoelaatbaar geacht wordt dat kandidaten ten onrechte zakken, kan de cesuur zo ingesteld worden dat het percentage ten onrechte gezakte kandidaten laag is.

[9] De totale detectiekosten ($C_{det}$) zijn gedefinieerd als:

$C_{det} = C_{ont.gezakt} P_{ont.gezakt} P_{>= cesuur} + C_{ont.geslaagd} P_{ont.geslaagd} P_{< cesuur}$

Hierbij zijn $C_{ont.\ gezakt}$ en $C_{ont.\ geslaagd}$ de kosten voor respectievelijk onterecht zakken of slagen, $P_{ont.\ gezakt}$ en $P_{ont.\ geslaagd}$ de percentages onterecht gezakte en onterecht geslaagde kandidaten en $P_{>= cesuur}$ de à priori waarschijnlijkheid dat een kandidaat over het vereiste taalvaardigheidsniveau beschikt.

# 5 Second opinion toets Kennis Nederlandse Samenleving (KNS)

## 5.1 Werkwijze

In paragraaf 5.1.1 wordt een second opinion gegeven over de validatiestudie uit het rapport "Verantwoording toets KNS' [ref9]. Achtereenvolgens worden geschreven: het succescriterium, de geschiktheid van dit succescriterium ten behoeve van de doelstelling van de toets en de representativiteit en onafhankelijkheid van het gekozen validatiemateriaal. In paragraaf 5.1.2 en 5.1.3 worden de door Ordinate en TNO uitgevoerde aanvullende validatiestudies beschreven. In de laatste paragraaf worden de conclusies over de toets KNS gegeven.

### *5.1.1 Validatie van de automatische scoring in de toets KNS*

*Succescriterium*
De kwaliteit van het machinale oordeel is gevalideerd door per opgave de totale toetsscore van de spraakherkenner te vergelijken met een totale toetsscore gebaseerd op een menselijke transcriptie van de antwoorden. De "human benchmark" is in dit geval de toetsscore verkregen op basis van de menselijke transcripties. De correlatie tussen het percentages correcte antwoorden (per opgave) volgens de spraakherkenner en op grond van menselijke oordelen is 0,89.

*Geschiktheid m.b.t. doel*
Er kleven twee nadelen aan de gekozen maat. Ten eerste is de correlatie per opgave berekend, terwijl de correlatie per kandidaat relevant is. Ten tweede legt de correlatiemaat (zoals ook bij de TGN opgemerkt) geen verband met de gekozen cesuur.

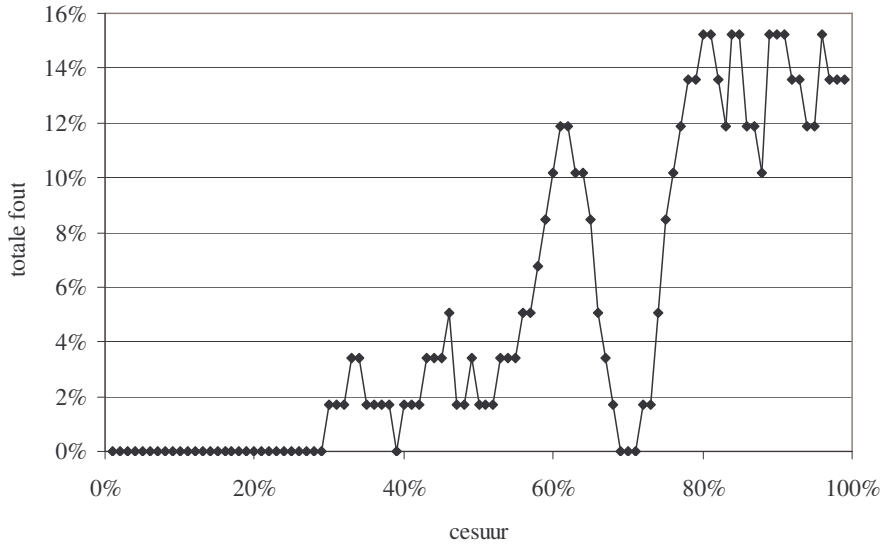
*Validatiemateriaal*
De akoestische modellen van de spraakherkenner zijn getraind met een onafhankelijke dataset (pretestdata van TGN) en de data lijken representatief te zijn voor de beoogde doelgroep. Het taalmodel van de spraakherkenner is echter getraind op dezelfde set waarop de validatie is uitgevoerd, waardoor het validatiemateriaal niet onafhankelijk is.

### *5.1.2 Aanvullende validatie door Ordinate*

Op verzoek van TNO heeft Ordinate de validatie herhaald op een onafhankelijke testset. Hierbij zijn de toetsscores per kandidaat berekend in plaats van per opgave. Voor deze validatie heeft Ordinate de antwoorden van 59 kandidaten door menselijke beoordelaars laten transcriberen. Op basis van de getranscribeerde antwoorden is vervolgens de menselijke toetsscore berekend. De gevonden correlatie tussen de menselijke en machinale toetsscore is 0,94, zie appendix G. Omdat ook de aanvullende validatie door Ordinate gebaseerd is op een correlatiecoëfficiënt wordt nog geen verband gelegd met de gekozen cesuur. Om deze reden heeft TNO met de door Ordinate geleverde data een aanvullende analyse uitgevoerd. Hierbij werd in eerste instantie aangenomen dat kandidaten die 80% scoren op de toets (op basis van de menselijke transcripties) over voldoende kennis van de Nederlandse samenleving bezitten, omdat dit de cesuur is die gebruikt is in [ref9]. De percentages terecht geslaagde en gezakte kandidaten zijn weergegeven in een DET-curve, zie appendix H.

*5.1.3 Aanvullende validatie door TNO*

Figuur 6 is, zoals gesteld, van toepassing bij een cesuur van 80%. Echter, deze is in tweede instantie gewijzigd in 70%. Helaas blijkt het, met de gegeven kleine dataset, onmogelijk om een DET-curve te berekenen voor een cesuur van 70%: in de dataset komen bij deze cesuur geen onterechte beslissingen voor, waardoor de analyse onmogelijk is (zie het strooidiagram in appendix G). Het is niet realistisch om te veronderstellen dat ook in de praktijk geen kandidaten zullen voorkomen die een score zullen behalen die rond de cesuur van 70% ligt. Om meer inzicht in deze materie te krijgen is de *totale* fout (percentage onterechte beslissingen) berekend die, op basis van de door Ordinate geleverde data, gemaakt zou worden bij een cesuur tussen 0 en 100%. De resultaten worden weergegeven in Figuur 1.




Figuur 1: Totale fout als functie van de cesuur voor de toets KNS (N=59)

Omdat de testset klein is, zijn de schattingen van de foutpercentages onnauwkeurig, met name onder de cesuur van 60% zijn er weinig datapunten (zie appendix G). Er lijkt geen fundamentele reden te bestaan waarom bij een cesuur van 70% een accuratere test zou worden verkregen dan, bijvoorbeeld, bij 60% of 80%. De totale fout voor de cesuur tussen 80-100% varieert tussen de 10-15%, voor een cesuur van 60% is de totale fout 12%. Het lijkt derhalve redelijk om voor de cesuur van 70% ook rekening te houden met een totale fout tussen de 10 en 15%.

