Verantwoording
Toets Gesproken
Nederlands

Titel	Verantwoording Toets Gesproken Nederlands
Projectnummer	10206.01
Auteurs	Anne Kerkhoff, Petra Poelmans (CINOP)
	John H.A.L. de Jong (LTS)
	Matthew Lennig (Ordinate)

Pettelaarpark 1
Postbus 1585
5200 BP 's-Hertogenbosch
Tel: 073-6800800
Fax: 073-6123425
www.cinop.nl

Verantwoording

Toets Gesproken Nederlands

Ontwikkeld in opdracht van het Ministerie van Justitie van het Koninkrijk der Nederlanden

CINOP: **Anne Kerkhoff, Petra Poelmans**
LTS: **John H.A.L. de Jong**
Ordinate: **Matthew Lennig**

Inhoud

Samenvatting

Achtergrond
In december 2003 gaf het Ministerie van Justitie CINOP de opdracht om een examensysteem te ontwikkelen voor de toetsing van mondelinge vaardigheden in het Nederlands van buitenlanders. Het examensysteem moest in de eerste plaats geschikt zijn voor de toetsing van de taalvaardigheid in het Nederlands van buitenlanders die zich duurzaam in Nederland willen vestigen. Contractueel werd vastgelegd dat de toets daarnaast ontwikkeld werd om eventueel ook ingezet te worden in het kader van een examenstelsel voor inburgering in Nederland dat gebaseerd zou zijn op de Europese taalportfoliomethodiek.
In dit rapport wordt de totstandkoming van de Toets Gesproken Nederlands (TGN) verantwoord. De toets is ontwikkeld door CINOP in nauwe samenwerking met Language Testing Services in Velp en het Amerikaanse bedrijf Ordinate.

Het toetssysteem
Gelet op de specifieke eisen aan een toetssysteem dat wereldwijd gebruikt moet worden, is gekozen voor een operationalisering met gebruikmaking van spraaktechnologie. De technologie van de Toets Gesproken Nederlands is gebaseerd op een toetssysteem dat Ordinate heeft ontwikkeld op basis van onderzoek naar spraakherkenning, statistische modellen, linguïstiek en toetstheorieën. Op basis van die kennis heeft Ordinate een automatisch scoringssysteem ontwikkeld dat specifiek geschikt is voor de scoring van de spraak van taalleerders. De technologie is operationeel voor de toetsing van Engels en Spaans en bevat een aantal taalonafhankelijke componenten die kunnen worden ingezet voor iedere taal waarvoor het spraakherkenningsysteem is getraind. Met behulp van spraak van ruim 1.500 leerders van het Nederlands en van ruim 800 moedertaalsprekers van het Nederlands is het scoringssysteem getraind voor het Nederlands.

Wat meet de toets?
De toets meet het gemak waarmee kandidaten in normaal conversatietempo gesproken Nederlands kunnen verstaan en begrijpen en hierop adequaat en verstaanbaar in het Nederlands kunnen reageren. Begrip van gesproken Nederlands (luistervaardigheid) is een preconditie bij alle items. Daarnaast worden bepaalde aspecten van de vaardigheid om begrijpelijk en zinvol Nederlands te spreken getoetst.

De niveaus: A1-min en A2
De normering van de TGN is gerelateerd aan de niveaus van het Gemeenschappelijk Europees Referentiekader (CEF, Common European Framework of Reference for Languages, Council of Europe, 2001). In principe heeft de toets een bereik vanaf géén enkele beheersing van het Nederlands tot en met een nagenoeg perfecte beheersing. Met het oog op de doelen die de opdrachtgever met de toets voor ogen heeft, is bij de ontwikkeling van de toets speciale aandacht besteed aan het meten van de laagste niveaus: van A1-min tot en met A2. Op basis van de adviezen van de Commissie Franssen aan de Minister voor Vreemdelingenzaken en Integratie zijn bij de constructie van de toets de volgende niveauomschrijvingen gehanteerd:

A1-min
Kan met behulp van losse woorden zaken van direct persoonlijk belang communiceren.
Gebruikt losse woorden, enkele standaarduitdrukkingen en elementaire beleefdheidsfrases, maar is vanwege uitspraak moeilijk te verstaan. Begrijpt eenvoudige direct tot hem/haar gerichte en met zorg gesproken vragen naar of mededelingen over personalia en een beperkt aantal concrete alledaagse begrippen. Kan vragen over dergelijke zaken soms ook met een of meer losse woorden beantwoorden. Conversatie is echter niet mogelijk.

A2
Communiceert basisinformatie over werk, achtergrond, familie, vrije tijd, et cetera.
Kan zichzelf in korte zinnen verstaanbaar maken, hoewel pauzes, valse starts, en herformuleringen evident aanwezig zijn. Uitspraak is over het algemeen helder genoeg om te worden verstaan ondanks een duidelijk buitenlands accent. Gebruikt een beperkt aantal eenvoudige structuren correct, maar maakt systematisch elementaire fouten. Kan woordgroepen verbinden met eenvoudige voegwoorden zoals “en”, “maar”, en “omdat”. Kan zich tot hem/haar richtende, duidelijk sprekende moedertaalsprekers verstaan, wanneer zonodig om herhaling gevraagd kan worden.

De afnamecondities
Bij afname van het examen op de Nederlandse ambassades en consulaten wordt gebruik gemaakt van een vaste telefoon. Voorafgaand aan de toets krijgen kandidaten een mondelinge instructie in hun moedertaal of in een andere taal die zij zeggen voldoende te beheersen om de instructies te kunnen begrijpen. De instructies worden gegeven door een lid van het ambassadepersoneel dat is aangewezen als examenleider. In gevallen waarin er géén examenleider beschikbaar is die een taal beheerst waarin zinvol met een kandidaat gecommuniceerd kan worden, krijgt de kandidaat de gelegenheid zelf iemand mee te brengen die als tussenpersoon kan functioneren. Bij de instructie aan kandidaten die kunnen lezen kan gebruik worden gemaakt van een schriftelijke ‘instructiekaart’. Wanneer de kandidaten hebben aangegeven de instructies te hebben begrepen, wordt via de telefoon een verbinding gelegd met de computer waarin het toetssysteem is opgeslagen. Nadat het systeem de kwaliteit van de verbinding heeft getest en goed heeft bevonden, vindt de toetsafname automatisch plaats.
Kandidaten krijgen achtereenvolgens vier groepen opgaven: zinnen herhalen, korte vragen beantwoorden, zinnen herhalen en tegenstellingen benoemen. Elke groep opgaven wordt voorafgegaan door een korte instructie en twee voorbeelden. De instructies en de voorbeelden zijn duidelijk gesproken door professionele stemacteurs.

In totaal krijgen kandidaten 48 opgaven waarvan er 45 worden gebruikt om de eindscore van de kandidaat te bepalen. Per soort opgave dient de eerste opgave alleen als kennismaking. Na afloop van de eigenlijke toets krijgen kandidaten, nog steeds via de telefoon, twee korte verhaaltjes te horen die ze moeten navertellen. De reacties op deze laatste twee opgaven worden niet automatisch gescoord en spelen geen rol bij het bepalen van de score van de kandidaat. Ze worden gebruikt bij de validering van de toets. De toets, inclusief de extra opgave, verhaaltjes vertellen, maar exclusief de instructie, duurt circa 12 minuten.

Na afronding van de toets kan de uitslag binnen enkele minuten via Internet worden opgevraagd via het unieke Toets Identificatie Nummer van de kandidaat. Bij toepassing van de toets in het kader van het Inburgeringexamen Buitenland zullen de uitslagen via e-mail aan de betreffende Nederlandse ambassades en de ministeries van Justitie en van Buitenlandse Zaken worden gestuurd.

Fraude
De richtlijnen van het Ministerie van Buitenlandse Zaken voorzien in maatregelen die fraude bij de identificatie van kandidaten en de afname van de toetsen helpen voorkomen. Daarnaast is de TGN zelf op een aantal manieren beschermd tegen mogelijke fraude bij de afname en de beoordeling. Tijdens de afname van de TGN krijgt elke kandidaat een unieke verzameling opgaven voorgelegd. De opgaven worden per kandidaat door willekeurige naar itemtype gestratificeerde selectie getrokken uit de totale opgavenbank. Door deze procedure krijgt elke kandidaat in principe een unieke deelverzameling items en is de kans dat twee toetsen meer dan 17% overlappen, erg klein.

De opgavenbank telt op dit moment circa 1.000 opgaven en zal wanneer de toets eenmaal in gebruik is genomen regelmatig worden ververst en aangevuld. Fraude op basis van voorkennis van opgaven is daarmee nagenoeg uitgesloten. Het automatische scoringssysteem maakt fraude bij de beoordeling van het examenwerk onmogelijk.

De opgaven
De TGN bevat drie soorten opgaven.

Herhaalopdrachten
De kandidaat krijgt de opdracht een gesproken zin letterlijk na te zeggen. De herhaalopdrachten vormen een steekproef van uitingen die men in gesproken taal kan tegenkomen. Zij zijn geput uit authentieke audiobronnen, zoals mondelinge interacties en radio-opnamen. Een groot aantal bestaat uit 'formulaic speech'. De stimuli worden op een alledaagse spontane manier uitgesproken zoals men ze ook in het normale spraakgebruik zou kunnen aantreffen. Stimuli variëren in lengte van twee tot maximaal dertien woorden. De zinnen worden aan de kandidaat in toenemende moeilijkheidsgraad aangeboden. De moeilijkheidsgraad komt over het algemeen overeen met de lengte van de zin. Bij korte zinnen kan de kandidaat steunen op het korte termijn geheugen en speelt mogelijk de vaardigheid om uitspraak te kunnen imiteren een rol. Bij langere zinnen is het korte termijn geheugen niet meer toereikend en moet de kandidaat de woorden herkennen en de zinsstructuur doorzien om de stimulus letterlijk te kunnen herhalen.

Voorbeelden zijn: "Daar heb ik nog nooit van gehoord" en "Volgende keer betaal ik".

Korte vragen
De kandidaat krijgt mondeling een korte vraag aangeboden en moet hierop een kort antwoord geven. Dit vereist de vaardigheid met begrip te kunnen luisteren naar een gesproken vraag en om een relevant en verstaanbaar gesproken antwoord te kunnen geven. Kort-antwoord-vragen vragen naar elementaire informatie, of eenvoudige gevolgtrekkingen met betrekking tot tijd, hoeveelheid, lexicale inhoud, of logica. Uitgangspunt is dat de vragen inhoudelijk moeten kunnen worden beantwoord door een twaalfjarige zonder specifieke kennis van Nederland. Met het oog op zeer laaggeschoolde kandidaten in de doelgroep zijn ook vragen die een beroep doen op basale rekenvaardigheid vermeden. Om de vragen te kunnen beantwoorden - vragen waarvan verondersteld wordt dat de kandidaat ze correct zou kunnen beantwoorden wanneer ze in hun eigen taal zouden worden gesteld - moet de kandidaat de woorden in de Nederlandse vraag identificeren, de woorden begrijpen in hun onderlinge betekenisrelatie, de gestelde vraag interpreteren, een antwoord in het Nederlands formuleren en dit op verstaanbare wijze produceren.

Voorbeelden zijn: "Kun je rijst eten of drinken?" en "Jan is ouder dan Piet. Wie is het jongst?".

Tegenstellingen
De kandidaat moet van een gegeven woord het tegengestelde zeggen. De woorden komen voor in de alledaagse spreektaal. Kandidaten moeten het aangeboden woord herkennen en begrijpen (receptieve woordenschat) en het tegengestelde vinden en uitspreken (productieve woordenschat). Vlotte associatie met het tegengestelde van een woord geeft derhalve een indicatie van receptieve en productieve beheersing van vocabulaire en is in alledaagse conversatie van belang.

Voorbeelden zijn: "Niet" en "Ochtend".

Alle opgaven, herhaalopdrachten, korte vragen en tegenstellingen, zijn ingesproken door moedertaalsprekers van het Nederlands. Tijdens een toetsafname krijgen kandidaten opgaven te horen die door verschillende moedertaalsprekers van het Nederlands zijn ingesproken.

De spraak is licht regionaal gekleurd doordat de sprekers - vrouwen en mannen - afkomstig zijn uit verschillende delen van het land.

Het scoringsmodel
Het automatische scoringssysteem levert vier deelscores. Twee daarvan zijn gericht op *wat* de kandidaten letterlijk zeggen (vocabulaire en zinsbouw) en twee zijn gericht op *hoe* de kandidaten spreken (uitspraak en vloeiendheid). De vier scores worden gecombineerd in een totaalscore. De minimale score is tien en betekent dat de kandidaat geen blijk heeft gegeven Nederlands te kunnen verstaan of te kunnen spreken. De hoogste score is tachtig en duidt erop dat de kandidaat in communicatie met Nederlanders geen enkele hindernis door het gebruik van het Nederlands ondervindt noch veroorzaakt voor de gesprekspartner. Hoe meer de score van kandidaten verschilt van de minimale score, hoe hoger hun vaardigheid.

De ontwikkeling van de opgaven
Er is gestart met de ontwikkeling van een groot aantal opgaven. Alle ontwerp-items werden gecontroleerd op vocabulaire met gebruikmaking van het Corpus Gesproken Nederlands. Dit corpus bevat opnamen van telefoongesprekken, spontane gesprekken, interviews en discussies van gesproken Nederlands op basis waarvan kan worden bepaald welke woorden veel voorkomen en daarom relatief belangrijk zijn. Alle opgaven werden in een schriftelijke commentaarronde ter beoordeling voorgelegd aan NT2-deskundigen.
Alle goedgekeurde opgaven werden ingesproken door tien verschillende vrouwelijke en mannelijke sprekers afkomstig uit diverse streken in Nederland. De instructies werden ingesproken door twee professionele stemacteurs, een mannenstem voor de algemene instructies bij het examen en een vrouwenstem voor de aanwijzingen bij de verschillende opgavensoorten. Alle opnamen werden gemaakt in een professionele geluidsstudio.

Ruim 1.300 opgaven werden in pretesten via vaste telefoonlijnen voorgelegd aan twee doelgroepen: 821 moedertaalsprekers (MS) en 1.522 niet-moedertaalsprekers (NMS). De pretesten werden op gelijke wijze samengesteld als de toetsen in de beoogde definitieve vorm: voor iedere kandidaat bevatte de pretest een verschillende selectie uit de opgaven. De pretesten vonden in Nederland plaats op ROC's omdat het niet mogelijk was binnen de beschikbare termijn de data in het buitenland te verzamelen. De cursisten van ROC's werden aangevuld met proefpersonen daarbuiten om spreiding in leeftijd en in vaardigheidsniveau te bereiken.

De gemiddelde leeftijd van de niet-moedertaalsprekers die aan de pretesten deelnamen was 31 jaar en liep uiteen van 8 tot 71. De verdeling man: vrouw was 26:64 (n=1.341). Ongeveer 8% van de NMS-deelnemers was niet-gealfabetiseerd. De NMS deelnemers waren afkomstig uit 121 verschillende landen. Ongeveer de helft verbleef twee jaar of minder in Nederland en één op de vijf had niet meer dan lagere school. Ongeveer 50% van de deelnemers aan de pretesten was twee jaar of minder in Nederland. De gemiddelde leeftijd van de MS deelnemers bedroeg 37 jaar. Deelnemers waren afkomstig uit alle delen van Nederland en van uiteenlopende opleidingsniveaus.

Aan de hand van de reacties van de pretestkandidaten is de kwaliteit van de ontwikkelde opgaven onderzocht en zijn de antwoordmodellen voor de korte vragen en de tegenstellingen vastgesteld. Circa 30% van de opgaven is uit de opgavenbank verwijderd omdat niet aangetoond kon worden dat ze voldeden aan de vooraf gestelde kwaliteitscriteria.

Tijdens de pretesten is speciale aandacht besteed aan de vraag of de gekozen opgavensoorten ook voor kandidaten met een zeer laag opleidingsniveau en ook voor analfabeten voldoende duidelijk waren.

De ontwikkeling van het automatische scoringssysteem
De verzameling van de pretestgegevens had als tweede doel de taalspecifieke componenten voor de spraakherkenning te ontwikkelen. De verzamelde spraakdata omvatten 132.000 uitingen, waaronder circa 59.000 van MS en circa 73.000 van NMS. De data zijn zeer rijk aan variatie: onder de MS zijn representanten van vele varianten naar de verschillende streken in Nederland inclusief dialecten, onder NMS zijn proefpersonen afkomstig uit alle werelddelen en 121 verschillende landen.

De totale dataset is verdeeld in subsets die zijn gebruikt om initiële taalmodellen te bouwen, om deze modellen te trainen en tenslotte om de werking van de modellen te valideren. Het doelwit van spraakherkenning in de gegeven context is de predictie van taalvaardigheid van taalgebruikers in de *perceptie van gebruikers van deze taal*. Uiteindelijk vormen daarom menselijke oordelen de norm waarvoor moet worden geoptimaliseerd.

Voor de onderscheiden maten waarop de deelscores zijn gebaseerd zijn daarom verschillende menselijke oordelen verzameld. Ten behoeve van de ontwikkeling van de correctheidsmaten (*wat* de kandidaten zeggen) zijn de responsen van proefpersonen (zowel MS als NMS) getranscribeerd door getrainde transcribeurs (moedertaalsprekers van het Nederlands). Wat de transcribeurs tonen te kunnen verstaan (doordat zij het opschrijven) moet uiteindelijk ook zijn wat door middel van spraakherkenning wordt verstaan. Overeenkomst tussen de score gebaseerd op de handmatige transcriptie en de score als gegenereerd door machine is het ultieme criterium voor de kwaliteit van het machinale oordeel. Ten behoeve van de ontwikkeling van de kwaliteitsmaten (*hoe* de kandidaten spreken) zijn menselijke oordelen verzameld over de uitspraak en de vloeiendheid van de antwoorden van de kandidaten. Op basis van deze oordelen zijn de machinale scores geschaald.