Er is op dit moment geen geschikte benchmark beschikbaar om het geschatte totale fout mee te vergelijken. Een denkbare benchmark zouden bijvoorbeeld foutscores zijn die optreden bij (mondelinge) examens met vergelijkbare vragen, met een menselijke examinator. In zo'n geval is, bijvoorbeeld door misverstanden of fouten van de examinator, ook sprake van een zekere mate van onnauwkeurigheid. Subjectief lijkt een foutpercentage van 10-15% wellicht aan de hoge kant, maar hierbij moet worden aangetekend dat eenduidige transcriptie van sprekers met een (zeer) laag taalvaardigheidsniveau zelfs voor deskundigen lastig kan zijn.

### *5.1.4 Conclusies toets KNS*

Het gekozen validatiemateriaal in de aanvullende studie is onafhankelijk en representatief, maar voor een nauwkeurigere schatting van de foutenpercentages is een grotere dataset nodig. Op basis van de kleine validatieset is een schatting gemaakt van het percentage onterechte beslissingen, echter omdat gegevens van vergelijkbare menselijk afgenomen examens ontbreken en een norm voor validatie niet is gesteld kan niet worden geconcludeerd dat 10-15% (te) veel of (te) weinig is.

# 6 Conclusies en aanbevelingen

In dit hoofdstuk worden de conclusies en aanbevelingen kort en bondig samengevat. De conclusies en aanbevelingen worden nader toegelicht in hoofdstuk 0.

**Conclusies met betrekking tot Toets Gesproken Nederlands (TGN)**

Op basis van de validatiestudies worden door TNO de volgende conclusies getrokken:

- Er is evidentie gevonden dat de toets voldoende betrouwbare (=consistente) oordelen geeft
- Kandidaten met een sterk buitenlands accent worden door de spraakherkenner niet benadeeld met lagere deelscores
- Het trainingsmateriaal waar de spraakherkenner mee getraind is, is voor de meest frequente taalachtergronden representatief voor de praktijk

MAAR

- Er is geen bewijs gevonden dat de kwaliteit van de toets voldoende is, noch dat de kwaliteit onvoldoende is. Er zijn aanwijzingen dat de toets op een zinnige manier taalvaardigheid meet, maar hiermee is niet gegarandeerd dat de kwaliteit van de toets voldoende is voor alle taalvaardigheidsniveaus.
- TNO heeft aanwijzingen gevonden dat de kwaliteit van de machinale oordelen minder goed is rond de A1min cesuur. Doordat gegevens omtrent de kwaliteit van vergelijkbare menselijke taalvaardigheidsbeoordelingen ontbreken (‘human benchmark’) is het niet mogelijk om te concluderen of de fouten die rond de A1min cesuur gemaakt worden acceptabel zijn.

**Conclusies met betrekking tot de toets Kennis Nederlandse Samenleving (KNS)**

Op basis van statistische analyse van een kleine testset worden door TNO de volgende conclusies getrokken:

- Op basis van een statistische analyse van de kleine set test gegevens (N=59) wordt geschat dat voor 10-15% van de kandidaten de toetsuitslag onterecht is.
- Voor een nauwkeurigere schatting van de foutenpercentages is een grotere dataset nodig.
- Omdat gegevens van vergelijkbare menselijk afgenomen examens ontbreken en geen kwaliteitsnormen zijn gesteld kan niet worden geconcludeerd dat 10-15% (te) veel of (te) weinig is.

**Gebruik en reikwijdte van de conclusies**

De bovenstaande conclusies hebben betrekking op de prestatie van de spraaktechnologie in het testsysteem en de validatie van de cesuurinstelling van de toets. De kwaliteit van de spraaktechnologie wordt als voldoende beschouwd als de fouten die de automatische toets maakt vergelijkbaar zijn met de fouten die menselijke examinators maken; zowel de oordelen van een geautomatiseerde toets als de oordelen van menselijke beoordelaars zullen nooit 100% foutloos zijn.. Echter, omdat onbekend is welk percentage fouten in deze situatie door mensen gemaakt worden, kan niet worden geconcludeerd dat de spraaktechnologie beter, even goed of slechter presteert dan een menselijke examinator. De geschiktheid van het systeem als alternatief voor menselijke examinators kan daarom niet worden aangetoond, de ongeschiktheid ook niet.

De conclusies betreffen het te verwachten percentage ten onrechte zakken of ten onrechte slagen. Met dit percentage alleen is nog niet te zeggen of de toets acceptabel is. De uitgangspunten die beleidsmatig gaan gelden betreffende mogelijkheden om in beroep te gaan tegen de uitslag, mogelijke second opinion van twijfelgevallen of het kiezen van de cesuur niveaus bepalen uiteindelijk of het te verwachten percentage acceptabel zal zijn.

**Eindconclusie**

- Op basis van de uitgevoerde testen bestaat de indruk dat de TGN toets een consistente uitslag geeft.
- Op basis van de uitgevoerde testen kan niet worden geoordeeld of de TGN toets wel of niet geschikt is voor gebruik, omdat kwaliteitsnormen niet zijn gesteld.
- Op basis van een statistische analyse van een kleine set testgegevens wordt geschat dat voor 10-15% van de kandidaten de KNS toetsuitslag onterecht is. Of dit foutenpercentage acceptabel is hangt af van de beleidsmatige uitgangspunten.

Er zijn mogelijkheden om zowel de toets, de cesuurinstelling, als de validatie te verbeteren. De noodzaak hiertoe hangt af van de beleidsmatige uitgangspunten en de richting is aangegeven in de aanbevelingen

**Aanbevelingen**

Op basis van de uitgevoerde evaluatie beveelt TNO aan over te gaan tot invoering van de toetsen en daarbij de volgende stappen te ondernemen:

5) Verzamel voor de TGN en KNS onafhankelijke data die voldoende representatief zijn voor de praktijk

6) Formuleer succescriteria (normen) voor de TGN en de toets KNS. Daartoe is het nodig om meerdere menselijke beoordelingen te verzamelen:
   - Bepaal een ‘human benchmark’: Dit zijn menselijke beoordelingen die beschouwd kunnen worden als alternatief voor de automatische toets
   - Bepaal een referentieoordeel ten opzichte waarvan de foutpercentage van de toets en de ‘human benchmark’ berekend kunnen worden

7) Valideer de toetsen door de foutpercentages (onterecht zakken/slagen) op de nieuwe datasets te bepalen, en deze volgens het vastgestelde succescriterium te vergelijken met de ‘human benchmark’ (aanbeveling 2).

8) Verbeter de TGN rond de A1min cesuur

# 7 Toelichting op conclusies en aanbevelingen

## 7.1 Verschillen in de gebruikte TGN validatieprocedures tussen TNO en CINOP

### *7.1.1 Algemeen*

Het doel van de toepassing van spraaktechnologie in de TGN is de automatisering van de beoordeling van het taalvaardigheidniveau van een kandidaat (ten opzichte van een vastgesteld minimum). De kwaliteit[10] van de toets wordt als voldoende beschouwd als de gemaakte fouten te vergelijken zijn met het niet-automatische alternatief: deskundige menselijk beoordelaars;.. Zowel de oordelen van een geautomatiseerde toets als de oordelen van menselijke beoordelaars zullen nooit 100% foutloos zijn.

Uit bestudering van de rapportage van CINOP wordt duidelijk dat de door CINOP gehanteerde validatie-procedures op enkele punten wezenlijk verschillen van de standaard-werkwijze voor spraaktechnologie. In toelichting hierop stelt CINOP dat de gevolgde aanpak gangbaar en geaccepteerd is voor toetsen en examens. Er is een duidelijk verschil in perspectief: TNO bekijkt de set toetsen als een product dat op spraaktechnologie gebaseerd is, en waarvan bewezen moet worden (op de algemeen geaccepteerde manier) dat de spraaktechnologie naar behoren werkt. CINOP benadert de validatie van de toetsen op dezelfde wijze als elke willekeurige toets zonder spraakherkenning. Daarnaast is er een verschil in de keuze van het evaluatiemateriaal. Zoals eerder uiteengezet wordt voor spraaktechnologisch onderzoek geëvalueerd op materiaal dat onafhankelijk is van training en ontwikkeling/optimalisatie van het systeem. In de door CINOP gehanteerde werkwijze is het trainingsmateriaal over het algemeen apart strikt gehouden, maar is het ontwikkel- en validatiemateriaal niet altijd duidelijk gescheiden. De reden dat dit niet altijd is gebeurd heeft te maken met de historische ontwikkeling van de toets. In de loop van de tijd zijn namelijk een aantal aanvullende studies uitgevoerd waarvoor drie extra datasets zijn (Den Haag, Amsterdam, MFA-FIT) verzameld.

Om de verschillen in aanpak te concretiseren worden beide aanpakken hieronder (enigszins vereenvoudigd) samengevat. De beschrijving van deze aanpak van CINOP is mede geformuleerd aan de hand van mondelinge uitleg door CINOP en LTS.

### *7.1.2 TNO aanpak TGN validatie*

De aanpak van TNO is de standaard aanpak voor validatie van spraaktechnologie. Dit houdt in dat je een gefundeerde eis stelt aan de prestaties, en vervolgens bewijst dat aan die eis voldaan wordt. In dit geval kan zo'n eis bijvoorbeeld zijn: de spraakherkenner mag niet meer fouten maken dan een alternatief, conventioneel examen. Door bij dezelfde steekproef van proefpersonen scores te bepalen op de nieuwe toetsen en een conventioneel examen is een degelijk (statistisch gefundeerd) bewijs te leveren. Als zo'n bewijs ontbreekt is de validatie niet toereikend, ook al zijn er andere aanwijzingen dat de technologie wel degelijk goed functioneert. Er ligt een duidelijke en harde bewijslast bij de partij die de spraaktechnologie introduceert. TNO stelt dat bewijs volgens de hardere normen van spraaktechnologie in principe wel degelijk te leveren is, en de vastgestelde problemen (met name de genoemde vergelijking met resultaten van

[10] Een ander terminologie die in dit rapport gebruikt wordt is *validiteit* (zie appendix A voor een definitie)

conventionele toetsen) op te lossen zijn. Het gebruik van spraaktechnologie maakt de toetsen wel degelijk principieel anders dan andere toetsen; er mag dus niet zonder meer worden aangenomen dat het validatie-recept van CINOP dezelfde garanties geeft voor de prestaties van de toetsen als bij conventionele toetsen.

*7.1.3 CINOP aanpak TGN validatie*

Ook CINOP stelt concrete eisen aan de prestaties (onder andere in de vorm van betrouwbaarheidsscores). Echter, de stelling is dat hard bewijs dat beter wordt gepresteerd dan een referentie-toets (met menselijke examinatoren) nauwelijks te geven is: strikt genomen is onbekend wat *daadwerkelijk* het taalvaardigheidsniveau van de proefpersonen is geweest. Als een kandidaat voor een conventionele toets zakt maar voor de TGN slaagt, is het alsnog mogelijk dat de TGN gelijk had en de menselijke examinator ongelijk; ook bij conventionele toetsen worden fouten gemaakt. In plaats van rigide bewijs, zoals in de spraakherkenning wordt vereist, wordt gezocht naar evidentie voor kwaliteit. Allerlei relevante aspecten van de toets worden tegen het licht gehouden, en aan tests onderworpen. Als hierbij geen problemen worden geconstateerd (prestaties beneden vastgestelde criteria) dan wordt gesteld dat de toets valide is.

## 7.2 Keuze van maten voor vaststelling van kwaliteit van spraaktechnologie

TNO stelt dat het bewijs omtrent kwaliteit van de spraaktechnologie in de TGN niet zou moeten zijn gebaseerd op basis van een correlatiematen, maar op analyse van zak/slaagpercentages van de kandidaten. Een hoge correlatiecoëfficiënt kan immers niet garanderen worden dat het aantal kandidaten dat onterecht zakt of slaagt laag is. Zowel TNO als CINOP hebben een analyse uitgevoerd waarbij het percentage ten onrechte gezakte en geslaagde kandidaten is geschat. Er is echter één belangrijk verschil; CINOP schat het percentage onterechte beslissingen op basis van een model, terwijl TNO het percentage onterechte beslissingen schat op basis van data. TNO noemt een beslissing terecht als deze overeenkomt met het meerderheids- of unanieme oordeel van drie menselijke beoordelaars. Zowel TNO als CINOP stellen dat de instelling van de cesuur een beleidskwestie is; de keuze hangt af van het belang dat door de opdrachtgever wordt gehecht aan de gemaakte typen fouten.

Op basis van de validatiestudies worden door TNO de volgende conclusies getrokken:

- Er is evidentie gevonden dat de toets voldoende betrouwbare (=consistente) oordelen geeft
- Kandidaten met een sterk buitenlands accent worden door de spraakherkenner niet benadeeld met lagere deelscores
- Het trainingsmateriaal waar de spraakherkenner mee getraind is, is voor de meest frequente taalachtergronden representatief voor de praktijk

MAAR

- Er is geen bewijs gevonden dat de kwaliteit van de toets voldoende is, noch dat de kwaliteit onvoldoende is. Er zijn aanwijzingen dat de toets op een zinnige manier taalvaardigheid meet, maar hiermee is niet gegarandeerd dat de kwaliteit van de toets voldoende is voor alle taalvaardigheidsniveaus.
- TNO heeft aanwijzingen gevonden dat de kwaliteit van de machinale oordelen minder goed is rond de A1min cesuur. Doordat gegevens omtrent de kwaliteit van vergelijkbare menselijke taalvaardigheidsbeoordelingen (‘human benchmark’) ontbreken het niet mogelijk om te concluderen of de fouten die rond de A1min cesuur gemaakt worden acceptabel zijn.