Om de kwaliteit van de scores van het automatische scoringssysteem te onderzoeken, zijn de gegevens van 139 daarvoor apart gehouden pretestkandidaten nader onderzocht. De reacties van de kandidaten, die dus niet betrokken waren bij de ontwikkeling van de toets en het automatische scoringssysteem, werden twee keer gescoord, één keer door de machine en één keer op basis van de transcripten en oordelen van daartoe getrainde menselijke beoordelaars. De resultaten laten zien dat de scores die tot stand komen op basis van automatische scoring nauw overeenkomen met de toetsscores die gebaseerd zijn op het werk van menselijke beoordelaars (.93).

Aanvullende onderzoeken en representativiteit van de steekproeven
Nadat op basis van de in de pretesten verzamelde gegevens een verzameling opgaven was geselecteerd waarmee toetsen kunnen worden samengesteld die op consistente wijze verschillen tussen personen - op grond van hun reacties op de opgaven – meten, is bepaald wat de relatie is tussen die verschillende toetsresultaten en de oordelen van menselijke beoordelaars over het taalvaardigheidsniveau van kandidaten aan de hand van het CEF. Bovendien is de kwaliteit van de toets nader onderbouwd. Er zijn daartoe drie aanvullende onderzoeken uitgevoerd. De volgende tabel geeft een overzicht van het aantal vertegenwoordigers van de beoogde doelgroep van de TGN die bij de verschillende onderzoeken betrokken waren en van de belangrijkste achtergrondkenmerken van de verschillende steekproeven.

Tabel 1: Overzicht NT2-leerders betrokken bij de ontwikkeling van de TGN

	Aantal	**Landen van herkomst**	**% ppn met taalvaardigheid Nederlands tussen 0 en A2**	**% ppn met een (zeer) laag opleidingsniveau (max. basisschool)**
Pretest	1522	121	Circa 65%	Circa 20%
Aanvullend experiment 1	216	onbekend	Circa 65%	onbekend
Aanvullend experiment 2	353	57	Circa 90%	Circa 35 %
Aanvullend experiment 3	461	82	Circa 90%	Circa 25%

Met name de steekproeven die betrokken waren bij het tweede en het derde aanvullende experiment kunnen wat betreft taalvaardigheidsniveau in het Nederlands en opleidingsniveau beschouwd worden als representatief voor de beoogde doelgroep van de TGN.

De gegevens die gebruikt zijn bij de ontwikkeling van de toets en het automatische scoringssysteem zijn strikt gescheiden van de gegevens die zijn gebruikt om de kwaliteit van de TGN te onderzoeken. De itemselectie en de ontwikkeling van het automatische scoringssysteem zijn geschied aan de hand van de pretestgegevens. De schaling en normering hebben plaatsgevonden aan de hand van de gecombineerde gegevens uit de pretest en het derde aanvullende experiment. De onderzoeken naar de kwaliteit van de uiteindelijke toets zijn gebaseerd op de uitkomsten van het tweede aanvullende experiment en op een daartoe speciaal apart gehouden set van 139 pretestkandidaten. De gegevens uit het eerste aanvullende experiment zijn uitsluitend gebruikt om het effect te onderzoeken van het telefoonnetwerk van het Ministerie van Buitenlandse Zaken.

Schaling en normering
Aangezien de uitslag van de Toets Gesproken Nederlands moet worden gerelateerd aan het CEF, is bij de schaling en normering gebruik gemaakt van de oordelen van menselijke beoordelaars over de mondelinge taalvaardigheid in het Nederlands van de deelnemende proefpersonen. In totaal waren 145 menselijke beoordelaars betrokken bij de ontwikkeling van de TGN. Onder hen waren tien speciaal daarvoor getrainde moedertaalsprekers, die géén specifieke ervaring hadden met het communiceren met leerders van het Nederlands en 42 ervaren NT2 docenten die een speciale beoordelaarstraining hadden gevolgd. De overige beoordelaars waren allen werkzaam als NT2-docent of begeleider, maar hadden geen specifieke training in het gebruik van het CEF als beoordelingsstandaard gevolgd. Alle beoordelaars gaven hun oordelen op basis van de niveaubeschrijvingen van het CEF.

Ten behoeve van de schaling en normering van de TGN – het relateren van de toetsscores aan het CEF – is uitgegaan van de data die verzameld zijn tijdens de pretest. Met het oog op de representativiteit van de steekproef is die dataset uit de pretest aangevuld met gegevens uit het derde aanvullende experiment. Lineaire transformaties leiden tot het volgende overzicht van de plaats van de CEF-ondergrenzen op de TGN-rapportageschaal.

Tabel 2: Relatie TGN-scores en CEF

Toetsscore volgens rapportageschaal TGN	CEF-niveau
80	C2
68 – 79	C1
57 – 67	B2
47 – 56	B1
37 – 46	A2
26 – 36	A1
16 – 25	A1min
10 – 15	Lager dan A1min

Het is aan de opdrachtgever om te bepalen welke eisen er precies aan toekomstige kandidaten zullen worden gesteld en waar de grenzen tussen 'zakken' en 'slagen' dienen te liggen. In de Verantwoording leveren de toetsontwikkelaars de informatie die daarvoor nodig is. Bijlage 13 bij de verantwoording geeft een overzicht van mogelijke beslissingen en de effecten daarvan.

De kwaliteit van de TGN

De kwaliteit van de opgaven en de mate waarin deze zijn afgestemd op de vaardigheid van de kandidaten bepalen tezamen met het aantal items de hoeveelheid informatie die een toets over de doelgroep kan opleveren. Het is wenselijk om de hoeveelheid informatie te maximaliseren op die punten van de scoreschaal waar beslissingen over kandidaten relevant zijn. Hoe groter immers de informatie op een bepaald punt op de scoreschaal, hoe kleiner de meetfout op dat punt. Iedere toets en ieder examen heeft een meetfout. De betrouwbaarheid van een toets geeft de verhouding aan tussen de spreiding van de scores (hoeveel verschil er tussen kandidaten wordt gemeten) en de fout die er bij het meten van die verschillen wordt gemaakt. Aangezien de informatie niet even groot is over de gehele lengte van de scoreschaal, zijn ook de daarvan afhankelijke standaard meetfout en de betrouwbaarheid niet over de hele schaal gelijk.

De *toetsbetrouwbaarheid* van de TGN in termen van homogeniteit (de mate waarin alle opgaven bijdragen aan dezelfde meting) bedraagt 0.94. De TGN doet wat dat betreft niet onder voor gezaghebbende taaltoetsen zoals de Test of Spoken English (TSE) van Educational Testing Service (ETS) en de onderdelen spreken en luisteren van het Staatsexamen Nederlands als Tweede Taal. De meetfout bij de onderscheiden grensscores varieert van iets minder dan drie punten op de scoreschaal bij A1-min tot iets meer dan drie punten bij A2 en loopt op tot bijna vier punten bij het hoogste niveau C2. Op de lagere niveaus is het betrouwbaarheidsinterval dus kleiner dan op de hogere niveaus. Dit past bij het door de opdrachtgever beoogde gebruik van de TGN die in de eerste plaats bedoeld is voor het toetsen van beheersing van de niveaus A1-min en A2.

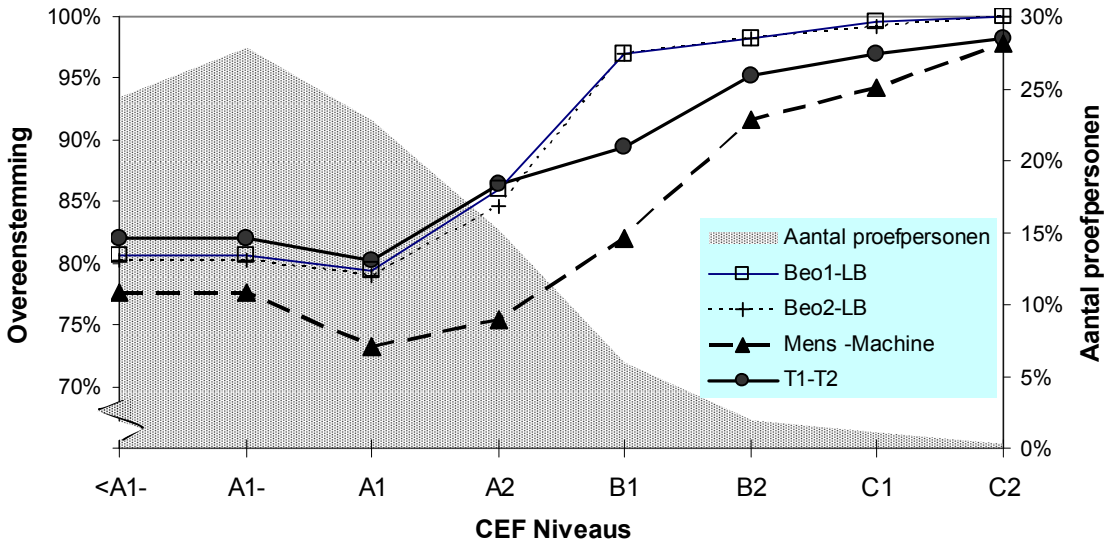
De *validiteit* van de TGN is onderzocht door de TGNscores te vergelijken met de oordelen van menselijke beoordelaars. Daarbij is steeds gebruik gemaakt van data die onafhankelijk zijn van de gegevens die gebruikt zijn bij de constructie van de toets. De uitkomsten van de analyses laten steeds opnieuw hoge correlaties zien tussen de TGNscores en diverse menselijke oordelen: 'naïeve' oordelen van ongetrainde docenten gebaseerd op hun globale indruk van de kandidaten, oordelen van getrainde docenten op basis van interviews met de kandidaten, oordelen van getrainde beoordelaars op de opdracht Verhalen navertellen en oordelen van getrainde beoordelaars op antwoorden van kandidaten op open vragen. De gevonden correlaties zijn steeds positief. Bovendien nemen zij overeenkomstig de verwachting toe met de mate van training van de beoordelaars en de mate waarin overeenkomst bestaat tussen het materiaal waarop zij hun oordelen baseren en de toetsopgaven van de TGN.

Verder blijkt dat moedertaalsprekers van het Nederlands - ongeacht hun opleidingsniveau, leeftijd of gebruik van dialect - topscores behalen, terwijl leerders van het Nederlands als Tweede Taal zeer gespreide scores behalen. Daarbij is er wel een samenhang gevonden tussen de TGNscores en de duur van verblijf in Nederland, maar geen of een zeer beperkte samenhang met land van herkomst, leeftijd, geslacht, en mate van alfabetisering. Ook deze gegevens ondersteunen de conclusie dat de TGN een valide instrument is om de mondelinge vaardigheden van leerders van het Nederlands te meten.

De kwaliteit van de TGN is ten slotte nog nader onderzocht door vergelijkingen te maken tussen zak-slaagbeslissingen die gebaseerd zijn op de TGNscores en zak-slaagbeslissingen die genomen zouden worden op basis van de oordelen van getrainde menselijke beoordelaars. Opnieuw is gebruik gemaakt van data die onafhankelijk zijn van de gegevens met behulp waarvan de TGN ontwikkeld is. Figuur 1 vat de uitkomsten van de verschillende analyses samen. Op de linker y-as is de mate van overeenstemming weergegeven tussen de zak-slaagbeslissingen van twee verschillende beoordelaars/instrumenten. Het gaat om de volgende vergelijkingen:

- BEO1-LB: het oordeel over het beheersingsniveau van een kandidaat door een getrainde interviewer tijdens een gestructureerd interview met de kandidaat versus het oordeel van een getrainde docent of begeleider tijdens een loopbaangesprek met dezelfde kandidaat;
- BEO2-LB: het oordeel over het beheersingsniveau van een kandidaat van een getrainde docent die als tweede beoordelaar aanwezig is bij een gestructureerd interview met de kandidaat versus het oordeel van een getrainde docent of begeleider tijdens een loopbaangesprek met dezelfde kandidaat;
- Mens-Machine: het oordeel van een getrainde menselijke beoordelaar over het beheersingsniveau van een kandidaat versus de automatisch gegeneerde score op de TGN van diezelfde kandidaat;
- T1-T2: de vergelijking tussen de uitkomsten van twee afnames van de TGN bij dezelfde kandidaat.

Op de rechter y-as is uitgezet waar de kandidaten zich bevonden volgens het *gemiddelde* oordeel van zowel de menselijke beoordelaars als de TGN bevonden: iets minder dan een kwart werd beoordeeld als “onder A1-min”, iets meer dan een kwart op niveau A1-min, weer iets minder dan een kwart op A1 en ongeveer 15% op A2. Het gearceerde deel van de figuur geeft weer dat de steekproef waarop de vergelijkingen zijn gebaseerd wat betreft hun geschatte taalvaardigheidsniveau in het Nederlands representatief is voor de doelgroep waarvoor de TGN is ontwikkeld: kandidaten met een beheersingsniveau rondom A2 of lager.



Figuur 1: Beoordelingen experiment Amsterdam

De lijnen in de grafiek geven de mate van overeenstemming tussen de zak-slaagbeslissingen van de verschillende beoordelaars weer. In het voor de TGN meest relevante gebied (van onder A1-min tot en met A2) zijn de overeenstemmingpercentages tussen twee machinale beoordelingen (T1-T2) steeds hoger dan die van de menselijke oordelen onderling. De overeenstemming tussen mens en machine is vanzelfsprekend het laagst, deze kan namelijk niet hoger zijn dan de laagste overeenstemming tussen mensen onderling.

Conclusies

Diverse schattingen van de betrouwbaarheid indiceren dat met de TGN voldoende betrouwbaar kan worden gemeten: de schattingen overtreffen ruimschoots de in de opdracht gestelde minimale streefwaarde van .80 en laten zich goed vergelijken met de betrouwbaarheidscoëfficiënten van toetsen die nationaal en internationaal als betrouwbaar worden beschouwd. Het automatische spraakherkenning- en scoringssysteem functioneert voldoende precies om oordelen te kunnen genereren die vergelijkbaar zijn met die van – goed getrainde – mensen. In de verzamelde gegevens is een samenhang gevonden tussen de toetsscores en relevante maten voor de beheersing van gesproken Nederlands in interactie met sprekers van het Nederlands. Er is in de verzamelde data géén aanwijzing gevonden dat de scores op de TGN worden beïnvloed door eigenschappen en kenmerken van kandidaten waarvan verondersteld kan worden dat ze niet samenhangen met taalvaardigheid.

1 Achtergronden

1.1 De vraag

In het Hoofdlijnenakkoord voor het kabinet CDA, VVD, D66 van 16 mei 2003 wordt gesteld dat wie zich duurzaam wil vestigen in Nederland actief aan de samenleving moet deelnemen en zich de Nederlandse taal eigen moet maken, zich bewust moet zijn van de Nederlandse waarden en de Nederlandse normen moet naleven. Iedere nieuwkomer die op vrijwillige basis naar Nederland komt en valt onder de doelgroepen van de Wet Inburgering Nieuwkomers, moet eerst in eigen land Nederlands op basisniveau leren en enige kennis over de Nederlandse samenleving opdoen als voorwaarde voor toelating. Eenmaal in Nederland aangekomen, moet hij of zij zich dan nog verder verdiepen in de Nederlandse taal en maatschappij. Nader af te bakenen groepen oudkomers, maar in ieder geval zij die onvoldoende Nederlands beheersen en afhankelijk zijn van een uitkering, moeten alsnog een inburgeringsexamen halen. Asielzoekers krijgen pas een definitieve verblijfsstatus na het halen van het examen.

Om aan het gestelde in het Hoofdlijnenakkoord te kunnen voldoen, deed het Ministerie van Justitie in september 2003 een offerteaanvraag voor de toetsontwikkeling uitgaan. Uit deze offerteaanvraag, de voorlichtingsbijeenkomst op 6 oktober 2003 en het verslag daarvan in een memorandum bleek dat de volgende instrumenten werden gevraagd:

A gestandaardiseerde toetsen waarmee het niveau van mondelinge en schriftelijke taalvaardigheid Nederlands van mensen die beogen zich in Nederland te vestigen, in het land van herkomst kan worden vastgesteld;

B gestandaardiseerde toetsen waarmee het niveau van mondelinge en schriftelijke taalvaardigheid Nederlands van nieuwkomers en oudkomers die reeds in Nederland verblijven, kan worden vastgesteld;

C praktijktoetsen waarmee het niveau van mondelinge taalvaardigheid Nederlands van nieuwkomers en oudkomers die reeds in Nederland verblijven, kan worden vastgesteld.

De onderdelen A, B, en C dienden volgens de aanwijzingen in de offerteaanvraag een sterk onderling verband te tonen, met dien verstande dat resultaten op één instrument moesten kunnen worden gerelateerd aan de resultaten op elk van de andere instrumenten. Bij voorkeur dienden de instrumenten genoemd onder A en B gelijk te zijn. De toetsen voor de mondelinge vaardigheden zouden óók gebruikt moeten kunnen worden voor personen die ongeletterd zijn in het Nederlands en in hun moedertaal. Bij de onder C genoemde instrumenten ging het om praktijktoetsen gericht op de functioneringsdomeinen Opvoeding en Arbeidsmarkt. De praktijktoetsen dienden centraal te worden ontwikkeld en moesten kunnen worden afgenomen door onderwijsinstellingen in Nederland. Voor alle delen van de offerte gold dat de uitslagen van de te ontwikkelen toetsinstrumenten gerelateerd moesten kunnen worden aan de schalen van het Gemeenschappelijk Europees Raamwerk voor talen (in het navolgende aangeduid met het Engelse acroniem CEF, Common European Framework). Welke beheersingsniveaus getoetst zouden moeten worden, was nog niet bekend. Daarover waren nog geen besluiten genomen.

1.2 De offerte

Op 17 oktober 2003 diende CINOP, dat daartoe een samenwerkingsverband was aangegaan met het Amerikaanse bedrijf Ordinate en met Language Testing Services in Velp, haar offerte in. Daarin wordt een examenstelsel voorgesteld volgens het model in Figuur 1.1.

<table>
<tr><td colspan="2"></td><td colspan="3">Inburgeringsexamen</td></tr>
<tr><td colspan="2"></td><td colspan="2">Nederland</td><td>Land van Herkomst</td></tr>
<tr><td colspan="2"></td><td>decentraal</td><td>centraal</td><td>centraal</td></tr>
<tr><td rowspan="2">Examen-ondedelen</td><td>Mondeling</td><td rowspan="2">Taalportfolio met (onder meer) Centraal Ontwikkelde Praktijktoetsen</td><td>Mondelinge Interactie</td><td>Mondelinge Interactie</td></tr>
<tr><td>Schriftelijk</td><td></td><td>Geletterdheid</td></tr>
</table>

Figuur 1.1 Model examenstelsel voor inburgering in buitenland en in binnenland volgens offerte CINOP (oktober 2003)

1.2.1 Voorstel voor het Inburgeringsexamen in Nederland

Uitgangspunt van de offerte van CINOP is het inburgeringsexamen in Nederland. Naar analogie van bijvoorbeeld de examens in het voortgezet onderwijs stelt CINOP ten behoeve van inburgeraars die reeds in Nederlands verblijven een examen voor dat bestaat uit een centraal deel en een decentraal deel. In het *decentrale deel* worden de geëiste mondelinge en schriftelijke taalvaardigheden getoetst door middel van portfolio-assessment. Dit deel van het inburgeringsexamen is - volgens het gepresenteerde model - gebaseerd op het Europees Taalportfolio voor volwassen tweede-taalleerders (Kerkhoff, A., 2002). Dit portfolio is in 2002 in opdracht van het Ministerie van OCW in nauwe samenwerking tussen CINOP, Citogroep, Bureau ICE en de Bve Raad ontwikkeld en in 2003 geaccrediteerd door de Raad van Europa (accrediteringsnummer 36.2003). Volgens de uitgangspunten van de Europese portfoliomethodiek is de eigenaar van een portfolio zelf verantwoordelijk voor de inhoud van zijn of haar portfolio en vrij in de keuze van de documenten/bewijzen die hij daarin wil bewaren om zijn taalvaardigheidniveau aan te tonen. In het door CINOP voorgestelde examenmodel zouden kandidaten verplicht kunnen worden in hun portfolio een minimum aantal bewijzen van met succes afgelegde (centraal ontwikkelde) praktijktoetsen op te nemen. Daardoor zou de inhoud van de portfolio's van individuele inburgeraars enigszins kunnen worden gestandaardiseerd. Kandidaten die géén behoefte hebben aan een persoonlijk dossier zouden de gelegenheid kunnen krijgen om een portfolio samen te stellen dat uitsluitend bestaat uit bewijzen van met succes afgelegde centraal ontwikkelde praktijktoetsen.

Het voorstel om het Europees Taalportfolio te kiezen als uitgangspunt voor het inburgeringsexamen in Nederland sluit aan bij de doelen van het regeringsbeleid inzake inburgering waarin 'eigen verantwoordelijkheid' en 'meedoen' centraal staan. De inzet van een portfolio, zoals voorgesteld in de offerte, zou er bovendien toe kunnen bijdragen dat inburgering een 'tweezijdig' proces wordt, doordat mensen uit de omgeving van inburgeraars een wezenlijke rol kunnen spelen bij de integratie en het verzamelen van bewijzen daarvan. De portfoliomethodiek past bij actuele ontwikkelingen in het onderwijs en op de arbeidsmarkt waarin noties als 'duaal leren' en 'competentiegericht' opleiden en beoordelen centraal staan. Een 'inburgeringportfolio' sluit naadloos aan bij de initiatieven van de overheid ter stimulering van Erkenning van Verworven Competenties.

Vanzelfsprekend stelt het beoogde civiel effect van het inburgeringsexamen hoge eisen aan de kwaliteit van het examen. Om de kwaliteit van het 'inburgeringportfolio' te versterken, worden in de offerte van CINOP een aantal instrumenten voorgesteld:

- een centraal door de Minister vast te stellen examenreglement;
- een accrediteringsstelsel voor portfolio-assessoren en;
- een centraal examen.

In de offerte van CINOP wordt voorgesteld voor het *centrale deel* een nieuwe toets te ontwikkelen die de mondelinge gespreksvaardigheid (luisteren en spreken) van de eigenaar van het portfolio meet.

1.2.2 Voorstel voor het Inburgeringsexamen Buitenland

Voor wat betreft het Inburgeringsexamen Buitenland stelt CINOP in haar offerte niet voor om uit te gaan van de Europese portfoliomethodiek. Behalve valide en betrouwbaar dienen examens immers ook uitvoerbaar te zijn. Een in de context van het Inburgeringsexamen Buitenland wereldwijd te organiseren examen op basis van een Europees taalportfolio leek pertinent onhaalbaar, zowel voor de doelgroep zelf als voor de opdrachtgever. In haar offerte stelt CINOP daarom voor de door de opdrachtgever te bepalen eisen aan de mondelinge en de schriftelijke vaardigheden van kandidaten in het buitenland te toetsen door middel van twee *centraal* te ontwikkelen onderdelen: één toets voor de mondelinge vaardigheden die geëist gaan worden en één toets voor de schriftelijke vaardigheden. Conform het gestelde in het offerteverzoek zou de toets voor de mondelinge vaardigheden ook geschikt moeten zijn voor inzet in het inburgeringsexamen in Nederland.

Het beoogde gebruik van de centraal te ontwikkelen toets voor mondelinge interactie in het buitenland brengt een aantal specifieke randvoorwaarden met zich mee waaraan met traditionele vormen van toetsing moeilijk of niet kan worden voldaan. De toetsafname moet wereldwijd en dagelijks kunnen worden gerealiseerd, onder toezicht van toetsleiders zonder specifieke expertise op het gebied van taaltoetsing. Daarnaast moet de toets efficiënt zijn wat betreft logistiek en doorlooptijd van het gehele proces vanaf inschrijving tot en met toetsuitslag. Het spreekt voor zich dat de toets in hoge mate fraudebestendig moet zijn. De genoemde randvoorwaarden hebben een belangrijke rol gespeeld bij de keuze voor het model van een bestaand geautomatiseerd toetssysteem ontwikkeld door Ordinate: de PhonePass technologie. Er waren ook meer inhoudelijke argumenten voor de keuze van het toetssysteem van Ordinate. De gangbare mondelinge taaltoetsen zijn sterk gericht op het zo direct mogelijk toetsen van functionele mondelinge taalvaardigheden in reële taalgebruiksituaties. Portfolio-assessment volgens het Europese model is bij uitstek een voorbeeld van een toetsmodel dat gebaseerd is op dat, door de socio-linguïstiek geïnspireerde, paradigma: kandidaten bewijzen hun gespreksvaardigheden in concrete en reële taalgebruiksituaties, waarin zij met voor hen relevante gesprekspartners communiceren over voor hen relevante onderwerpen. Het toetssysteem van Ordinate is daarentegen gebaseerd op onderzoek op het gebied van spraakherkenning, statistische modellen, linguïstiek en toetstheorieën. Met deze kennis is een systeem voor automatische scoring ontwikkeld specifiek voor het beoordelen van de spraak van taalleerders. De bijbehorende opgaven zijn vooral gericht op het verzamelen van informatie over de processen die zich in het hoofd van de taalgebruiker afspelen en doen nauwelijks of geen beroep op de meer socio-linguïstische competenties van de taalleerder. Beide paradigma's hebben hun eigen sterke kanten en hun eigen beperkingen als het gaat om het betrouwbaar en valide beoordelen van taalvaardigheid. Voor het examen in Nederland heeft de combinatie van de twee paradigma's duidelijk meerwaarde. Voor kandidaten die zich buiten Nederland op het examen moeten voorbereiden, en die daar doorgaans nauwelijks of geen gelegenheid hebben gehad om kennis te maken met de meer sociale aspecten van het Nederlands, zou een examen waarin ook meer socio-linguïstische competenties worden getoetst al gauw leiden tot culturele bias, onterechte moeilijkheden. Een valide en betrouwbaar instrument op basis van het toetssysteem van Ordinate heeft ook wat dat betreft voordelen.

1.3 De opdracht

Medio december 2003 kreeg CINOP de opdracht om samen met Ordinate en LTS de in haar offerte voorgestelde centrale toetsen op basis van PhonePass technologie te ontwikkelen. Onderdeel C werd toegekend aan een consortium van Bureau ICE, Citogroep en ITTA. Op dat moment had de opdrachtgever nog geen besluit genomen over de precieze vorm van het inburgeringsexamen in Nederland en over de rol van het Europees Taalportfolio daarbinnen. Wel werd contractueel vastgelegd dat de toetsen op basis van de PhonePass technologie ontwikkeld werden in het kader van het streven van de overheid naar een examenstelsel in Nederland op basis van de Europese portfoliomethodiek. Wat betreft het niveau van de te ontwikkelen toetsen werd vastgelegd dat de toets voor de mondelinge vaardigheden een bereik zou moeten hebben van A1 tot en met B2. Verder werd met betrekking tot de betrouwbaarheid van de te ontwikkelen instrumenten contractueel een split-half betrouwbaarheid van 0.80 als ondergrens gesteld.

CINOP is meteen na de gunning van de opdracht in december 2003 gestart met de ontwikkeling van de toetsen. In februari 2004 publiceerde de Adviescommissie Normering Inburgeringseisen, de Commissie Franssen, haar advies 'Inburgering getoetst'. De commissie adviseerde daarin niveau 'A1-min' als norm voor het inburgeringsexamen buitenland. Op dat niveau worden volgens de omschrijving van de Commissie geen schriftelijke vaardigheden verondersteld (in hoofdstuk 2 wordt uitgebreider ingegaan op de omschrijving van dit niveau). Medio mei 2004 werd de ontwikkeling van de toetsen ten behoeve van de meting van de schriftelijke vaardigheden op verzoek van de opdrachtgever stopgezet. De tot op dat moment ontwikkelde opgaven en de gegevens die verzameld waren in de pretests, zijn zonder nadere bewerking of analyse opgeslagen en blijven beschikbaar voor de opdrachtgever. Zij vormen echter geen onderdeel van dit rapport.

Deze rapportage is beperkt tot de verantwoording van de Toets Gesproken Nederlands, hierna 'TGN', die CINOP, Ordinate en LTS gezamenlijk hebben ontwikkeld.

1.4 Projectactiviteiten

Eind december 2003 is gestart met de uitvoering van het project. We geven hier een beknopt overzicht van de productiefasen van de TGN.

Voorbereiding
In de periode december 2003 - januari 2004 hebben CINOP, Ordinate en LTS de specificatie van toets en items, zoals beschreven in de offerte, uitgewerkt en nadere afspraken gemaakt omtrent de uitvoering van het project. Tegelijkertijd werd gestart met het zoeken naar projectmedewerkers (itemschrijvers, itembeoordelaars, stemacteurs voor het inspreken van opgaven, beoordelaars, transcribeurs), pretestkandidaten en hulpmiddelen.

Ontwikkelen pretestopgaven
In februari 2004 is een set van in totaal 2.131 opgaven ontwikkeld, opgenomen volgens de overeengekomen specificaties en in Ordinate's toetssysteem geïnstalleerd.

Pretest
In de periode eind maart - begin mei 2004 namen in totaal 836 moedertaalsprekers en 1.522 leerders van het Nederlands als tweede taal deel aan de pretesten. Van alle kandidaten werden relevante achtergrondgegevens gevraagd.

Transcriberen
Parallel aan de dataverzameling werd gestart met de transcriptie van de tijdens de pretest verzamelde reacties van de kandidaten en moedertaalsprekers. Transcribeurs waren speciaal voor dit doel getrainde moedertaalsprekers van het Nederlands.

Menselijke beoordeling
Een groep getrainde beoordelaars - allemaal moedertaalsprekers van het Nederlands - beoordeelde in de zomer van 2004 de reacties van de pretestkandidaten aan de hand van daartoe ontwikkelde beoordelingsschalen, gerelateerd aan het CEF.

Ontwikkelen automatische beoordeling
Op basis van de verzamelde pretestdata en met gebruikmaking van Ordinate's bestaande technologie werd een systeem ontwikkeld ten behoeve van de automatische analyse, beoordeling en scoring van de reacties van kandidaten op de Nederlandse toets.

Eerste rapportage
In de periode juni 2004 - november 2004 werden de verzamelde gegevens geanalyseerd. In november en in december 2004 werden eerste versies van een rapport over de ontwikkeling van de toets besproken met de resonansgroep bestaande uit deskundigen op het gebied van spraakherkenning en toetsing. De commentaren van de leden van de resonansgroep, alsmede reacties van derden, waren de directe aanleiding om aanvullende projecten uit te voeren om de kwaliteit van de ontwikkelde toets nader te onderzoeken en te onderbouwen.

Experiment telefoonlijnen
In het najaar van 2004 werd duidelijk dat veel Nederlandse ambassades en posten in het buitenland niet kunnen beschikken over het algemene publieke telefoonnet (Public Switched Telephone Network, hierna PSTN). In plaats daarvan maken zij gebruik van een speciaal en beveiligd netwerk van het Ministerie van Buitenlandse Zaken (hierna: MFA-net). In december 2004 werd een experiment uitgevoerd waarin de invloed van het gebruik van het netwerk van Buitenlandse zaken werd onderzocht. Hierover waren namelijk geen gegevens beschikbaar omdat de pretest via PSTN werd afgenomen. De verzamelde gegevens toonden aan dat het gebruik van het MFA-net significante invloed had op de scores van kandidaten. Omwille van de leesbaarheid wordt dit experiment hierna aangeduid als 'experiment telefoonlijnen'.

Experiment Amsterdam
Naar aanleiding van reacties vanuit de resonansgroep is in januari 2005 gestart met de voorbereiding van een nieuw experiment waarin werd nagegaan of de uitkomsten van de pretesten met betrekking tot de betrouwbaarheid en validiteit van het ontwikkelde instrument óók generaliseerbaar zijn naar de specifieke doelgroep van het Inburgeringsexamen in het buitenland: kandidaten met een zéér laag taalvaardigheidniveau in het Nederlands die nauwelijks of geen onderwijs Nederlands hebben genoten. Een tweede doelstelling van het experiment was het verzamelen van aanvullende gegevens ter onderbouwing van de cesuur A1-min. Omwille van de leesbaarheid van wat volgt, wordt dit experiment hierna aangeduid als 'experiment Amsterdam'.

MFA-Fit
Vragen uit de resonansgroep waren de aanleiding om in februari 2005 te starten met de voorbereiding van een experiment dat aanvullende informatie moest opleveren met betrekking tot de kwaliteit van de verzamelde data, de validering van de toets en de wijze waarop gecorrigeerd kan worden voor het effect van het gebruik van MFA-net zoals geconstateerd in het 'experiment telefoonlijnen'. Omwille van de leesbaarheid wordt dit experiment hierna aangeduid als 'experiment MFA-Fit'.

Eindrapportage
In de periode mei/juni 2005 werd een tweede versie van het rapport over de ontwikkeling van het Inburgeringsexamen Nederlands besproken met de opdrachtgever en met de leden van de resonansgroep. Op basis van de verzamelde reacties is de verantwoording van de toets in de zomer van 2005 definitief gemaakt.

Gedurende de totale looptijd van het project heeft minimaal een keer per maand voortgangsoverleg plaatsgevonden met de opdrachtgever.

1.5 Opzet rapport

Omwille van de leesbaarheid worden in dit eindrapport gegevens die in verschillende fases van het project zijn verzameld, soms naast elkaar gepresenteerd.

Hoofdstuk 2 geeft een beschrijving van de TGN en behandelt achtereenvolgens het construct, de opgavenontwikkeling en de techniek van de automatische beoordeling.

Hoofdstuk 3 beschrijft de opzet, de uitvoering en de resultaten van het pretesten van de ontwikkelde opgaven op moedertaalsprekers en niet-moedertaalsprekers van het Nederlands. De resultaten gaven aanleiding tot enkele aanvullende vraagstellingen ten aanzien van de technologische randvoorwaarden en psychometrische aspecten.

Hoofdstuk 4 beschrijft de opzet en eerste resultaten van de drie aanvullende experimenten: het experiment telefoonlijnen, het experiment Amsterdam en het experiment MFA-Fit. De resultaten van de verschillende experimenten worden in de volgende hoofdstukken beschreven.

Hoofdstuk 5 behandelt de schaling en normering van de TGN. Hierbij worden zowel gegevens uit de pretest als uit de aanvullende onderzoeken gebruikt.

Hoofdstuk 6 geeft een overzicht van alle gegevens die bij de pretest en bij de aanvullende onderzoeken zijn verzameld ten aanzien van de betrouwbaarheid en de validiteit van de TGN. Hierbij komen de interne samenhang van de TGN, de samenhang met externe maten voor mondelinge vaardigheden in het Nederlands en de relatie tussen de toetsscores en achtergrondvariabelen van de kandidaten aan de orde.