De uitgevoerde analyses en de gebruikte data die zijn beoordeeld in deze ‘second opinion’ hebben een aantal nadelen. Ten eerste ontbreekt een ‘human benchmark’. Hiervoor zijn referentie-oordelen nodig: een oordeel dat het werkelijke taalvaardigheidsniveau van de kandidaat benadert. Deze referentiewaarden kunnen niet bestaan uit de bevindingen van slechts één enkele examinator per kandidaat; deze kan zich immers ook vergissen. Echter, door gemiddeldes van meerdere oordelen te hanteren kan een nauwkeurige benadering van de daadwerkelijke taalvaardigheid van de kandidaat worden verkregen. Door nu zowel het machinaal verkregen oordeel als een menselijke oordeel dat op een conventionele manier is verkregen te vergelijken met exact hetzelfde referentieoordeel kan een eenduidig succescriterium gesteld worden. De mogelijke tegenwerping dat zelfs een gemiddelde van vele examinatoren nog niet noodzakelijk de waarheid oplevert is irrelevant: als het referentieoordeel maar geaccepteerd wordt als een goede benadering van het werkelijke taalvaardigheidniveau, dan mag elke afwijking van deze referentie als onterechte uitslag worden aangemerkt. Ten tweede is bijna al het materiaal in de studie gebruikt voor ontwikkeling van de toets (zoals het trainen van de modellen van het automatische scoringscomponent, het bepalen van de schalings- en normeringsparameters, het bepalen van de itembank enz.). De enige dataset die niet betrokken is in ontwikkeling van de toets zijn de Amsterdam data. Het nadeel van deze data is:

- Het gaat om slechts twee soorten menselijke beoordelingen, het is moeilijk om op basis van deze beoordelingen zowel een ‘human benchmark’ als een referentieoordeel te bepalen
- Het gaat om relatief weinig data (voor 94 kandidaten zijn drie menselijke en machinale beoordelingen aanwezig)
- Het is onbekend in hoeverre de data representatief zijn voor de praktijk (het is bijvoorbeeld aannemelijk dat rond de A2 cesuur in de praktijk meer kandidaten voorkomen met A2 niveau)

## 7.3 Validatie aanpak toets KNS en het belang van onafhankelijke data

Het doel van de toepassing van spraaktechnologie in de toets KNS is het automatisch bepalen of een antwoord correct is of niet. De toepassing van spraaktechnologie kan succesvol worden genoemd als de fouten die de spraakherkenner maakt voldoende klein zijn ten opzicht van (conventionele) alternatieven om antwoorden te registreren.

De validatiestudie die is uitgevoerd voor de toets KNS is in eerste instantie niet correct uitgevoerd, omdat de validatie-set gebruikt is voor het trainen van het taalmodel van de spraakherkenner. Om deze reden heeft Ordinate de validatie nogmaals uitgevoerd voor een onafhankelijke validatie-set van 59 sprekers. Op basis van de door Ordinate geleverde data heeft TNO een validatie uitgevoerd waarbij het percentage onterecht zakken en slagen is berekend. Hierbij is aangenomen dat een kandidaat over voldoende kennis over de Nederlandse samenleving bezit als er een score van 80% of meer wordt gehaald op de toets KNS (op basis van de antwoorden die gescoord zijn op basis van menselijke beoordelingen). Echter, de voorgestelde cesuur is in tweede instantie gewijzigd in 70%. In de (kleine) testset komen weinig kandidaten voor die een toetsscore hebben rond de 70%. Op basis van statistische analyse van deze kleine, onafhankelijke testset worden door TNO de volgende conclusies getrokken:

- Op basis van een statistische analyse van de kleine set test gegevens (N=59) wordt geschat dat voor 10-15% van de kandidaten de toetsuitslag onterecht is.
- Voor een nauwkeurigere schatting van de foutenpercentages is een grotere dataset nodig

- Omdat gegevens van vergelijkbare menselijk afgenomen examens ontbreken en geen kwaliteitsnormen zijn gesteld kan niet worden geconcludeerd dat 10-15% (te) veel of (te) weinig is.

## 7.4 Aanbevelingen

Zoals eerder gesteld wordt het door TNO wel degelijk mogelijk geacht om te bewijzen dat de kwaliteit van een toets voldoende is. Hierin bestaat in taalvaardigheidstoetsen nog geen traditie, simpelweg omdat de op spraaktechnologie gebaseerde taalvaardigheidstoets een nieuw fenomeen is. TNO stelt vast dat de garanties waarop toepassers van spraaktechnologie normaal gesproken kunnen rekenen in dit geval niet onmiddellijk te geven zijn. Op grond van de bevindingen beveelt TNO aan over te gaan tot invoering van de toetsen en daarbij de volgende stappen te ondernemen:

5) Verzamel onafhankelijke data die volledig representatief zijn voor de praktijk
   - Geluidsfiles van een ruime hoeveelheid toetsafnames die ontstaan gedurende het gebruik van de TGN
   - Geluidsfiles van een ruime hoeveelheid toetsafnames die ontstaan gedurende het gebruik van de toets KNS
   - Neem bij de dataverzameling van de TGN twee verschillende steekproeven (datasets). Voor dataset 1 wordt een willekeurige steekproef van de kandidaten genomen, voor dataset 2 worden alleen kandidaten met een taalvaardigheidsniveau rond de cesuur geselecteerd. Dataset 1 is volledig representatief is voor de praktijk, terwijl dataset 2 gebruikt kan worden om de toets rond de cesuur te valideren en optimaliseren.
   - De vereiste hoeveelheid geluidsopnamen bestaat voor zowel de toets KNS als de TGN uit 200-300 toetsafnamen voor dataset 1 en 100-200 voor dataset 2, dus een totaal van 600-1000 toetsafnamen.

6) Formuleer succescriteria (normen) voor de TGN en de toets KNS
   - Laat alle opnamen beoordelen door (minimaal) vier menselijke beoordelaars die in staat zijn om een betrouwbaar oordeel over de taalvaardigheid van de kandidaat te vormen.
   - Gebruik per kandidaat steeds één menselijk oordeel als “human benchmark,” waarmee de geautomatiseerde toetsen kunnen worden vergeleken
   - Bepaal op basis van de overige oordelen (minimaal 3) een referentieoordeel. Omdat dit het gezamenlijk oordeel van meerdere deskundigen is, wordt geaccepteerd dat dit oordeel het ware taalvaardigheidsniveau van de kandidaat benadert
   - Bepaal een criterium voor succes van de geautomatiseerde toetsen relatief ten opzichte van de human benchmark.

7) Valideer de toetsen door de foutpercentages (onterecht zakken/slagen) op de nieuwe datasets te bepalen, en deze volgens het vastgestelde succescriterium te vergelijken met de human benchmark (aanbeveling 2).

8) Verbeter de TGN rond de A1min cesuur
   - Hertrain de spraakherkenner met meer materiaal van lage taalvaardigheidsniveaus.
   - Voor hertraining van de spraakherkenner kan gebruik gemaakt worden van de data die beschikbaar zijn gekomen in de aanvullende experimenten die na de pretest zijn uitgevoerd (Den Haag, Amsterdam, MFA-FIT), vooropgesteld dat

er nieuwe (onafhankelijke) validatie-datasets beschikbaar komen (aanbeveling 1)

- Hertrain/herschaal andere componenten van het automatische scoringssysteem op vergelijkbare manier als de spraakherkenner.
- Pas eventueel meer complexe classificatie-algorithmen dan een eenvoudige cesuur toe.