2 Beschrijving van de Toets Gesproken Nederlands

2.1 Doelstelling

De Minister voor Vreemdelingenzaken en Integratie heeft de Toets Gesproken Nederlands (TGN) laten ontwikkelen om de luister- en spreekvaardigheid te toetsen van personen die in aanmerking willen komen voor een Machtiging tot Voorlopig Verblijf in Nederland. De Toets Gesproken Nederlands is ook geschikt om te worden ingezet in het inburgeringsexamen in Nederland.

Op het moment waarop dit rapport wordt vastgesteld, heeft de wetgever nog geen definitieve besluiten genomen over de eisen die aan de onderscheiden doelgroepen gesteld zullen worden. Bij de samenstelling van het rapport is ervan uitgegaan dat het parlement zou instemmen met de volgende exameneisen:

1. Van oudkomers en nieuwkomers in Nederland zal beheersingsniveau A2 van het CEF vereist worden voor de mondelinge vaardigheden. Voor wat betreft de schriftelijke vaardigheden zal van nieuwkomers een beheersingsniveau van A2 geëist worden, van oudkomers A1.
2. Personen die in aanmerking willen komen voor een Machtiging tot Voorlopig Verblijf zullen op het A1-min niveau moeten kunnen functioneren wat mondelinge vaardigheden betreft. 'A1-min' is de aanduiding van een niveau dat zich bevindt ónder het laagste niveau dat in het CEF wordt beschreven.

In de context van het Inburgeringsexamen Buitenland is de Toets Gesproken Nederlands het enige instrument met behulp waarvan de taalvaardigheid van de kandidaten wordt gemeten. Het tweede onderdeel van het Inburgeringsexamen Buitenland meet de Kennis van de Nederlandse Samenleving van de kandidaten.
De precieze vorm en inhoud van het Inburgeringsexamen in Nederland is op het moment van schrijven van dit rapport nog niet bekend. Het examen zal bestaan uit twee onderdelen: mondelinge taalvaardigheid Nederlands en Kennis van de Nederlandse Samenleving. Bij de constructie van de TGN is aangenomen dat de TGN in het examenonderdeel Nederlands gecombineerd zal worden met praktijktoetsen in een examenstelsel dat gebaseerd is op de Europese portfoliomethodiek (zie hoofdstuk 1).

2.2 Toets Format en afnamecondities

De TGN kan in principe met behulp van elke telefoon met een vaste verbinding worden afgelegd. De toetsconstructeurs verzorgen 'back office' het beheer van de toetsen en het onderhoud van de opgavenbank. De beoordeling en scoring verlopen door middel van een automatisch scoringsproces.
Om de toets te maken, telefoneert een kandidaat naar het toetssysteem van Ordinate. Nadat de examenleider heeft gezorgd voor verbinding met het toetssysteem van Ordinate en het intoetsen van het persoonlijke Toets Identificatie Nummer van de kandidaat, test het toetssysteem de kwaliteit van het stemvolume en de verbinding. Zonodig volgen er aanwijzingen, bijvoorbeeld om harder te spreken of om de microfoon beter voor de mond te plaatsen. Die aanwijzingen zijn geformuleerd in het Nederlands. De examenleider is aanwezig om eventuele hulp te bieden.
Wanneer de technische aspecten van het examen goed zijn bevonden, begint de eigenlijke toets die ongeveer 12 minuten duurt. Tijdens het examen wordt geen gebruik gemaakt van schriftelijke materialen of opgaven.

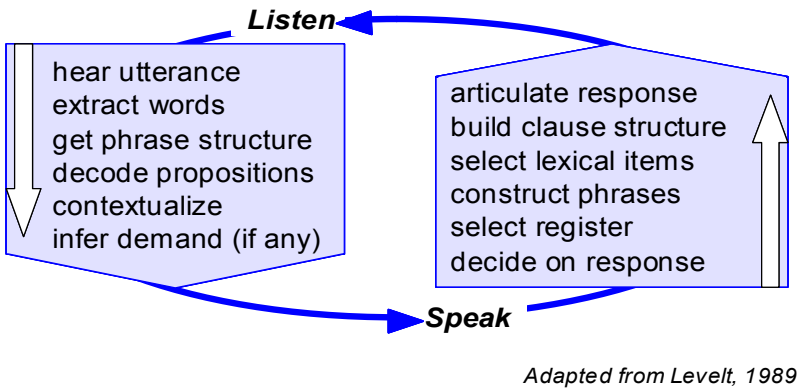
Het toetssysteem presenteert de kandidaat via de telefoon een aantal opgaven in het Nederlands waarop de kandidaat in het Nederlands moet reageren. De toets bestaat uit 4 onderdelen: 'Zinnen Herhalen', 'Korte Vragen', opnieuw een onderdeel 'Zinnen Herhalen en 'Tegenstellingen'. Elk onderdeel begint met een korte instructie in het Nederlands en twee voorbeelden. De instructies bij de opgaven zijn kort en worden duidelijk in het Nederlands gesproken door twee verschillende stemmen. Na de instructie en de voorbeelden volgen de opgaven. De vier onderdelen van de toets worden automatisch beoordeeld door het scoringssysteem van Ordinate. Na het laatste onderdeel volgen twee opdrachten 'Verhalen navertellen. Deze opdrachten worden niet automatisch beoordeeld en spelen geen rol bij het bepalen van de score van een kandidaat. Ze zijn bestemd voor de validering van de toets.

Kandidaten krijgen voorafgaand aan de toets mondelinge instructies van een daartoe getrainde examenleider. Bij de instructie van kandidaten die kunnen lezen wordt daarbij gebruik gemaakt van een instructievel (zie Bijlage 1). Van dit instructievel zijn vertalingen beschikbaar in het Arabisch, Frans, Indonesisch, Portugees, Spaans, Turks, Chinees, Engels, Russisch en Thai. Kandidaten voor het Inburgeringsexamen in Nederland krijgen de mondelinge instructies in het Nederlands aan de hand van het Nederlandse instructievel. Bij de mondelinge instructie van de kandidaten in het buitenland wordt gebruik gemaakt van de moedertaal van de kandidaten of een andere taal die zij voldoende beheersen om de instructies te volgen. In gevallen waarin er géén examenleiders beschikbaar zijn die een taal beheersen waarin zinvol met een kandidaat gecommuniceerd kan worden, krijgt de kandidaat de gelegenheid zelf iemand mee te brengen die als tussenpersoon kan functioneren. Deze tussenpersoon verlaat - nadat de kandidaat de instructies heeft begrepen - de ruimte waarin het examen wordt afgenomen. Bijlage 2 bevat het voorlopige draaiboek voor de afname van het Inburgeringsexamen Buitenland.

Bij de voorbereiding op de TGN kunnen kandidaten zich vertrouwd maken met de itemtypes en de instructies door middel van oefentoetsen. Ten behoeve van de doelgroep van het examen zijn drie oefentoetsen beschikbaar. Als algemeen uitgangspunt wordt gehanteerd dat elke kandidaat minimaal één keer oefent voordat hij aan de toets deelneemt. De oefentoetsen zijn wat betreft opzet en instructies identiek aan het echte examen. Ze zijn echter korter dan de toets die de kandidaat op de ambassade zal maken, doordat ze van elk onderdeel minder opgaven bevatten. Oefentoetsen zijn verkrijgbaar via boekhandels in Nederland en via Internet. Kandidaten kunnen de oefentoetsen vanaf elke plek ter wereld met behulp van een vaste telefoonverbinding afleggen. Scores kunnen met een persoonlijke code via Internet worden opgezocht. Via de website www.naarnederland.nl kunnen (aspirant-)kandidaten en andere belangstellenden in de toekomst nadere informatie vinden over de wetgeving en de examens. Kandidaten zullen daar ook suggesties vinden over de wijze waarop zij zich op de examens kunnen voorbereiden.

2.3 Toetsconstruct

De Toets Gesproken Nederlands is gebaseerd op psycholinguïstische modellen van taalgedrag en beoogt het gemak te meten waarmee een kandidaat in staat is Nederlands dat wordt gesproken in een normaal conversatietempo te verstaan, het gesprokene te begrijpen en te interpreteren en er zinvol in verstaanbaar Nederlands op te reageren. Succesvolle deelname aan een gesprek berust op de beheersing en uitvoering van een aantal deelprocessen. Het gaat daarbij om het volgen van hetgeen wordt gezegd, het herkennen van de lexicale elementen, het reconstrueren van de syntactische structuur, het decoderen van de boodschap en deduceren van de noodzaak of de gewenstheid van een reactie. Dit luisterproces vindt plaats gedurende de voortgang van het gesprek. Heeft men eenmaal een besluit tot een reactie genomen, dan moet men dit formuleren in een betekenisstructuur, de benodigde lexicale elementen ophalen, de bijbehorende zinstructuur bouwen en het geheel verklanken. Figuur 2.1 brengt dit proces in beeld.

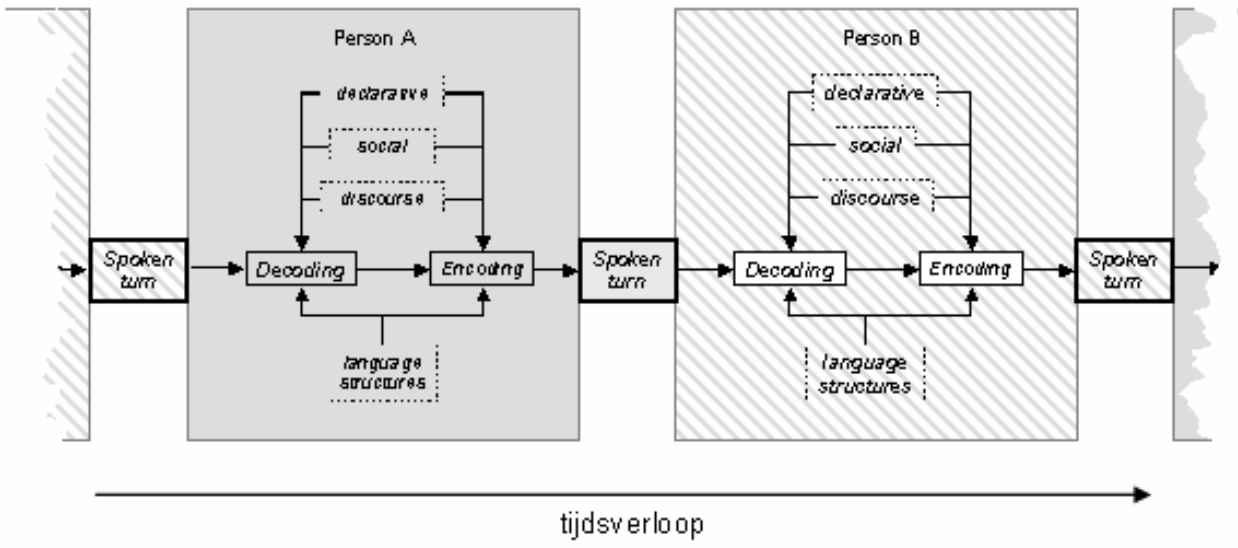


Figuur 2.1: Conversatieproces, naar Levelt 1989

Kandidaten krijgen tijdens het afleggen van de TGN een aantal losse opgaven achter elkaar te horen. Alle opgaven worden in een natuurlijk tempo gesproken door moedertaalsprekers van het Nederlands. Tijdens de toetsafname krijgt een kandidaat verschillende sprekers te horen; mannen en vrouwen met verschillende – lichte – regionale accenten van het Nederlands. De kandidaten moeten op elke opgave een passende reactie geven: zij moeten herhalen wat ze gehoord hebben, een antwoord formuleren op de gestelde vraag of een tegenstelling noemen bij het woord dat zij te horen krijgen. Om te voorkomen dat de toets, die beoogt mondelinge taalvaardigheid te meten, onterechte eisen zou stellen aan bijvoorbeeld de kennis van de wereld, het opleidingsniveau of de culturele identiteit van de kandidaten - die afkomstig kunnen zijn uit alle delen van de wereld en zeer zullen variëren wat betreft onderwijservaring en wereldoriëntatie - is bij de constructie van de TGN het uitgangspunt geweest dat de toets uitsluitend opgaven mag bevatten die door elke kandidaat correct beantwoord kunnen worden *indien taal geen hinderpaal vormt.* In de volgende paragrafen zal nader ingegaan worden op de opgaven van de TGN.

De mate van correctheid en vloeiendheid van de reacties van de kandidaten geeft informatie over de mate van automatisering van de processen die ten grondslag liggen aan hun taalgedrag in het Nederlands. Door de directe interactie tussen de aangeboden stimulus en de respons van de kandidaat worden er eisen gesteld aan de mate waarin kandidaten beschikking hebben over de talige middelen om te reageren en wel op verstaanbare wijze en binnen acceptabele tijd. Immers, in een normale conversatie is tijdig reageren vereist, anders zijn andere sprekers ons voor. Zo zijn er bijvoorbeeld slechts 40 milliseconden tussen het moment waarop een spreker zijn interventie mentaal codeert en de fonologische realisatie (Van Turennout, Hagoort and Brown, 1998). Het totaal beschikbare interval bij beurtwisseling, ‘turn-taking’, is zo’n 500 milliseconden (Bull and Aylet, 1998). Het spreekt daarom vanzelf dat een succesvolle deelnemer aan een gesprek een hoge mate van automatisering nodig heeft. De spreker moet zijn aandacht kunnen reserveren voor ‘wat’ hij wil zeggen en weinig aandacht hoeven te besteden aan ‘hoe’ hij die gedachte zal verwoorden. Automatisering betreft onder andere de vaardigheid om lexicale elementen op te halen, hiermee zinnen te structureren en deze te produceren zonder bewuste aandacht voor de linguïstische code (Cutler, 2003; Jeschinak, Hahne and Schrievers, 2003; Levelt, 2001).

De TGN meet in hoeverre kandidaten de mentale deelprocessen en combinaties daarvan die in het model voor mondelinge taalproductie van Levelt (zie Figuur 2.1) worden onderscheiden automatisch kunnen uitvoeren. Figuur 2.2 plaatst het model dat in Figuur 2.1 werd weergegeven in een context waarin ook meer sociaal-communicatieve aspecten van het taalgedrag worden weergegeven.



Figuur 2.2: Decoderen en coderen als een tijdgebonden kettingproces bij mondelinge interactie

De TGN beoogt het gemak te voorspellen waarmee een kandidaat aan een conversatie in het Nederlands kan deelnemen: het gemak waarmee hij de spraak van anderen kan verwerken en hier adequaat op kan reageren. Doordat de TGN het decoderende en coderende vermogen van spraak door de kandidaat test in een 'real-time' situatie, kan gesteld worden dat de TGN informatie geeft over de mate waarin processen al dan niet geautomatiseerd zijn. Een zekere mate van automatisering is voorwaardelijk voor een succesvolle communicatie. Om in mondelinge communicatiesituaties vlot te functioneren en te voldoen aan alle sociale en retorische aspecten van mondelinge interactie, is het nodig dat taalgebruikers géén aandacht meer nodig hebben voor de kennis van de linguïstische aspecten (Carroll 1961, Carroll 1986, Schneider en Shiffrin (1977). Resultaten op de TGN vormen dan ook een voorspelling van de algemene conversatievaardigheid van de kandidaat. Daarbij wordt aangenomen dat:

- personen die het Nederlands beheersen een (nagenoeg) perfecte score zullen behalen op de TGN en wel ongeacht hun leeftijd of opleidingsniveau;
- personen die het Nederlands niet of in beperkte mate beheersen op de TGN een hogere score halen naarmate zij een betere beheersing van het Nederlands hebben, waarbij geldt dat deze score geen onterechte afhankelijkheid vertoont met achtergrondvariabelen zoals het land van herkomst, het opleidingsniveau, of de mate van geletterdheid.

2.4 Scoring

Bij het leren van een nieuwe taal ontwikkelen de beheersing van de linguïstische code en van de fonologie zich in een tot op zekere hoogte onafhankelijk tempo van bewuste kennis naar een steeds verder geautomatiseerd proces (Higgs and Clifford, 1982; De Jong and Van Ginkel, 1992). In de TGN worden daarom afzonderlijke deelscores gegeven voor ***wat*** de kandidaat zegt en voor ***hoe*** de kandidaat dit zegt.

2.4.1 Wat een kandidaat zegt: inhoudelijke correctheid

Met betrekking tot de inhoudelijke correctheid van de reacties van de kandidaten worden twee subvaardigheden beoordeeld: 'zinsbouw' en 'woordenschat'.
De score voor 'zinsbouw' wordt in de TGN bepaald op basis van herhaalopdrachten: de kandidaten spreken een zin na die ze via de telefoon hebben gehoord. Mensen kunnen zich reeksen van vijftien tot twintig woorden correct herinneren wanneer deze een inhoudelijk verband hebben, maar maken al na vijf of zes woorden fouten wanneer dit verband ontbreekt (Baddely 1986, 2000; Poelmans, 2003). De woorden in een betekenisvolle zin zoals "Daar heb ik nog nooit van gehoord" hebben uiteraard een inhoudelijk verband, maar dat geldt alleen als de hoorder de taal waarin de zin is gesteld kan verstaan. Om langere uitingen te kunnen herhalen, moet men deze verstaan, decoderen, de betekenis opnieuw coderen en vervolgens de uiting reproduceren. Correcte herhaling vergt dus zowel receptieve als productieve beheersing van de linguïstische code met betrekking tot de structuur van Nederlandse zinnen.

De TGN bevat twee soorten opgaven die informatie geven over de 'woordenschat' van de kandidaten: 'Korte vragen' en 'Tegenstellingen'. Bij een korte vraag als bijvoorbeeld "Wat is langer, een arm of een vinger" dient de kandidaat te begrijpen dat het om een vergelijking gaat, op welk aspect wordt vergeleken, en welke zaken worden vergeleken. Op grond van dit begrip kan de kandidaat het juiste antwoord reproduceren. De woordenschat wordt hier dus met name reproductief gemeten. Bij de 'Tegenstellingen' moet een kandidaat het aangeboden woord in de stimulus verstaan en begrijpen (receptieve woordenschat) en vervolgens de gevraagde tegenstelling vinden in zijn lexicon en produceren. Hier wordt de woordenschat productief gemeten. De score die een kandidaat voor de deelvaardigheid 'woordenschat' behaalt, wordt bepaald door de omvang van zijn receptieve en zijn productieve woordenschat: in hoeverre kent de kandidaat de vorm en de betekenis van de woorden die in de items voorkomen en in hoeverre kent hij de woorden die nodig zijn om de antwoorden te geven en kan hij die produceren?

2.4.2 Hoe de kandidaat iets zegt: kwalitatieve correctheid

Ook met betrekking tot de kwaliteit van de reacties van de kandidaten worden in de TGN twee subvaardigheden beoordeeld: 'vloeiendheid' en 'uitspraak'. Beide aspecten van taalvaardigheid worden beoordeeld via de reacties van de kandidaten op de herhaalopdrachten.
In de context van taalverwerving wordt de term 'vloeiendheid' soms gebruikt om mondelinge taalbeheersing in brede zin aan te duiden, zoals in "*Zij spreekt vloeiend Nederlands*". Bij deze toets wordt met 'vloeiendheid' echter gedoeld op een waarneembaar deelaspect bij de uitvoering van spreektaken. In deze zin definieerde Lennon (1990) vloeiendheid als *"... an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently"* (p. 391). Met deze definitie, die gelijk is aan de 'phonological fluency' zoals beschreven door Pennington (1989), geeft Lennon aan dat de waarneembare vloeiendheid een aanwijzing vormt voor de mate van een geautomatiseerd verlopend coderingsproces.

Uitspraak betreft de productie van klinkers, medeklinkers en klemtoon in een taal volgens de regels van die taal. Afwijken van die regels bemoeilijkt de communicatie met de sprekers van die taal. De vaardigheid is afhankelijk van de mate waarin men door ervaring met sprekers van de taal of door een gericht leerproces kennis heeft opgedaan over de gebruikelijke uitspraakvarianten.

2.4.3 De totaalscore

De informatie die de basis vormt van elk van de hierboven beschreven deelscores staat los van de informatie die aan de basis ligt van de andere deelscores.

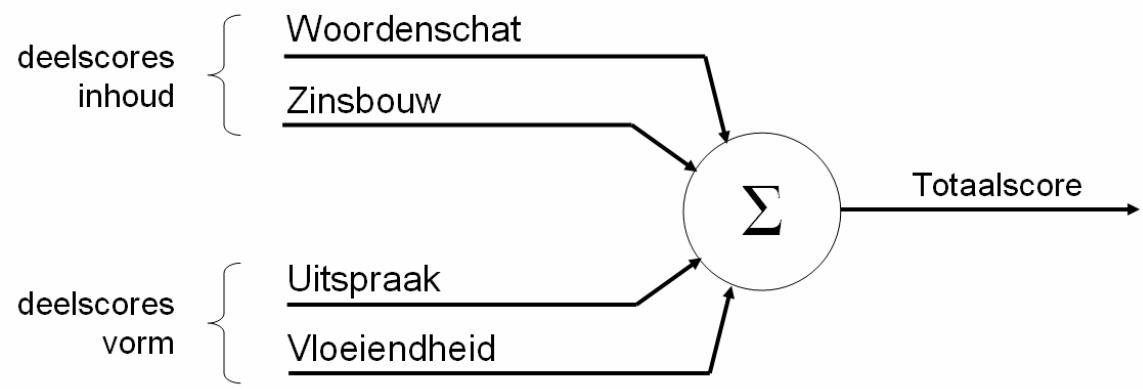
Elke deelscore is gebaseerd op verschillende onafhankelijke aspecten van de gesproken reacties van kandidaten op de verschillende opgaven van de TGN. De deelscores op zinsbouw en woordenschat verschillen uiteraard wat het materiaal betreft in de toetsonderdelen. De verschillen tussen de deelscores op de overige twee subvaardigheden, 'vloeiendheid' en 'uitspraak', betreffen basismaten die gebruikt worden bij het vaststellen van de score. Hierop wordt later in dit hoofdstuk dieper ingegaan.
Tabel 2.1 laat zien hoe de verdeling van de opgaven van de TGN en de vier onderscheiden deelvaardigheden samen gaan. De drie opgavensoorten worden uitvoeriger toegelicht in paragraaf 2.5.

Tabel 2.1: Relatie tussen items en vaardigheden

Onderdeel	Itemtype	Aantal items	Aantal items gescoord	Deelvaardigheid
Deel A	Zinnen herhalen (1)	12	11	Uitspraak Vloeiendheid Zinsbouw
Deel B	Kort-antwoordvragen	14	13	Woordenschat
Deel C	Zinnen herhalen (2)	12	12	Uitspraak Vloeiendheid Zinsbouw
Deel D	Tegenstellingen	10	9	Woordenschat

Bij de bepaling van de eindscore worden de inhoudsdimensie, die aangeeft of de kandidaat de stimulus correct heeft begrepen en er adequaat op heeft gereageerd en de kwaliteitsdimensie, die aangeeft in welke mate de spraak van de kandidaat lijkt op de spraak van moedertaalsprekers, ieder voor 50% meegenomen. Beide dimensies krijgen in het scoringsmodel een even groot gewicht omdat beide dimensies miscommunicatie kunnen veroorzaken wanneer er geen minimaal resultaat behaald wordt. Overdreven aandacht voor uitspraak kan de vloeiendheid van de spraak bijvoorbeeld hinderen en kan zo een vlotte conversatie in de weg staan. Het mag duidelijk zijn dat een onjuist gebruikte of beperkte woordenschat en grammatica het communiceren eveneens kunnen hinderen.

Figuur 2.3 laat zien dat de scores op de vier deelvaardigheden elk met een zelfde gewicht de totaalscore bepalen.



Figuur 2.3: De scores op de deelvaardigheden bepalen met eenzelfde gewicht de totaalscore

2.5 Toetstaken en itemsoorten

De Toets Gesproken Nederlands is ontwikkeld naar het voorbeeld van de SET 10 – 'Spoken English Test' - van Ordinate. Bij de keuze van toetstaken en itemsoorten hebben de doelstellingen van de opdrachtgever en de kenmerken van de doelgroep van het Inburgeringsexamen Buitenland er echter toe geleid dat we op een aantal punten zijn afgeweken van het Engelse voorbeeld.
De SET10 heeft in principe een bereik van nauwelijks of geen beheersing van het Engels tot en met beheersingsniveau C2 van het CEF. De TGN is bij het door de opdrachtgever beoogde gebruik vooral bedoeld om de lagere beheersingsniveaus te onderscheiden. Bij de constructie van de toetstaken is daar rekening mee gehouden door te kiezen voor de ontwikkeling van relatief 'gemakkelijke' opgaven. Zo zijn de herhaalopdrachten in de TGN gemiddeld korter dan de herhaalopdrachten in de SET10 en is bij de korte vragen rekening gehouden met kandidaten die niet kunnen rekenen. Bovendien ontbreken er in de TGN twee itemtypes die wel in de SET10 zijn opgenomen. Het betreft de onderdelen 'voorlezen' en 'zinnen maken'. 'Voorlezen' wordt in de SET10 gebruikt om informatie te verzamelen over de uitspraak en de vloeiendheid van kandidaten. In de TGN komt dit itemtype niet voor omdat de opdrachtgever ook niet-geletterden aan het examen wil laten deelnemen. Vanwege het lage niveau waarop de toets onderscheidend moet kunnen zijn, is het onderdeel 'Sentence Builds' ook niet in de TGN opgenomen. Sentence Builds geven in de SET10 informatie over de uitspraak, de vloeiendheid en de zinsbouw van kandidaten. Overleg met ervaren alfabetiseringsdocenten van ROC's én een experiment onder een groep allochtone deelnemers aan een alfabetiseringstraject leerde dat dit soort opgave niet geschikt is voor ongeletterde kandidaten. Concepten als 'zinsdeel' en 'in de juiste volgorde plaatsen' bleken voor veel deelnemers van alfabetiseringscursussen onbekend, en tijdens een korte instructie niet over te dragen.

Er zijn drie soorten opgaven geselecteerd die voor de hele doelgroep geschikt leken. Deze drie itemsoorten - 'Herhaalopdrachten', 'Korte vragen' en 'Tegenstellingen' - worden hierna beschreven. Daarna wordt de itemsoort 'Verhalen navertellen' beschreven.

2.5.1 Herhaalopdrachten

Taak
Letterlijk een gesproken zin herhalen.

Mondelinge instructie
De kandidaat hoort via de telefoon de volgende instructie:

Nazeggen.
U hoort steeds een zin. Zeg de zin precies na.
*Bijvoorbeeld: een stem zegt: "*Dat is een mooi verhaal*" en u zegt: "*Dat is een mooi verhaal*".*
Nu is het uw beurt. Luister naar de zin en zeg precies na wat u hoort.

Materiaal
Herhaalopdrachten vormen een steekproef van uitingen die men in gesproken taal kan tegenkomen. Zij zijn geput uit authentieke audiobronnen, zoals mondelinge interacties en radio-opnamen. Een groot aantal bestaat uit 'formulaic speech'. De stimuli worden op een alledaagse spontane manier uitgesproken zoals men ze ook in het normale spraakgebruik zou kunnen aantreffen. Stimuli variëren in lengte van twee tot maximaal dertien woorden. De zinnen worden aan de kandidaat in toenemende moeilijkheidsgraad aangeboden. De moeilijkheidsgraad komt over het algemeen overeen met de lengte van de zin.

Voorbeelden
Daar heb ik nog nooit van gehoord.
De volgende keer betaal ik.

Verantwoording
Deze taak betreft imitatie. Bij korte zinnen kan de kandidaat steunen op het korte termijn geheugen en is waarschijnlijk vooral de vaardigheid om uitspraak te kunnen imiteren van belang. Zodra de zin echter meer dan ca 7 woorden bevat, wordt de zogenaamde 'word span' overschreden (Miller, G.A. 1956, Chomsky & Miller, 1963; Miller& Isar, 1964; Baddely, 1986; Baddely, 2000; Poelmans, 2003). De hoeveelheid woorden die een kandidaat kan verstaan, onthouden en reproduceren hangt dan af van de lengte van de woorden maar ook van de bekendheid met die woorden. Bovendien toonden Salamé en Baddeley (1982) aan dat 'irrelevant speech' de uitvoering van repetitietaken negatief beïnvloedt. Delen van de uitingen die men niet verstaat, i.e. herkent, kunnen worden opgevat als 'irrelevant speech'. Inzicht in de structuur van zinnen en vertrouwdheid met de grammatica dragen voorts bij aan het vermogen de zinnen correct te begrijpen en weer te geven (Gibson, 1991; 1998). De uitvoering van dit itemtype is dus niet alleen afhankelijk van 'verstavaardigheid' maar ook van de vertrouwdheid met de taal en de daarin voorkomende woorden én van het begrip ervan. Zinnen herhalen is overigens niet ongebruikelijk in het dagelijkse taalverkeer en wordt bij interactie gebruikt om het gesprek op gang te houden om begrip te tonen (Van Baaren et al, 2003).

2.5.2 Korte vragen

Taak
Met begrip luisteren naar een gesproken vraag en een relevant verstaanbaar gesproken antwoord geven.

Mondelinge instructie:
De kandidaat hoort via de telefoon de volgende instructie:

Vragen
U hoort steeds een korte vraag. Geef op elke vraag een kort antwoord.
*Bijvoorbeeld: Een stem zegt:"*Is januari een dag of een maand*?"*
*En u zegt:"*maand*" of "*een maand*".*
*Of u hoort: "*Een auto, heeft die twee wielen of vier wielen*?"*
*En u zegt: "*vier*" of "*vier wielen*".*
Nu is het uw beurt: Luister naar de vraag en geef dan antwoord.

Materiaal
Korte vragen vragen naar elementaire informatie, of eenvoudige gevolgtrekkingen met betrekking tot tijd, hoeveelheid, lexicale inhoud of logica. De vragen veronderstellen géén vertrouwdheid met specifieke kennis van de Nederlandse cultuur, van Nederlandse gewoonten, geschiedenis et cetera. Uitgangspunt is dat de vragen inhoudelijk moeten kunnen worden beantwoord door kandidaten zonder specifieke kennis van Nederland. Met het oog op zeer laaggeschoolde kandidaten in de doelgroep zijn ook vragen die een beroep doen op - zelfs basale - rekenvaardigheid vermeden.

Voorbeelden
Kun je rijst eten of drinken?
Jan is ouder dan Piet. Wie is het jongst?

Verantwoording
Een vraag begrijpen en daarop antwoord geven behoort tot de niveau-omschrijvingen van alle niveaus in het CEF. Korte vragen toetsen of kandidaten vragen - die in het Nederlands worden gesteld - begrijpen en of zij het antwoord op die vragen in het Nederlands kunnen geven.

Omdat er alleen vragen worden gesteld waarvan verondersteld mag worden dat de kandidaten het antwoord erop in hun eigen taal zouden moeten kennen, toetsen de opgaven in hoeverre kandidaten in staat zijn om de woorden in een Nederlandse vraag te identificeren, te begrijpen in hun onderlinge betekenisrelatie, de gestelde vraag te interpreteren, het juiste antwoord te formuleren en dat op verstaanbare wijze te produceren.

Nadere kenmerken
Het te verwachten antwoord moet kort zijn en bepaald, maar er mag méér dan één antwoord goed zijn. Bij het eerst gegeven voorbeeld kan de kandidaat antwoorden met *rijst kun je eten*, maar *eten* of *dat kun je eten* is ook goed. De vragen bevatten maximaal vijfentwintig woorden en bestaan vaak uit meer dan een enkele zin. Daarbij is gestreefd naar variatie in de structuur van de vragen. De vragen zijn verder niet formeel en zo veel mogelijk geformuleerd in spreektaal. Er zijn geen vragen ontwikkeld die met 'ja' of 'nee' dan wel 'goed' of 'fout' kunnen worden beantwoord.

2.5.3 Tegenstellingen

Taak
Van een begrip het tegengestelde zeggen.

Mondelinge instructie
De kandidaat hoort via de telefoon de volgende instructie:

Tegenstellingen.
U hoort steeds een woord. U zegt het tegenovergestelde.
*Bijvoorbeeld: u hoort "*hoog*", dan zegt u "*laag*", of u hoort "*niet*". Dan zegt u: "*wel*".*
Nu is het uw beurt. Luister naar het woord en zeg het tegengestelde woord.

Voorbeelden
Niet
Ochtend

Materiaal
De opgaven bestaan uit voorzetsels, zelfstandige naamwoorden, bijvoeglijke naamwoorden en werkwoorden waarvoor duidelijke tegenstellingen bestaan in gesproken Nederlands. Bij veel woordparen kunnen de twee leden allebei optreden als stimulus *en* als respons. In dat geval worden deze woorden in de itembank gemarkeerd voor onderlinge uitsluiting zodat ze niet samen in eenzelfde toets kunnen voorkomen.

Verantwoording
Het geven van een tegenstelling vereist zowel receptieve als productieve kennis van vocabulair. De kandidaat moet de stimulus verstaan en de betekenis ervan begrijpen (receptief) en vervolgens het woord noemen dat met de stimulus een tegenstelling vormt (productief). In mondelinge interactie wordt veelvuldig beroep gedaan op het kennen van tegenstellingen. De mogelijke conversatie: "*Kun jij morgenochtend? Nee, wel morgenmiddag."* bevat al twee tegenstellingen.

Nadere kenmerken
De woorden komen voor in de alledaagse spreektaal. Met het oog op de automatische spraakherkenning die moet plaatsvinden, worden geen paren opgenomen waarvan de tegenstelling wordt gevormd door 'on' en dergelijke voor de stimulus te plaatsen. Kandidaten krijgen de instructie met een tegengesteld woord te antwoorden. Antwoorden waarbij 'niet' voor de stimulus wordt geplaatst (hoog - niet hoog) worden als fout gescoord.

2.5.4 Verhalen navertellen

Voor validatie doeleinden wordt de TGN afgesloten met een open opdracht: Verhalen navertellen. Deze opdracht maakt geen onderdeel uit van de eigenlijke toets. De opdracht van het onderdeel Verhalen navertellen is als volgt gedefinieerd:

Taak
Een kort verhaal beluisteren en in eigen woorden navertellen.

Mondelinge instructie
De kandidaat hoort via de telefoon de volgende instructie:

Verhalen navertellen.
U hoort een kort verhaal. U moet het verhaal navertellen. U krijgt daarvoor 30 seconden. Vertel zoveel mogelijk. Denk bijvoorbeeld aan: Wie deden er mee? Wat gebeurde er? Waar was het? En, hoe liep het af?

Materiaal
De korte verhalen tellen in totaal 40 tot 90 woorden. Zij zijn globaal volgens het volgende stramien opgebouwd:

- Handeling 1 en introductie van hoofdpersoon en eventueel bijrol.
- Handeling 2 mogelijk een reactie op handeling 1, of een gevolg ervan.
- Afsluiting: nieuw ontstane situatie, gevoelens van hoofdpersoon en/of bijrol.