# Referenties

[ref1] Commissie Franssen/ Franssen, J. et al., Inburgering getoetst. Advies over het niveau van het inburgeringsexamen in Nederland. Den Haag, 2004.

[ref2] Verantwoording Inburgeringexamen Nederlands als tweede taal. Ten behoeve van de opdrachtgever, CINOP in samenwerking met Language Testing Services en Ordinate, mei 2005.

[ref3] Verantwoording Inburgeringexamen Nederlands als tweede taal. Ten behoeve van de opdrachtgever, CINOP in samenwerking met Language Testing Services en Ordinate, 19 september 2005.

[ref4] SET-10;Test Description & Validation Summary, Ordinate Corporation, 2004.

[ref5] Spoken Spanish Test; Test Description & Validation Summary (v.1.1), Ordinate Corporation.

[ref6] “The DET curve in assessment of detection task performance”, A. Martin, G. Doddington, T. Kamm, M. Ordowski & M. Przybocki, Proceedings of Eurospeech 1997.

[ref7] Staatsexamen NT2. Het gewenste niveau. Deel 2B Cesuurindicatie Spreken en Schrijven Programma II. Aanvullend onderzoek, CITO, December 2004.

[ref8] “De NT2 Profieltoets in de praktijk”, L. Bekkers, B. Bossers, M. Jetten en J. Soeting, CITO/ICE, 1999.

[ref9] Verantwoording Inburgeringexamen Kennis van de Nederlandse Samenleving. Ten behoeve van de opdrachtgever, CINOP in samenwerking met Language Testing Services en Ordinate.

[ref10] van Leeuwen, D.A., Martin, F.A., Przybocki, M.A., Bouten, J.S., “NIST and NFI-TNO Evaluations of Automatic Speaker Recognition”, Computer Speech and Language, 2005.

[ref11] Cohen, J.A., “Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220, 1968.

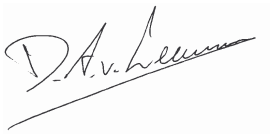
[ref12] Landis, J.R., Koch, G.G., “The measurement of observer agreement for categorical data. Biometrics, 33, 159-174, 1977.

# 8 Ondertekening

Soesterberg, oktober 2005

Judith Kessens

Sander van Wijngaarden

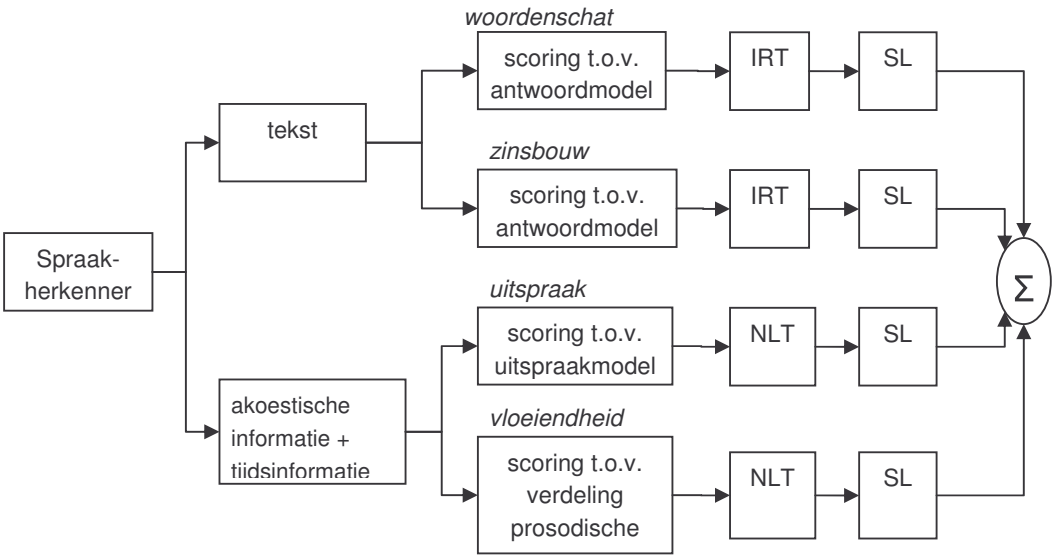


David van Leeuwen

# A Terminologie

| | | |
|---|---|---|
| acceptabele fouten | = | fouten die toelaatbaar zijn op basis van de beleidsmatig uitgangspunten |
| betrouwbaarheid | = | consistentie van de uitkomst van de toets bij herhaalde meting |
| cesuur toetscore | = | toetsscore waarboven/onder een kandidaat voor de toets slaagt/zakt |
| correlatie | = | mate waarin twee variabelen volgens lineair verband samenhangen |
| DET-curve | = | curve die het verband (Detection Error Trade-off) weergeeft tussen het percentage onterecht geslaagde en gezakte kandidaten |
| ontwikkeling | = | instelling/optimalisering van de parameters van het systeem |
| onafhankelijke data | = | data die niet gebruikt zijn voor training/ontwikkeling van het systeem |
| representatieve data | = | data die overeenkomen met de praktijk |
| training | = | schatting van de parameters van een model |
| validatie | = | meting van de fouten die het geoptimaliseerde systeem maakt |
| validatiestudie | = | studie die evidentie of bewijs levert omtrent validiteit van de toets |
| validiteit | = | mate waarin de toets de waarheid meet |
| human benchmark | = | systeem dat vergelijkbaar is met het automatische systeem en waarbij de geautomatiseerde taken door mensen worden uitgevoerd. |

# B Opbouw en training van het automatische scoringscomponent in de TGN




Figuur 2: Opbouw automatische scoringscomponent TGN

*Opbouw*

In de TGN is automatische spraakherkenning ingezet als onderdeel van het automatische scoringssysteem. Voor de inhoudelijke en kwalitatieve scoring zet de spraakherkenner het akoestische signaal in verschillende formaten om: 1) tekst en 2) tijdsinformatie en informatie over het akoestisch signaal. Tijdens automatisch scoring wordt de informatie die de spraakherkenner genereert getransformeerd tot een score zoals is te zien in Figuur 2. Hierbij wordt gebruik gemaakt van een aantal modellen en transformaties die hieronder achtereenvolgens worden beschreven.

1) *Tekst.* Voor de inhoudelijke scoring is het nodig om te weten wat de inhoud van het antwoord is. De spraakherkenner herkent gevulde pauzes (eh, ehm) en niet-spraak geluiden en verwijdert deze voordat de inhoudscoring plaatsvindt. Verder wordt spraak aan het begin of einde van het correcte antwoord genegeerd. Tijdens inhoudelijke score voor woordenschat wordt een dichotome score voor inhoudelijke correctheid berekend: "1" als het antwoord in het geheel correct is en "0" als dat niet het geval is. Voor de inhoudelijke scoring voor zinsbouw wordt gebruik gemaakt van een polytome scoring die afhankelijk is van de lengte van de zin. Als antwoordmodel wordt een database gebruikt die alle correcte antwoorden bevat. Vervolgens worden de inhoudelijke deelscores niet-lineair getransformeerd door een "Item Response Theory (IRT) component.
2) *Tijdsinformatie en informatie over het akoestisch signaal.* Voor de kwalitatieve scoring is gedetailleerde tijdsinformatie en informatie over het akoestisch signaal nodig. Tijdens oplijning wordt de spraakherkenner gebruikt om te bepalen welk deel van het akoestisch signaal overeenkomt met een woord, syllabe of foneem (klank). Op basis van de tijdsinformatie en de akoestische informatie kan een aantal prosodische maten berekend worden (o.a. begin en einde van een woord, de duur en

lengte van pauzes, syllabes en fonemen). De prosodische maten worden gescoord t.o.v. de "prosody performance distributions". De uitspraakscore wordt berekend t.o.v. uitspraakscores verkregen met referentie akoestische modellen. De kwalitatieve maten worden vervolgens getransformeerd door een "Niet Lineaire Transformatie" (NLT).