Voorbeeld
'Fred reed naar huis. Hij was niet blij want het gesprek met de laatste klant was niet zo goed verlopen. Fred had geen goede indruk op die klant gemaakt. Die zou vast niets van hem willen kopen. Toen hij de sleutel in het slot stak, besefte Fred dat hij ook nog zijn tas bij de klant had laten staan.'

De navertelde verhaaltjes worden uitsluitend gebruikt voor validatie van de toets en kunnen ook aan gebruikers van de toets een indruk geven van het taalniveau van de betreffende kandidaten. Om de instructies zoveel mogelijk te beperken én om te voorkomen dat kandidaten de laatste twee opgaven overslaan, worden kandidaten niet geïnformeerd over de specifieke functie van de laatste twee opgaven.

2.6 De samenstelling van de toets

Elke toets bestaat uit 50 opgaven. De toetsen worden samengesteld uit een itembank via willekeurige naar itemtype gestratificeerde selectie, waardoor elke toets in principe een unieke deelverzameling items bevat. De taken worden in vier onderdelen gepresenteerd. Ieder onderdeel wordt voorafgegaan door benoeming van de naam van het onderdeel, instructies en voorbeelden. Bij aanvang van elke nieuwe taak wordt ook een oefenitem aangeboden. De reacties op deze oefenitems worden niet gescoord.

De herhaalopdrachten worden in twee sets van twaalf opgaven gepresenteerd om te voorkomen dat kandidaten hun concentratie verliezen. De items worden bovendien per set van twaalf in volgorde van oplopende moeilijkheid gepresenteerd.

Korte vragen en tegenstellingen worden respectievelijk in een set van veertien en in een set van tien opgaven aangeboden.
Bij de trekking van de opgaven voor opname in een onderdeel van de toets worden drie strata gehanteerd: gemakkelijk, gemiddeld en moeilijk.

Opgaven uit het stratus 'gemakkelijk' worden het eerst aangeboden, vervolgens opgaven uit de stratus 'gemiddeld' en tenslotte de opgave uit het stratus 'moeilijk'. De moeilijkheidsgraad van de opgaven is bepaald op basis van de pretests.

Tabel 2.2 geeft een beeld van de samenstelling van de toets.

Tabel 2.2: Toetssamenstelling

Onderdeel	**Taak**	**Aantal opgaven**	**Aantal gescoorde opgaven**
Deel A	*Instructie incl. twee voorbeelden* Herhaalopdrachten (deel 1)	12	11
Deel B	*Instructie incl. twee voorbeelden* Korte vragen	14	13
Deel C	*Instructie incl. twee voorbeelden* Herhaalopdrachten (deel 2)	12	12
Deel D	*Instructie* Tegenstellingen	10	9
Deel E	*Instructie* Verhalen navertellen	2	
Totaal		50	45

2.7 Itemontwikkeling

Nadat de itemspecificaties waren vastgesteld, is gestart met de ontwikkeling van de opgaven. Auteurs waren medewerkers van het team Talen & Burgerschap van de afdeling Onderwijsinnovatie van CINOP en LTS.

2.7.1 Het vocabulair in de opgaven

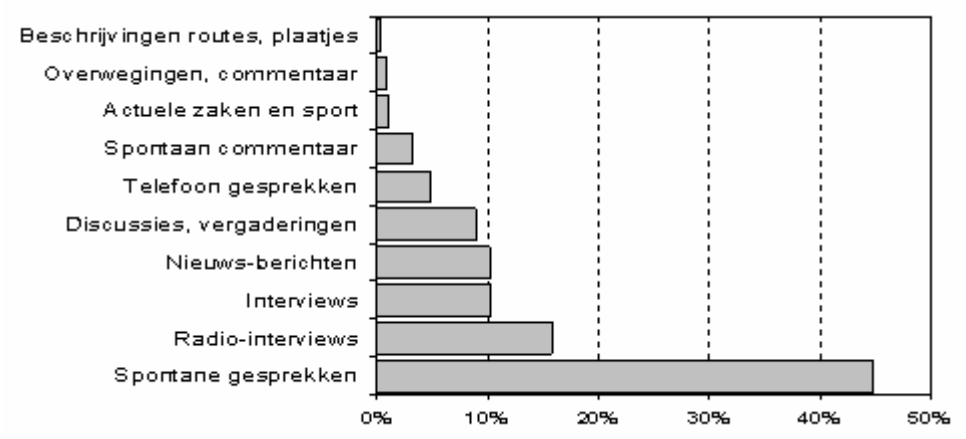
Alle ontwerpitems werden in eerste instantie gecontroleerd op woordfrequentie met gebruikmaking van de in januari 2004 beschikbare pre-release van het Corpus Gesproken Nederlands (Release 6). Dit corpus bevat opnamen van gesproken Nederlands in Nederland en in België. De pre-release bevatte nog enkele onvolkomenheden, daarom is de definitieve itembank in mei 2005 ook tegen het definitieve corpus afgezet. Later in deze paragraaf wordt hier verder op ingegaan.

In het Corpus Gesproken Nederlands (CGN) is de volgende, voor ons doel geschikte, selectie gemaakt uit de beschikbare taalsamples:

- van alle beschikbare leeftijden;
- van alleen de streken in Nederland;
- van de onderstaande teksttypen;
- spontane gesprekken (face-to-face);
- interviews;
- telefoon gesprekken;
- radio-interviews;
- discussies, vergaderingen;
- beschrijvingen van routes en plaatjes;
- spontaan commentaar;
- actuele zaken en sport;
- nieuwsberichten;
- overwegingen/commentaar.

De teksttypen “zakelijke onderhandelingen”, “lezingen/toespraken”, en “hardop voorgelezen teksten” werden op inhoudelijke gronden uitgesloten; het teksttype “lessen” was op het moment waarop de items werden ontwikkeld nog niet beschikbaar.

Het geselecteerde deel van het in januari 2004 beschikbare corpus bevatte circa 2.4 miljoen woorden verdeeld over ruim 48 duizend lemma’s. Hiervan werd een in frequentie aflopende tabel aangelegd. Bijna de helft van de woorden in het geselecteerde deel van het corpus (45%) was afkomstig uit het teksttype ‘Spontane gesprekken’.Van de lemma’s in het geselecteerde deel van het corpus was ruim 62% afkomstig uit spontane gesprekken. Figuur 2.4 toont de herkomst naar tekstsoort van de woorden in het geselecteerde deel van het CGN.



Figuur 2.4 Verdeling bronnen voor lexicon in geselecteerd deel van CGN

Vervolgens werd een beperking met betrekking tot de frequentie aan de in de toets te gebruiken woorden opgelegd: de woorden in de toets zouden een frequentie van 10 of hoger moeten hebben in het geselecteerde deel van het corpus. Dit kwam ongeveer overeen met de 7.000 meest frequente lemma’s.

Tabel 2.3 vermeldt aantallen lemma’s en woorden en de verdeling over de teksttypen in het geselecteerde deel van het corpus.

Tabel 2.3: Verdeling van lemma's en woorden over de teksttypen in het geselecteerde deel van het CGN (release 6)

	Spontane gesprekken	Radio-interviews	Interviews	Nieuwsberichten	Discussies, vergaderingen	Telefoon gesprekken	Spontaan commentaar	Actuele zaken en sport	Overwegingen, commentaar	Beschrijvingen routes, plaatjes	Sommering over geselecteerde teksttypen
Aantal lemma's	6.217	4.124	3076	5684	3909	872	2348	2113	5416	2373	7247
% alle lemma's	13%	8%	6%	12%	8%	2%	5%	4%	11%	5%	15%
% selectie lemma's	86%	57%	42%	78%	54%	12%	32%	29%	75%	33%	100%
Aantal woorden (x1.000)	1.047	238	110	368	208	8	74	24	237	20	2333
% van gehele corpus	43.2%	9.8%	4.5%	15.2%	8.6%	0.3%	3.1%	1.0%	9.8%	0.8%	96.4%
% van selectie	44.9%	10.2%	4.7%	15.8%	8.9%	0.3%	3.2%	1.0%	10.1%	0.9%	100%

Met gebruikmaking van een daartoe ontwikkeld hulpprogramma CHECKTEST.EXE werden alle itemteksten (vragen en verwachte antwoorden) getoetst op woordenschat. Het programma vraagt als input een DOS tekstbestand en opgave van een minimale lemmafrequentie en levert als output een rapport bestaande uit een lijst van alle woordvormen die wel in de itemteksten maar niet in het geselecteerde deel van het corpus voorkomen samen met het lemma, het lemmarangnummer en de lemmafrequentie (zie Bijlage 3).

Aangezien het corpus ten tijde van de ontwikkeling van de items nog relatief klein was en ook nog fouten bevatte, werd de lijst gescreend op aannemelijkheid. Wanneer bijvoorbeeld het lemma "zuid" als te weinig frequent werd gerapporteerd hebben wij het corpus nader geïnspecteerd. Zo bleek bij het genoemde voorbeeld dat naast het lemma "zuid" een lemma "zuiden" was opgenomen dat wel een frequentie van 10 of meer had. In zo'n geval werd een woord alsnog beschouwd als geschikt om in de toets op te nemen. Ook werden sommige woorden om inhoudelijke reden toegelaten. Een woord als "uitgang" (lemmafrequentie 9 in release 6) hoort volgens ons tot de noodzakelijke en relevante woordenschat van taalleerders. Het CGN diende in de toenmalige vorm als hulpmiddel bij het inspecteren van het vocabulair, niet als criterium.

In mei 2005 werd de definitieve itembank vergeleken met de eerste officiële uitgave van het Corpus Gesproken Nederlands. Het corpus bevatte toen in de voor de samenstelling van de toets geselecteerde domeinen 4,7 miljoen woorden. Daarvan behoren 4,5 miljoen woorden tot de 7.000 meest frequente woorden in het corpus. De minimale lemmafrequentie voor deze woorden is 17. Bijlage 3 bevat alle woorden die in de uiteindelijke itembank voorkomen met een frequentie lager dan 17 in de officiële uitgave van het CGN.

2.7.2 Itemrevisie

In januari 2004 werden door in totaal 12 auteurs 2.500 items geschreven. Alle items werden in een schriftelijke commentaarronde ter beoordeling voorgelegd aan twee NT2-deskundigen en aan het toenmalige hoofd van de afdeling toetsconstructie van Ordinate, linguïste en moedertaalspreker van het Nederlands.

Aan de beoordelaars van de items werden de volgende algemene criteria voorgelegd:

- items moeten gewone spreektaal bevatten: informeel of enigszins formeel;
- items mogen geen inhoudelijke vraagstelling bevatten die niet door een moedertaalspreker van twaalf jaar kan worden beantwoord;
- items mogen niet of nauwelijks een beroep doen op andere vaardigheden of kennis dan uitsluitend de vaardigheid Nederlands te verstaan en te spreken;
- items mogen geen 'gevoeligheden' raken: cultureel noch persoonlijk (geen items over bijvoorbeeld dood, gehandicapten, etnische groepen);
- items mogen geen laagfrequente woorden bevatten; hierop is de tekst al gecontroleerd met behulp van het Corpus Gesproken Nederlands. Ook eventuele voorstellen voor wijziging zullen met behulp van dit corpus worden gecontroleerd;
- voor specifieke criteria waaraan de items binnen een opgavensoort moeten voldoen: zie de betreffende itemspecificatie.

Tabel 2.4 geeft een overzicht van het soort commentaar dat werd gegeven op het eerste ontwerp van de items en de frequenties ervan. In totaal werd op 25% van de items commentaar van een of meer van de NT2-deskundigen ontvangen.

Tabel 2.4: Overzicht soorten commentaar op items

Type opmerking	**Voorbeeld item** *(verwacht antwoord)*	**Percentage**
Herhaalopdrachten		
Te directe aanspreking	Ik kan u niet verstaan.	5.0
Lijkt op instructie	Wilt u wat harder spreken.	6.0
Geen spreektaal	In de bibliotheek kun je boeken lenen.	2.0
Te stereotiep Nederlands	De koeien staan dit jaar wel erg vroeg in de wei.	1.5
Mogelijk te emotioneel	De dokter kwam te laat.	1.0
Onderwerp niet geschikt	Wat was dat examen moeilijk.	1.0
Te idiomatisch	Doe maar twee tientjes.	2.0
Kort antwoord vragen		
Verwachte bias opleiding	Eindigt een zin met een punt of met een komma? (*punt)*	1.5
Geen eenduidig goed antwoord	Hoeveel wielen heeft een fiets? (*twee)*	1.0
Niet geschikt voor automatische beoordeling	Wat legt een kip? Een ei of een ui? (*ei)*	0.5
Culturele bias	Op welke datum begint het jaar? *(1 januari)*	1.0
Niet algemene feitenkennis	Waar wordt kaas van gemaakt? *(melk)*	1.0
Tegenstellingen		
Teveel mogelijke alternatieven	Ouderwets (*modern)*	1.0
Geen tegenstelling	Peper (*zout)*	0.5
Totaal % items met commentaar		25.0

Gelet op het beperkte aantal experts werd het niet juist geacht het commentaar te wegen naar het aantal deskundigen dat een probleem met een item signaleerde. Ieder commentaar werd overwogen en gaf aanleiding tot een van drie volgende beslissingen (tussen haakjes het percentage van het totale aantal becommentarieerde items):

- het commentaar kon worden weerlegd (ca 5%);
- het commentaar werd verwerkt in een wijzigingsvoorstel (ca 15%);
- het commentaar leidde tot verwerping van het item (ca 5%).

Alle items evenals de toets- en iteminstructies werden vervolgens opgenomen in een professionele geluidsstudio onder regie van een van de itemconstructeurs en een geluidstechnicus. De opgaven werden ingesproken door tien verschillende vrouwelijke en mannelijke sprekers afkomstig uit diverse streken in Nederland.

Daarnaast werden de instructies ingesproken door twee professionele stemacteurs, een mannenstem voor de algemene instructies en een vrouwenstem voor de specifieke aanwijzingen bij de verschillende itemtypes.

De gehele set opgaven, voorzien van de op grond van het commentaar genomen beslissingen en wijzigingen en de bijbehorende geluidsopnamen, werd in een vergadering van twee NT2-deskundigen en één itemconstructeur doorgenomen. In deze vergadering werden de meeste voorstellen voor revisie overgenomen. Naast de reeds op basis van de geschreven vorm verworpen items werd echter nog circa 5% van de totale set verworpen op basis van de gesproken opname. Daarmee werd in totaal circa 10% van de ontworpen items afgewezen. De meeste verworpen items betroffen Kort-antwoordvragen en Tegenstellingen. Omdat de te gebruiken woordenschat beperkt is, kunnen voor deze itemtypen niet gemakkelijk alternatieven worden ontwikkeld. Afgewezen herhaalopdrachten konden wel gemakkelijk worden aangepast of vervangen. Deze werden in een nieuwe sessie in de geluidstudio opgenomen.

2.8 De verzamelde items vóór de pretest

Na de hiervoor geschetste procedures hadden we ten behoeve van de pretest een verzameling van 2.131 items die aan alle criteria voldeden. Tabel 2.5 geeft een overzicht.

Tabel 2.5 De samenstelling van de itembank voor de pretest

Itemsoort	Aantal items
Herhalingen	1.102
Korte vragen	605
Tegenstellingen	424
Totaal	**2.131**

2.9 Niveaudefinitie

In principe heeft de TGN een bereik van nauwelijks of geen beheersing van het Nederlands tot beheersing van CEF-niveau B2 of hoger. De opdrachtgever heeft de TGN laten ontwikkelen om de luister- en spreekvaardigheid in het Nederlands te meten van aspirant nieuwkomers in de landen van herkomst. De TGN is ook geschikt voor de toetsing van nieuwkomers en nader te bepalen groepen oudkomers die al in Nederland wonen. Bij het afronden van dit eindrapport zijn de eisen die aan de kandidaten gesteld gaan worden nog niet wettelijk vastgesteld. Bij de constructie van de TGN is daarom uitgegaan van de adviezen daaromtrent van de Commissie Franssen. Voor wat betreft de eisen die gesteld worden aan inburgering in het buitenland geldt dan A1-min als norm voor beheersing van spreek- en luistervaardigheid. Voor inburgering in Nederland gaat het om de verplichting voor deze mondelinge vaardigheden tot beheersing van niveau A2. In deze paragraaf worden beide beheersingsniveaus nader omschreven en worden de omschrijvingen gepresenteerd die in het kader van dit project zijn geformuleerd. De geformuleerde definities zijn gebruikt ten behoeve van de menselijke beoordeling van de mondelinge vaardigheid van de kandidaten, het bepalen van de grensscores en de validering van de toets.

2.9.1 Het niveau A1-min

Op grond van overwegingen van functionaliteit, redelijkheid en selectiviteit naar opleiding en motivatie adviseerde de commissie Franssen in februari 2004 het vereiste niveau voor mensen die beogen zich in Nederland te vestigen op A1-min te stellen.

De commissie zegt over de keuze voor het niveau A1-min (2004a:34):
"Indien de overheid niet bereid is om hoge investeringen in materiaalontwikkeling te doen, is alleen een forse beperking van het niveau nog denkbaar om tot een acceptabel niveau voor het examen te komen, mede gelet op de juridische randvoorwaarden. Daartoe is een derde variant ontwikkeld. Deze variant noemen we het A1-min niveau. Het is gebaseerd op de omschrijving van de Australian Second Language Proficiency Ratings (ASLPR) van Ingram en Wylie (1984). Het taalniveau wordt daarin aangeduid als 'formulaic proficiency'(0+): men is in staat om in de meest vertrouwde en voorspelbare omgevingen taal te gebruiken, waarbij men vooral een beperkt repertoire aan standaarduitingen toepast (...).

De vereisten voor dit scenario hebben ook slechts betrekking op de mondelinge vaardigheden.
Men kan slechts een beperkt aantal vertrouwde woorden en basiszinnen begrijpen die betrekking hebben op de directe, persoonlijke levenssfeer en op de allereerste levensbehoeften; en alleen in direct contact met Nederlandssprekenden die gewend zijn zich aan te passen
Men kan zich slechts in zeer beperkte mate uitdrukken, eigenlijk alleen met behulp van losse woorden en standaardformuleringen ('formulaic speech'), op een gering aantal terreinen die verband houden met de directe, persoonlijke levenssfeer."

De ASLPR wordt sinds 1997 aangeduid met ISLPR (International Second Language Proficiency Ratings). Het niveau 0+ wordt door de auteurs van de ISLPR gedefinieerd als het middelste gedeelte van het 0-niveau van de ILR schaal (Inter-Agency Language Roundtable) ook bekend onder de naam FSI schaal (Foreign Service Institute). Dit 0-niveau van de ILR schaal wordt op de voor het Amerikaans onderwijs van de ILR afgeleide ACTFL schaal onderverdeeld in drie niveaus: *Novice Low, Novice Mid* en *Novice High*. Deze drie niveaus komen op de ISLPR overeen met de niveaus 0 (*zero proficiency*) 0+ (*Formalaic proficiency*) en 1- (*Minimum Creative Proficiency*). Volgens een van de co-auteurs van het CEF (Brian North) en een van de co-auteurs van de ILR en ACTFL schalen (Pardee Lowe) komt het *Novice mid* niveau overeen met A2.1 van het CEF. Dit ligt dus belangrijk hoger dan het door de commissie gecreëerde niveau A1-min.

Het is helder dat de commissie met de gegeven omschrijving doelt op een niveau lager dan het niveau A1 op de schaal van de Raad van Europa. North (2000) heeft in zijn onderzoek waarop de schaling van het CEF is gebaseerd inderdaad descriptoren voor een niveau onder A1 aangetroffen. Bij de publicatie van het CEF is dit echter niet in de schaal opgenomen, omdat het aantal descriptoren te gering werd bevonden.

Uitgaande van het advies van de Commissie Franssen hebben wij ten behoeve van de ontwikkeling van het inburgeringsexamen Nederlands in het buitenland op grond van bijlage 5 van dat advies en in combinatie met het beperkte aantal descriptoren genoemd door North (2000), de volgende omschrijving van het niveau A1-min geformuleerd:

'Mondelinge interactie A1-min'
Kan met behulp van losse woorden zaken van direct persoonlijk belang communiceren.
Gebruikt losse woorden, enkele standaarduitdrukkingen en elementaire beleefdheidsfrases maar is vanwege uitspraak moeilijk te verstaan. Begrijpt eenvoudige direct tot hem/haar gerichte en met zorg gesproken vragen naar of mededelingen over personalia, en een beperkt aantal concrete alledaagse begrippen. Kan vragen over dergelijke zaken soms ook met een of meer losse woorden beantwoorden. Conversatie is echter niet mogelijk.

Bovenstaande omschrijving is gehanteerd door de menselijke beoordelaars en bij de validering van de toets.

Ter vergelijking citeren we hier de schaal van de American Council of Teachers of Foreign Languages (ACTFL, 1999; Breiner et al., 2000) waarin het niveau 'Novice Low' als volgt wordt gedefinieerd:
"Speakers at the Novice-Low level have no real functional ability and, because of their pronunciation, they may be unintelligible. Given adequate time and familiar cues, they may be able to exchange greetings, give their identity, and name a number of familiar objects from their immediate environment. They are unable to perform functions or handle topics pertaining to the Intermediate level, and cannot therefore participate in a true conversational exchange

2.9.2 Het niveau A2

Voor wat betreft het te eisen mondelinge beheersingsniveau van het Nederlands van personen die in Nederland onder de Wet Inburgering Nieuwkomers vallen, adviseert de Commissie Franssen niveau A2. De commissie zegt over dit niveau (2004b):
"Het niveau A2 is te beschouwen als een eenvoudig beheersingsniveau, waarmee men in beperkte mate kan communiceren over vertrouwde en alledaagse zaken. Het biedt volgens de Commissie een goede uitgangspositie voor een verdere integratie in de samenleving. Het niveau omvat de meest gangbare, eenvoudige taalfuncties en kent een woordenschat van ongeveer 2.000 woorden, zowel hoogfrequente woorden als domeinspecifieke woorden afkomstig uit de voor de inburgeraar relevante domeinen. Er is daarmee een redelijke mate van communicatie mogelijk met de directe omgeving van de inburgeraar. Om te kunnen participeren in de Nederlandse samenleving is een dergelijk taalvaardigheidniveau noodzakelijk."

De commissie geeft de volgende omschrijvingen van dat niveau (zie voor details bijlage 3 van het rapport van de commissie Franssen et al, 2004b):

Luisteren:
"Men kan zinnen en de meest frequente woorden begrijpen die betrekking hebben op gebieden die van direct persoonlijk belang zijn (bijvoorbeeld basisinformatie over zichzelf en zijn/haar familie, winkelen, plaatselijke omgeving, werk). Men kan de belangrijkste punten in korte, duidelijke, eenvoudige boodschappen en aankondigingen volgen."

Spreken:
"Men kan een reeks uitdrukkingen en zinnen gebruiken om in eenvoudige bewoordingen familie en andere mensen, leefomstandigheden, opleiding en huidige of meest recente baan te beschrijven."

Gespreksvaardigheid:
"Kan communiceren over eenvoudige en alledaagse taken die een eenvoudige en directe uitwisseling van informatie over vertrouwde onderwerpen en activiteiten betreffen. Kan zeer korte sociale gesprekken aan, alhoewel hij/zij gewoonlijk niet voldoende begrijpt om het gesprek zelfstandig gaande te houden."

Uitgaande van het advies van de Commissie Franssen hebben we ten behoeve van de ontwikkeling van de Toets Gesproken Nederlands de volgende beschrijving van niveau A2 geformuleerd:

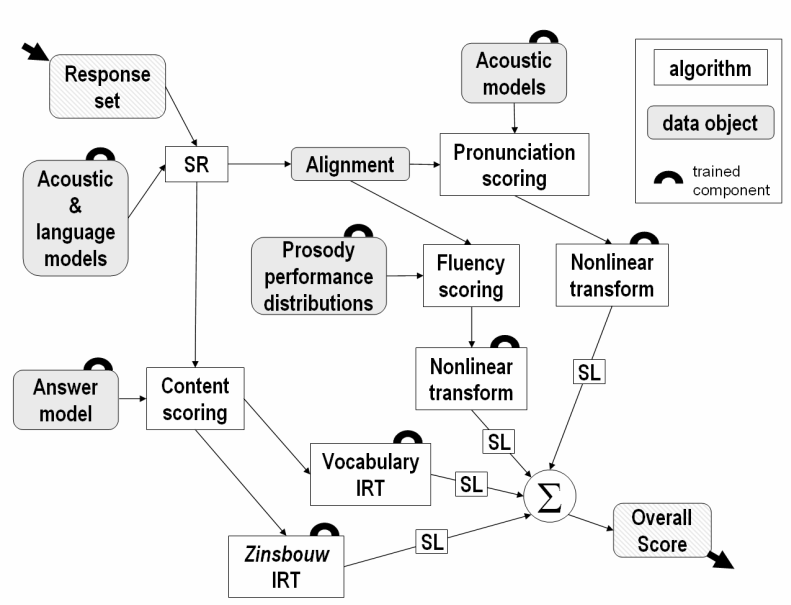
'Mondelinge interactie A2'
Communiceert basisinformatie over werk, achtergrond, familie, vrije tijd, et cetera.
Kan zichzelf in korte zinnen verstaanbaar maken, hoewel pauzes, valse starts, en herformuleringen evident aanwezig zijn. Uitspraak is over het algemeen helder genoeg om te worden verstaan ondanks een duidelijk buitenlands accent. Gebruikt een beperkt aantal eenvoudige structuren correct, maar maakt systematisch elementaire fouten.

Kan woordgroepen verbinden met eenvoudige voegwoorden zoals "en", "maar", en "omdat". Kan zich tot hem/haar richtende, duidelijk sprekende moedertaalsprekers verstaan, wanneer zonodig om herhaling gevraagd kan worden.

Bij de ontwikkeling van de Toets Gesproken Nederlands is deze omschrijving gehanteerd bij de validering van de toets en bij de beoordeling door menselijke beoordelaars.

2.10 Het 'runtime' scoringssysteem

In deze paragraaf wordt het 'runtime' scoringssysteem dat Ordinate gebruikt voor het automatisch scoren van de Toets Gesproken Nederlands beschreven. Het systeem wordt geïllustreerd in Figuur 2.5. Woorden die in deze beschrijving *cursief* zijn weergegeven, verwijzen naar onderdelen van Figuur 2.5. De legenda bevindt zich in de rechter bovenhoek van de figuur.



Figuur 2.5: Ordinate's Runtime Scoringsysteem

Het automatische scoringssysteem heeft één input en één output. De input bestaat uit de *Response set*: de digitaal opgenomen antwoorden op de toetsitems van één bepaalde kandidaat. De output van het scoringssysteem is een *Overall score:* een totaalscore die kan variëren van 0 tot 90 en die wordt afgerond op gehele getallen. Scores groter dan 80 worden gerapporteerd als 80 en scores kleiner dan 10 worden gerapporteerd als 10.

De rechthoeken in Figuur 2.5 en in de hierna te presenteren figuren tot en met Figuur 2.10 stellen algoritmes voor. De afgeronde rechthoeken stellen datastructuren voor. Rechthoeken met het dikke halve-cirkel symbool verwijzen naar systeemcomponenten die getraind zijn met data die tijdens de pretest verzameld zijn.

Data stromen in Figuur 2.5 van de linkerbovenhoek naar de rechterbenedenhoek. De input van het scoringssysteem, de *Response set*, bestaat uit een aantal audiobestanden in het formaat ITU G.711 (volgens aanbeveling G.711 van de Internationale Telecommunicatie Unie, Genève, Zwitserland, november 1988). De *Response set* bevat alle antwoorden voor een enkele toetsafname, afgelegd door een individuele kandidaat. De *Response set* wordt eerst langs de Spraakherkenner geleid (hierna *SR*: Speech Recognizer).

De eerste stap van de *SR* is 'preprocessing', dan wel 'voorbehandelen'. Daarbij worden overlappende Hamming vensters van 25 ms met een venstervoortgang van 10ms op het audiosignaal gelegd. Bij elk frame worden 26 mel-frequentie cepstral coëfficiënten berekend waaronder 13 delta cepstral coëfficiënten. Deze worden herkend door gebruik te maken van een Viterbi search door een netwerk van Gaussian mixture hidden Markov modellen.

De *SR* produceert *Alignment, '*oplijningsdata*'*. *Alignment* vormt de input voor de *Pronunciation scoring,* de beoordeling van uitspraak aan de hand van een set van *Acoustic models*, akoestische modellen. *Alignment* wordt ook gebruikt door het algoritme voor *Fluency scoring*, de beoordeling van vloeiendheid, aan de hand van een set *Prosody performance distributions* (prosodie prestatie distributies). Zowel Uitspraak als Vloeiendheid ondergaan dan *Nonlinear transform* (non-lineaire transformaties) waarvoor de parameters worden bepaald gedurende training. Wanneer kandidaten minder dan 5 herhaalopdrachten uitvoeren waarvan minimaal 25 procent van de correcte antwoordstring kan worden herkend, wordt er geen score voor vloeiendheid en uitspraak toegekend.

De *SR* produceert ook de a posteriori meest waarschijnlijke lexicale transcriptie van elke uiting in de *Response set*. Deze lexicale 'string' wordt gebruikt door het algoritme voor *Content scoring* (inhoudscoring) op grond van een *Answer model* (antwoordmodel), dat getraind is aan de hand van data.

Voor elk antwoord op een Herhaalopdracht berekent het algoritme voor *Content scoring* een discrete polytome inhoudsscore die de Levenshtein afstand is tussen de correcte antwoordstring en de herkende string. De *SR* herkent aarzelingen (gevulde pauzes) en het algoritme voor inhoudscoring verwijdert deze van de herkende string alvorens de Levenshtein afstand te berekenen.

Voor Korte vragen en Tegenstellingen, berekent het algoritme voor *Content scoring* een dichotome score van 1 voor een correct antwoord en 0 voor een incorrect antwoord. Het algoritme voor *Content scoring* negeert spraak aan het begin en einde van een uiting wanneer het correcte antwoord zich in het midden van de uiting bevindt. Als een respons een correct én een incorrect antwoord bevat, wordt het antwoord een score van 0 toegekend. Aan deze wijze van scoring kleeft in principe het nadeel dat zelf-correcties die resulteren in een correct antwoord (bijvoorbeeld: "rijst kun je drinken … uh… drinken… uh eten natuurlijk") ten onrechte als fout worden gescoord. Uit inspectie van alle antwoorden op Korte vragen die tijdens de pretests zijn verzameld, blijkt dat dergelijke zelfcorrecties zelden voorkomen. Tijdens de pretests zijn in totaal 21378 antwoorden op korte vragen verzameld. Van die totale set reacties waren er in totaal 179 reacties (0.84%) die zowel het goede als het foute antwoord bevatten. Scoring van deze 179 reacties door menselijke beoordelaars leverde in totaal 35 antwoorden op die beschouwd konden worden als 'juiste zelf-correcties' en die derhalve door het automatische scoringsmodel ten onrechte als 'fout' werden gescoord. Het betreft 0.16% van alle reacties. Uitkomsten van onderzoek naar zelf-correcties van eerste- en tweede taalgebruikers ondersteunen de beslissing om genoegen te nemen met deze beperking van het automatische scoringsmodel (zie bijv. Van Hest, 1996). Er bestaat een negatieve samenhang tussen het taalvaardigheidsniveau van taalgebruikers en de mate waarin zij gebruik maken van zelf-correcties.

Content scoring produceert discrete scores voor zinsbouw en woordenschat. Op grond van beide deelscores worden onafhankelijk van elkaar vaardigheidsschattingen gegenereerd met een 1-parameter IRT model. Het algoritme voor *Scaling and Limiting* (SL, 'schaling inperking'*)* herschaalt de parameters voor Zinsbouw, Woordenschat, Uitspraak en Vloeiendheid lineair (zie Hoofdstuk 5) en perkt ze in een eerste stap in tot een bereik van 0 tot 90 . Daarmee worden de vier deelscores (voor Uitspraak, Vloeiendheid, Zinsbouw en Woordenschat) geproduceerd.

De inperking van de deelscores van 0 tot 90 wordt toepast om te vermijden dat extreme deelscores bij de berekening van de totaalscore een te groot gewicht krijgen. Vervolgens worden zoals reeds eerder gezegd de vier deelscores met gelijke gewichten gecombineerd tot een *Overall score* (Totaalscore). Tenslotte worden voor de rapportage de deelscores en de Totaalscore nogmaals ingeperkt, ditmaal tot een bereik van 10 tot 80. Deze inperking vindt plaats omdat de betrouwbaarheid buiten dit interval snel afneemt.

2.10.1 Training van de componenten van het automatische scoringssysteem

Deze paragraaf beschrijft de training van elke getrainde component in Figuur 2.5. Om het overzicht van de beschrijving te bewaren worden vaak de Engelse termen uit het model in de beschrijvende tekst gebruikt. Benoemde componenten worden weer cursief weergegeven.

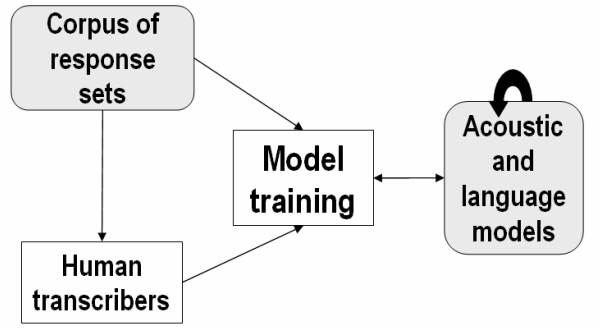
Figuur 2.6 illustreert het trainingsproces voor de modellen voor *Acoustic and language models*. In de figuren toont een dikke gebogen pijl steeds de component die in die specifieke figuur wordt getraind. De trainingsdata worden gevormd door responsen van proefpersonen (*Corpus van response sets).*

De totale verzameling spraak die tijdens de pretests is verzameld, bestaat uit 132.000 uitingen. Hiervan zijn 59.000 uitingen afkomstig van moedertaalsprekers (hierna 'MS') en 74.000 uitingen van niet moedertaalsprekers (hierna 'NMS'). Zowel de groep moedertaalsprekers als de groep NT2-leerders is erg divers. De NMS zijn bijvoorbeeld afkomstig uit 121 verschillende landen en de moedertaalsprekers komen uit verschillende delen van het land. Een gedetailleerde beschrijving van de achtergrondkenmerken van deze twee groepen en een gedetailleerde beschrijving van de pretest is te vinden in Hoofdstuk 3.