Tenslotte worden alle deelscores lineair geschaald en gelimiteerd door de "Scaling and Limiting (SL) component". De totale scores variëren tussen 0 en 90. Scores groter dan 80 worden gerapporteerd als 80 en scores kleiner dan 10 worden gerapporteerd als 10.

*Training*

Voordat het automatische scoringssysteem gebruikt kan worden dienen alle gebruikte modellen getraind te worden. Voor de ontwikkeling van de toets en training van het automatische scoringssysteem is een database opgenomen. Deze database – de pretest database – bestaat uit opnames van in totaal 132.000 antwoorden van 836 moedertaal sprekers (MS) en 1518 niet-moedertaal sprekers (NMS). Voor validatie-doeleinden is een subset bestaande uit 139 sprekers apart gehouden, die niet gebruikt is voor de training van het automatische scoringssysteem en de ontwikkeling van de toets. De verdeling van het aantal antwoorden en sprekers over validatie en training set is weergegeven in Tabel 6.

| | | **sprekers** | **antwoorden** |
|---|---|---|---|
| **validatie** | **MS** | - | - |
| | **NMS** | 139 | 7.000 |
| **training** | **MS** | 836 | 59.000 |
| | **NMS** | 1379 | 66.000 |

Tabel 6: Aantal sprekers en antwoorden uit pretest gebruikt voor validatie en training

Tijdens de trainingsfase worden drie soorten modellen getraind:

1) Modellen van de spraakherkenner

De spraakherkenner bestaat uit akoestische modellen en taalmodellen. Om optimale herkenning te bereiken wordt gebruikt gemaakt van een item-specifiek taalmodel, dat wil zeggen dat voor ieder item aparte à priori waarschijnlijkheden geschat worden. De akoestische en taalmodellen van de spraakherkenner zijn getraind met spraakmateriaal dat zowel NMS en MS bevat. Naast aanpassing van de akoestisch modellen aan de NMS zijn geen andere aanpassingen gedaan om de niet-moedertaal specifieke uitspraakvariatie te modelleren.

2) De referentiemodellen

Voor de inhoudelijke scoring wordt aan de hand van een antwoordmodel bepaald of een antwoord correct is of niet. Om het antwoordmodel te trainen is door menselijke beoordelaars bepaald of een antwoord juist is of niet. Voor de kwalitatieve scoring worden andersoortige referentiemodellen gebruikt. De uitspraak wordt gescoord ten opzichte van referentie akoestische modellen. Voor de vloeiendheid de distributies van de prosodische kenmerken als referentie gebruikt.

3) De schalingsmodellen

Voor de inhoudelijke deelscores (woordenschat en zinsbouw) zijn de schalingsmodellen IRT modellen. Tijdens training worden de parameters van deze IRT modellen geschat. Hiertoe worden alle antwoorden in de trainingset met de spraakherkenner herkend. De herkende uitvoer van de spraakherkenner worden gebruikt om de IRT modellen te trainen. Voor de kwalitatieve deelscores (vloeiendheid en uitspraak) worden NLT modellen getraind. Invoer zijn de prosodische en de uitspraakmaten. De gewenste uitvoer zijn de vloeiendheids- en uitspraakscores van de menselijke beoordelaars. Voor

training van alle schalingsmodellen wordt gebruik gemaakt van een deelverzameling van de pretest dataset (95% NMS, 5% MS). Met 2/3 van dit materiaal worden de schalingsmodellen getraind, 1/3 deel van het materiaal dient voor testen. In totaal ging het om ongeveer 150-250 iteraties per model in 200 runs.

# C Opbouw en training automatische scorings-component toets KNS

Voor de ontwikkeling van de toets en training van de modellen van de spraakherkenner is een database - de pretest database - opgenomen bestaande uit 25,800 antwoorden op 148 toets-items. Deze antwoorden zijn verkregen voor 228 mannelijke en 527 vrouwelijke sprekers met variërende taalachtergrond en taalvaardigheidniveau. Bij deze toets gaat het alleen om de inhoudelijke correctheid van de antwoorden, de kwalitatieve scores zijn niet van belang. Hiermee wordt het aantal modellen dat getraind moet worden gereduceerd:

a) Modellen van de spraakherkenner.
Voor de spraakherkenner worden dezelfde akoestische modellen gebruik als in de TGN. Het taalmodel is weer een item-specifiek taalmodel dat is getraind op basis van menselijke transcripties van de antwoorden in de pretest database.

b) Referentiemodellen
Aan de hand van een antwoordmodel wordt bepaald of een antwoord correct is of niet. Het antwoordmodel bevat alle correcte antwoorden. Daartoe is door menselijke luisteraars voor alle antwoorden uit de pretest database beoordeeld of een antwoord correct is of niet. Verschil met de TGN is dat hierbij ook uitspraakvarianten van het antwoord zijn meegenomen; bijvoorbeeld zowel de uitspraak “fies” en “vies” voor het woord “fiets” is goed gerekend.

c) Schalingsmodellen
De ruwe toetsscores worden met behulp van IRT-modellen omgezet naar de uiteindelijke toetsscore.

# D Rekenvoorbeeld correlatie

Op basis van de data die staan geplot in Figuur 6.2 op pagina 92 van [ref3] kan het percentage ten onrechte gezakte[11] en geslaagde[12] kandidaten berekend worden voor de A2 cesuur. Voor de A1min cesuur zijn de berekeningen niet uitgevoerd aangezien er erg weinig kandidaten een niveau kleiner dan A1min hebben (naar schatting 6 kandidaten). Er wordt de aanname gedaan dat een toetsscore van 37 of meer overeenkomt met een taalvaardigheidsniveau van A2. Deze relatie tussen toetsscore en CEF-niveau is overgenomen uit tabel 7.10 uit [tabel 5.5, ref 9][13]. Het percentage ten onrechte geslaagde kandidaten is 9% en het percentage ten onrechte gezakte kandidaten is 21%. De totale fout, oftewel het percentage onterechte toetsuitslagen, is 12%. Deze percentages zijn niet af te leiden uit de correlatiecoëfficiënt van 0,94.

| | |
|---|---|
| **% ten onrechte zakken** | 9% |
| **% ten onrechte slagen** | 21% |
| **totale fout** | 12% |

Tabel 7: Samenvatting rekenvoorbeeld

[11] Het percentage ten onrechte gezakte kandidaten is berekend t.o.v. het aantal kandidaten met niveau A2, zie appendix F

[12] Het percentage ten onrechte geslaagde kandidaten is berekend t.o.v. het aantal kandidaten met niveau <A2, zie appendix F

[13] Omdat het niet zeker is dat er een één-op-één relatie bestaat tussen menselijke toetsscore en werkelijk CEF taalvaardigheidsniveau dienen er geen conclusies getrokken te worden op basis van de getallen in dit rekenvoorbeeld; het dient slechts ter illustratie dat een hoge correlatiecoëfficiënt niet een garantie is voor een succesvolle zak/slaag-toets

# E Aanvullende analyses voor mens-mens, mens-machine en machine-machine oordelen

*Percentages overeenstemming voor alle mogelijke mens-machine paren*
TNO heeft CINOP gevraagd de percentages overeenstemming voor alle mogelijke mens-machine paren te berekenen. Uit deze analyse blijkt dat de mate van overeenstemming van de afzonderlijke paren niet veel verschilt van de mate van overeenstemming tussen het maximale oordeel van de twee menselijke beoordelingen en het maximale oordeel van de twee TGN toetsen.

*Cohen's kappa*
TNO heeft CINOP ook gevraagd om de mate van overeenstemming uit te drukken in een andere maat, namelijk de Cohen's κ. Deze maat corrigeert voor overeenstemming op basis van kans ($P_{kans}$). De overeenstemming op basis van kans is de overeenstemming die wordt bereikt met een machine die willekeurige beoordelingen produceert. Cohen's κ drukt uit in hoeverre de overeenstemming hoger is dan te verwachten is op basis van kans en is als volgt gedefinieerd [ref11]:

$$\text{Cohen's } \kappa = (P_{agree} - P_{kans})/(1 - P_{kans})$$

De resultaten van deze berekeningen en de overeenkomstige kwalificaties volgens Landis & Koch [ref12] staan vermeld in onderstaande tabel 8.

| **Oordeel 1** | **Oordeel 2** | **A1min** | | **A2** | |
|---|---|---|---|---|---|
| | | **Cohen's κ** | **kwalificatie** | **Cohen's κ** | **kwalificatie** |
| **Interviewer** | **Loopbaan** | 0.51 | matig | 0.58 | matig |
| **Beoordelaar** | **Loopbaan** | 0.48 | matig | 0.58 | matig |
| **Mens-Max. v. 3** | **TGN-Max. v. 2** | 0.31 | beperkt | 0.48 | matig |
| **TGN-1** | **TGN-2** | 0.57 | matig | 0.69 | goed |

Tabel 8: Cohen's kappa voor mens-mens, mens-machine en machine-machine oordelen

# F DET-curves

In een DET (Detection Error Trade-off) – curve wordt het verband tussen de percentages ten onrechte gezakte en geslaagde kandidaten weergegeven voor verschillende drempelwaarden van de toets. Het taalvaardigheidsniveau van een kandidaat wordt benaderd door een referentieoordeel dat gebaseerd is op meerdere menselijke beoordelingen van het taalvaardigheidsniveau. Op basis van de toetsuitslag bij een specifieke drempelwaarde, de cesuur (A1min of A2) en het taalvaardigheidniveau kan het percentage ten onrechte geslaagde ($P_{ont.geslaagd}$) en gezakte kandidaten ($P_{ont.gezakt}$) worden geschat. Essentieel voor het gekozen criterium is dat de schattingen voor de percentages onterecht zakken/slagen worden gescheiden. De percentages worden daarom niet berekend t.o.v. het totale aantal kandidaten dat meedoet aan de toets, maar t.o.v. het aantal kandidaten dat respectievelijk wel of niet over het vereiste taalvaardigheidniveau beschikt:

$$P_{ont.gezakt} = N_{ont.\ gezakt} / N_{\ taalvaardigheid >= cesuur} \quad (1)$$

$$P_{ont.geslaagd} = N_{ont.\ geslaagd} / N_{\ taalvaardigheid < cesuur} \quad (2)$$

Als bijvoorbeeld 40 van de 100 kandidaten volgens het referentieoordeel over het vereiste taalvaardigheidniveau beschikken en 10 van deze kandidaten ten onrechte zakken dan is het geschatte percentage onterecht zakken 25% (10/40).

De geschatte percentages onterecht gezakte en geslaagde worden tegen elkaar uitgezet in een DET-curve [ref6]. In een DET-curve wordt gebruik gemaakt van een niet-lineaire schaal omdat de curves rechte lijnen worden als de data normaal verdeeld zijn. De berekening van het betrouwbaarheidinterval in de DET-curve is gebaseerd op de aanname dat de zak/slaag-scores binomiaal verdeeld zijn [ref10]. De *standard error* (s) wordt als volgt berekend:

$$s_{ont.gezakt}^2 = P_{ont.gezakt}(100\% - P_{ont.gezakt}) / N_{>= cesuur} \quad (3)$$

$$s_{ont.geslaagd}^2 = P_{ont.geslaagd}(100\% - P_{ont.geslaagd}) / N_{< cesuur} \quad (4)$$

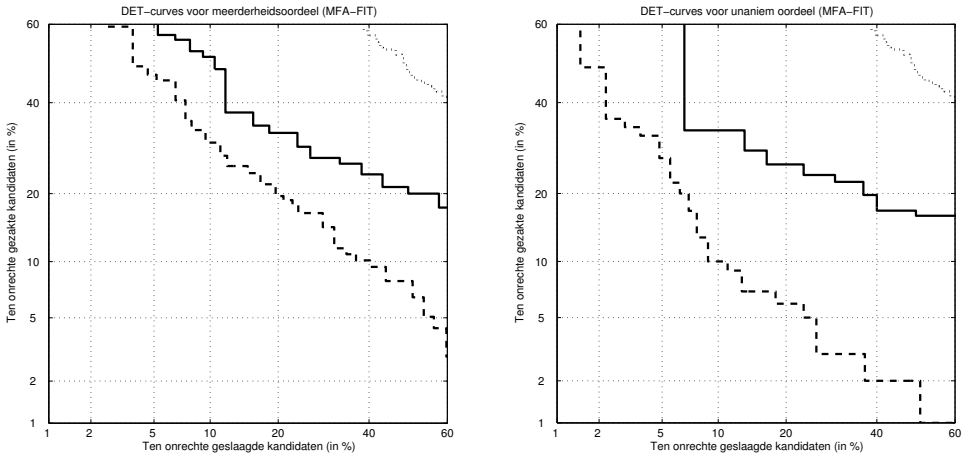
Het 95% betrouwbaarheidsinterval wordt gegeven door:

$$P_{ont.\ gezakt} \pm 1.96\ s_{ont.gezakt} \text{ en } P_{ont.\ geslaagd} \pm 1.96\ s_{ont.geslaagd} \quad (5)$$

De DET-curves voor de A1min en de A2 cesuur zijn berekend voor zowel de MFA-FIT en Amsterdam data om drie redenen: 1) deze datasets lijken het meest representatief te zijn voor de doelgroep(en) van de toets, 2) voor beide datasets is voor een groot aantal kandidaten een drietal betrouwbare menselijke oordelen voorhandig, en 3) de toets is drie keer afgenomen (1 oefentoets + 2 toetsen). Alleen kandidaten die zowel een oefentoets als twee versies van de toets hebben uitgevoerd èn waarvoor 3 menselijke beoordelingen beschikbaar zijn, zijn geanalyseerd. De score die behaald is op de oefentoets is niet gebruikt. De scores van de twee versies van de toetsen zijn beiden als datapunt meegenomen. In totaal gaat het om 243 kandidaten (486 datapunten) voor de MFA-FIT data en om 94 kandidaten (188 datapunten) voor de Amsterdam data.

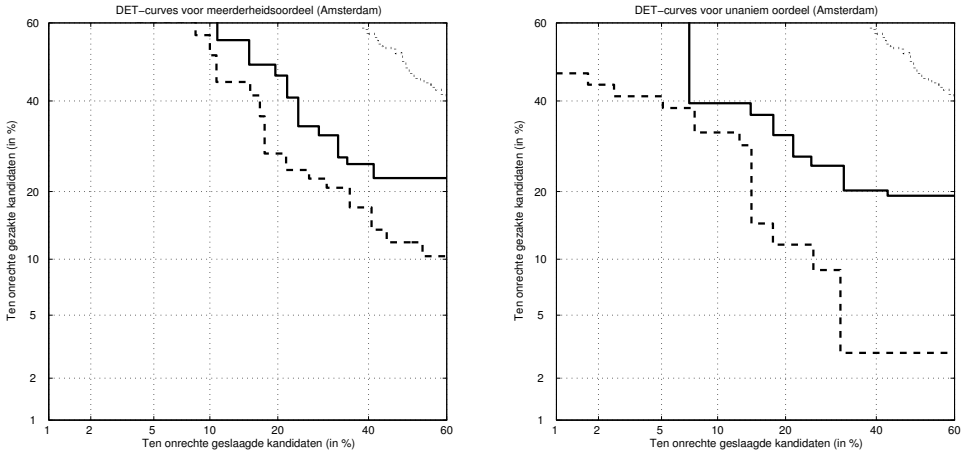
Er is gekozen voor twee typen referentieoordelen: het meerderheidsoordeel van de 3 beoordelaars en het unanieme oordeel van de 3 beoordelaars (niet unanieme oordelen worden niet meegenomen in deze analyse). De menselijke beoordelingen waren afkomstig van; 1) een interviewer (op basis van een gestructureerd interview), 2) een onafhankelijke beoordelaar die bij hetzelfde interview aanwezig is, 3) een beoordelaar die een oordeel geeft op basis van een audio-opname van hetzelfde interview (MFA-FIT), of een beoordelaar die het oordeel geeft op basis van een loopbaangesprek (Amsterdam).

De DET-curves zijn weergegeven in Figuur 3a voor het meerderheidsoordeel en in Figuur 3b voor de unanieme oordelen. De curves voor de A1min zijn weergegeven met '—' en de curves voor A2 met '----'. Daarnaast is met '····' de DET-curve weergegeven die overeenkomt met gokkans[14].




Figuur 3a: DET-curves voor a) meerderheidsoordeel en b) unanieme oordeel voor MFA-FIT data

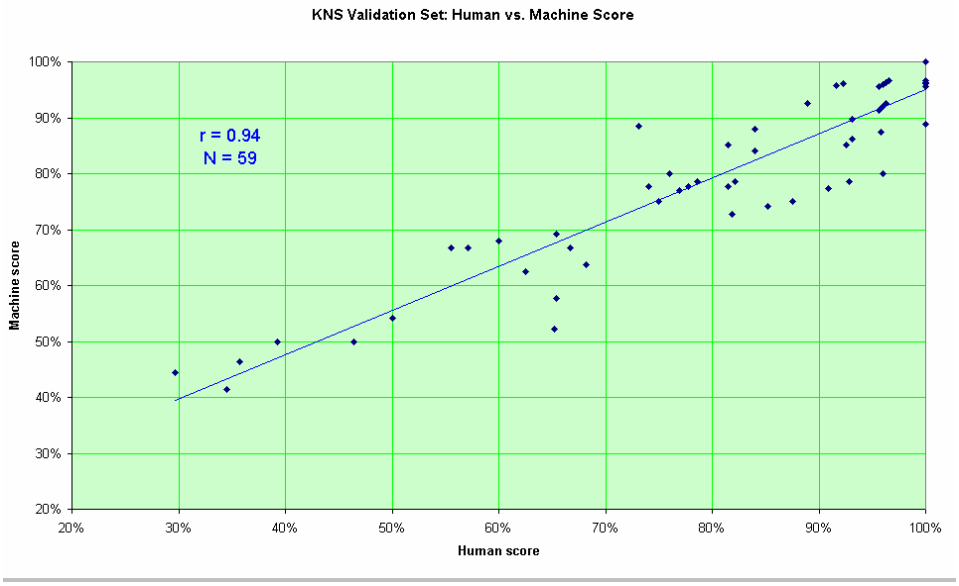
(('- - -' = A2,'—' =A1min, '····'= gokkans)




Figuur3b: DET-curves voor a) meerderheidsoordeel en b) unanieme oordeel voor Amsterdam data
('- - -' = A2,'—' =A1min, '····'= gokkans)

[14]Deze DET-curve is berekend op basis van een simulatie van een toets die willekeurig scoort (650 datapunten).

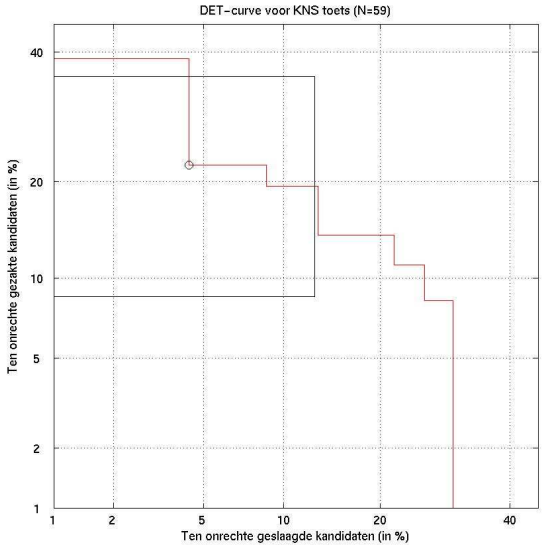
# G Correlatie menselijke en machinale scores voor toets KNS




Figuur 3: Strooidiagram voor menselijke versus machinale scores (per kandidaat, N=59)

In Figuur 3 staat het strooidiagram voor de menselijke versus machinale scores. Het is te zien dat er met name voor de lage cesuren weinig datapunten zijn. Ook deze figuur laat zien dat het gewenst is een grotere validatie-set te gebruiken.

# H DET-curve toets KNS voor een cesuur van 80%




Figuur 4: DET-curve voor de toets KNS (N=59)

De rechthoek in Figuur 4 geeft het 95%-betrouwbaarheidsinterval (zie appendix F) aan voor een punt op de DET-curve, aangegeven met het symbool ‘o’. Aangezien het aantal datapunten klein is (N=59) is het 95%-betrouwbaarheidsinterval groot. Het 95%-betrouwbaarheidsinterval kan gereduceerd worden door meer datapunten te gebruiken.