Om de *Acoustic and language models* en de overige getrainde componenten van het automatische scoringssysteem te trainen en te testen worden verschillende subsets gebruikt. In de Figuren 2.6 tot en met 2.10 wordt de subset op basis waarvan getraind wordt steeds aangeduid als '*Corpus of response sets'*. Naast dit *Corpus of response sets* ten behoeve van de training werd één subset van 139 NT2-leerders doelbewust opzij gezet en niet betrokken bij de ontwikkeling van het scoringssysteem van de toets. Deze subset werd tijdens de valideringfase gebruikt om de component 'spraakherkenning' binnen het scoringssysteem onafhankelijk te kunnen evalueren. Dit wordt verder beschreven in Hoofdstuk 6. Het *Corpus of response sets* en de afzonderlijke subset ten behoeve van de validering werden willekeurig geselecteerd uit de totale hoeveelheid data die tijdens de pretest waren verzameld.

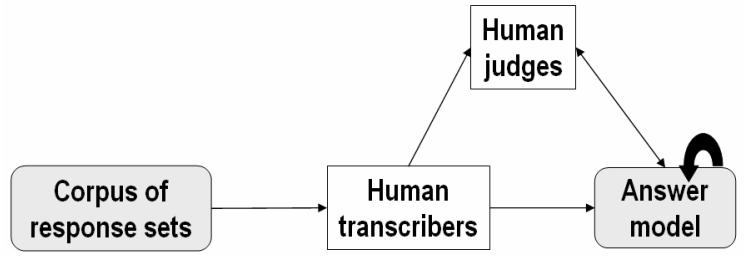
*Human transcribers (*transcribeurs) luisterden naar elke uiting in het *Corpus of response sets* en transcribeerden woord voor woord wat de kandidaat had gezegd. Transcribeurs waren speciaal voor dit doel getrainde moedertaalsprekers van het Nederlands. Aspirant transcribeurs volgden Ordinate's standaard selectie- en trainingprocedures, zoals ook gebruikt voor de ontwikkeling van Ordinate's tests voor Engels en Spaans, en werkten met specifiek voor dit doel ontwikkelde software. Na de training van de transcribeurs werden er op grond van de kwaliteit van hun werk twaalf geselecteerd voor de uitvoering van de transcripties.

De twee soorten input, de audiobestanden en de transcripties van deze audiobestanden, vormen de twee belangrijkste soorten input voor het *Model training* algoritme. Een derde soort input voor het *Model training* algoritme zijn de initiële modellen die moeten worden getraind. Training van de akoestische modellen is een iteratief proces. Bij elke iteratie worden de modellen beter. Het iteratieve proces gaat zolang door tot een maximaal resultaat is bereikt of tot de laatst behaalde verbetering onder een drempelwaarde valt. Figuur 2.6 geeft het trainen van de modellen weer.



Figuur 2.6: Training van Acoustic and language models

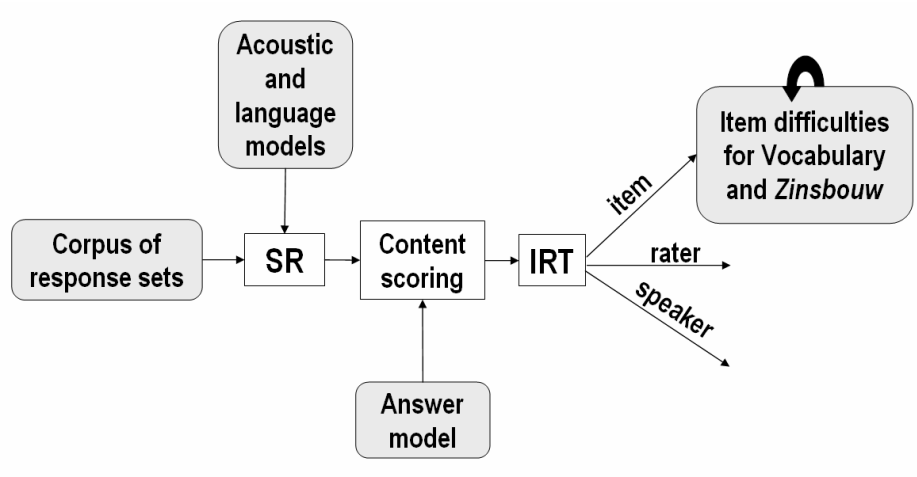
Zoals hierboven beschreven, transcribeerden menselijke transcribeurs het *Corpus of response sets*. De transcripties van de antwoorden op de Korte vragen en Tegenstellingen werden vervolgens beoordeeld door menselijke beoordelaars, *Human judges.* Zij bepaalden of het gegeven antwoord correct of incorrect was. Voor Herhalingen is het correcte antwoord uiteraard de letterlijke tekst van de opgave.



Figuur 2.7: Training van Answer Model

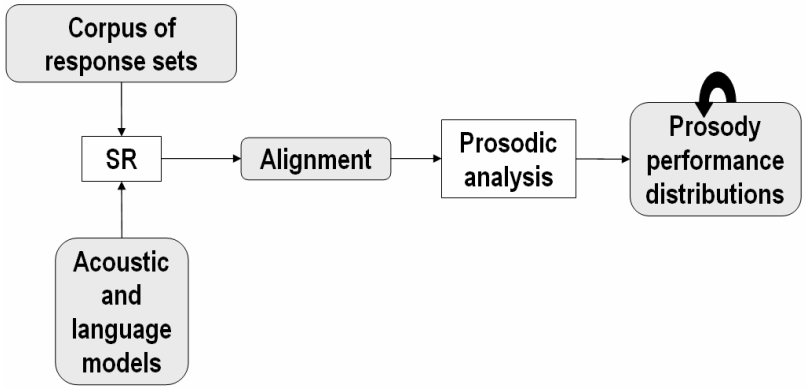
Wanneer het *Answer Model* en de *Acoustic and language models* eenmaal zijn getraind, wordt het *Corpus of response sets* herkend door de spraakherkenner (*SR*). De *SR* produceert de meest waarschijnlijke lexicale string voor een bepaalde response (een bepaald antwoord). Het *Content Scoring* algoritme vergelijkt deze lexicale string dan met de lexicale strings in het *Answer Model* voor dat item en produceert een discrete inhoudscore. Deze discrete inhoudscores worden berekend voor alle antwoorden op de Herhalingen, Korte vragen en Tegenstellingen in het *Corpus of response sets* en vormen de input voor de Item Response Theorie (*IRT)* component. De IRT berekent *Item difficulties* (item parameters) op grond van een 1-parameter Rasch Model. De item-parameters zijn nodig voor de bepaling van een score die onafhankelijk is van de aan de kandidaat gepresenteerde verzameling opgaven. Kandidaten krijgen in een toets immers een willekeurig uit de itembank getrokken set opgaven. Opgaven verschillen onderling in moeilijkheidsgraad. Wanneer een kandidaat toevallig een set moeilijke opgaven krijgt, zou hij minder kans op een voldoende score hebben dan een andere kandidaat die toevallig een set makkelijke opgaven krijgt. De itemparameters bevatten informatie over de onderlinge verschillen in moeilijkheid tussen de opgaven. Op deze wijze kan het scoringsmodel ‘rekening houden’ met de moeilijkheid van de door de kandidaat beantwoorde set opgaven.

De schatting van de IRT parameters voor Woordenschat en Zinsbouw wordt geïllustreerd in Figuur 2.8. De geschatte parameters worden gebruikt voor het bepalen van de test onafhankelijke scores voor Woordenschat en Zinsbouw.



Figuur 2.8: Schatting IRT Parameters voor Woordenschat en Zinsbouw

Figuur 2.9 laat zien hoe de component *Prosody performance distributions* (prosodie prestatie distributies) getraind wordt. Spraakherkenning levert een a*lignment* op. Deze *alignment* produceert temporele informatie met betrekking tot het begin en het einde van elk woord, elke pauze, elke lettergreep en elk foneem. Deze data ondergaan dan *Prosodic analysis* om prosodische kenmerken zoals segmentduur te produceren. Uit de data wordt een schatting van de distributie van elk prosodisch kenmerk afgeleid. Deze *Prosody performance distributions* worden gebruikt om Vloeiendheid te scoren.



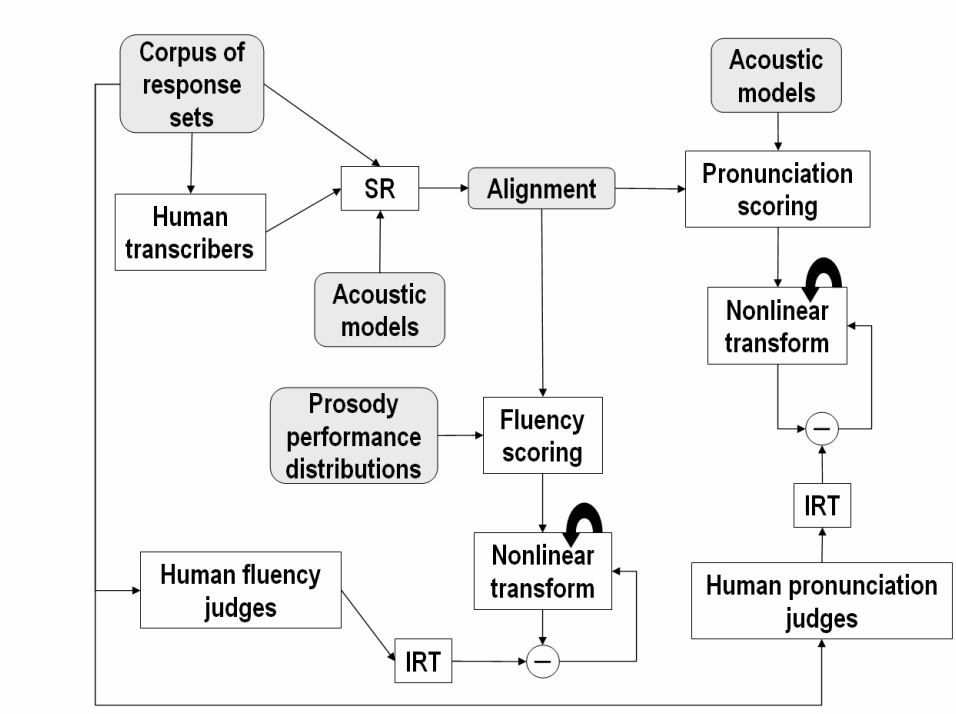
Figuur 2.9: Training van Prosody Performance Distributions

Wanneer de *Prosody Performance Distributions* eenmaal geschat zijn, kunnen de parameters van de twee non-lineaire transformaties (één voor vloeiendheid en één voor uitspraak) worden geoptimaliseerd.

Figuur 2.10 illustreert de training van de non-lineaire transformaties *(Nonlinear transforms)*. De non-lineaire transformaties zetten een multidimensionele input om naar een geschaalde output die een schatting vormt van menselijke beoordelingen van respectievelijk Uitspraak en Vloeiendheid. De bij de training gebruikte menselijke beoordelingen werden gegeven door een groep van zes getrainde onafhankelijke beoordelaars voor uitspraak *(Human pronunciation judges)* en vloeiendheid *(Human fluency judges)*.

Deze beoordelaars beoordeelden de uitspraak en de vloeiendheid van de reacties op de Herhaalopdrachten van een subset van niet moedertaalsprekers (95%) en moedertaalsprekers (5%). Dit type subset (een mix van NMS en MS uitingen) maakt deel uit van Ordinate's standaard procedure voor het ontwikkelen van de deelscores voor uitspraak en vloeiendheid. De beoordelaars hanteerden daarbij een set consistente beoordelingscriteria. Deze zijn gebaseerd op de criteria die Ordinate heeft ontwikkeld voor het Engels en voor het Spaans, maar zijn aangepast aan het Nederlands (zie Bijlage 4 en 5). Beoordelaars maakten gebruik van Ordinate's telefonisch beoordelingssysteem. De beoordelaars horen eerst een opgave uit het onderdeel Zinnen herhalen en krijgen vervolgens, in willekeurige volgorde, antwoorden gepresenteerd van een serie kandidaten op deze stimulus. De beoordelaars toetsen na ieder antwoord hun score in op de telefoon. Deze menselijke oordelen werden geschaald met een één-parameter *IRT* model. Deze geschaalde scores worden ingezet als targets bij de training van de non-lineaire transformaties (*Nonlinear transforms)*.

Net als bij het trainen van de *Acoustic and language models* is de training van de non-lineaire transformaties een iteratief proces. Bij elke iteratie worden de parameters van de non-lineaire transformaties gewijzigd teneinde de output van de transformatie de geschaalde menselijke beoordelingen *(IRT)* dichter te laten benaderen. Tweederde van de training data (6.975 uitingen) werden gebruikt om de *Nonlinear transform* te trainen. Het overblijvende deel van de data (3.139 uitingen) werd gebruikt als stopcriterium. Overfitting kan namelijk een probleem zijn. Wanneer de fit voor de 3.139 uitingen die niet gebruikt werden om de *Nonlinear transform* te trainen begint te verminderen, wordt de procedure gestopt. Gewoonlijk zijn er tussen 150 en 250 iteraties nodig voordat het stopcriterium bereikt is. Het hele proces wordt dan opnieuw gestart met nieuwe willekeurige beginwaarden om na te gaan of een beter lokaal optimum kan worden bereikt. Hiervoor waren 200 runs van elk 150 tot 250 iteraties nodig. De parameters die resulteren in de beste fit worden geselecteerd en worden gebruikt in het runtime scoringssysteem.



Figuur 2.10: Training van de Non-lineaire transformaties voor uitspraak en vloeiendheid

2.10.2 Criterium voor de kwaliteit van het scoringssysteem

Het automatische scoringssysteem dat hier is beschreven, is een complex systeem dat bestaat uit meer dan twaalf componenten, waarvan sommige zelf weer bestaan uit complexe subsystemen. Acht componenten zijn getraind, dat wil zeggen, zij gebruiken parameters die geleerd zijn van data tijdens een training, dat is de periode waarin het systeem nog in ontwikkeling is. Aangezien de componenten statistisch van aard zijn en aangezien zij getraind worden aan de hand van stochastische optimalisatietechnieken, zijn componenten later in de trainingprocedure soms in staat te compenseren voor onzekerheid in de output van eerdere componenten van het systeem. Uiteindelijk moet het scoringssysteem de 45 te scoren antwoorden van een kandidaat verwerken en komen tot een eindscore. Bij het evalueren van het automatische scoringssysteem is het criterium voor de kwaliteit van het systeem de deugdelijkheid van dit ene cijfer: de eindscore. De evaluatie van het scoringssysteem wordt beschreven in hoofdstuk 6.

3 De pretest

Met de pretest werden vier hoofddoelen nagestreefd, te weten:
1. Ontwikkeling van de taalspecifieke componenten van het scoringssysteem.
2. Toets- en itemanalyse.
3. Schaling van de itembank.
4. Validering van de toetsscores.

In dit hoofdstuk worden alleen resultaten met betrekking tot de eerste twee doelen gepresenteerd. Voor de schaling en validering gaven de resultaten van de pretest aanleiding tot vervolgonderzoek. Schaling en validering worden daarom na de presentatie van dit vervolgonderzoek (Hoofdstuk 4) besproken in respectievelijk Hoofdstuk 5 en Hoofdstuk 6.

3.1 Materiaal voor de pretest

Na de in hoofdstuk 2 geschetste ontwikkeling en revisie van items (zie paragraaf 2.7) waren er in totaal 2131 items beschikbaar die aan alle gestelde eisen voldeden. Omdat wij er niet zeker van waren dat het mogelijk zou blijken om in de periode die daarvoor beschikbaar was, maart en april 2004, voldoende proefpersonen te werven om alle ontworpen opgaven te pretesten, werd besloten de pretest te 'faseren' en in een eerste fase slechts de helft van de items op te nemen. Wanneer gedurende het verloop van de pretest zou blijken dat er voldoende proefpersonen beschikbaar waren om ook de andere helft van de ontwikkelde items te testen, dan zouden de betreffende items alsnog worden klaargemaakt voor pretesting. Dat is echter niet mogelijk gebleken.

Er werd gestart met het pretesten van ruim 50% van de opgaven. Uit de gehele voorraad werden in totaal via een gestratificeerde willekeurige selectie 1328 items geselecteerd. Uit dit voor de eerste fase van de pretest geselecteerde deel van de itemvoorraad werden toetsen samengesteld door willekeurige naar itemtype gestratificeerde selecties, waarmee in principe iedere toets in de pretest een unieke deelverzameling items bevatte. Voor de niet-moedertaalsprekers (NMS) werden de pretesten op gelijke wijze samengesteld als toetsen in de beoogde definitieve vorm. NMS kregen in totaal 48 verschillende items voorgelegd. Moedertaalsprekers (MS) kregen 72 verschillende items voorgelegd.

Tabel 3.1 geeft een overzicht van het totale aantal geconstrueerde items, het aantal voor de pretest geselecteerde items en het per test voor MS en voor NMS opgenomen aantal items. De aantallen zijn per soort opgave weergegeven.

Tabel 3.1: Overzicht van aantallen items (in de pretest)

Itemsoort	Totaal geconstrueerd	In pretest-verzameling	Per toets NMS	Per toets MS
Herhalingen	1102	691	24	36
Korte vragen	605	366	14	21
Tegenstellingen	424	271	10	15
Totaal	2131	1328	48	72

Daarnaast bevatte elke toets twee opgaven (Verhalen navertellen) ten behoeve van de validering (zie paragraaf 2.6).

3.2 Procedure van de pretest

De pretesten werden via vaste telefoonlijnen voorgelegd aan moedertaalsprekers en niet moedertaalsprekers van het Nederlands. Gelet op het verwachte aantal deelnemers (2000 NMS en 1000 MS) zouden bij de hierboven genoemde aantallen items per toets 2000*48 = 96.000 responsen van NMS worden verzameld en 1000*72= 72.000 responsen van MS. Gemiddeld zouden daarmee voor ieder van de 1328 items ruim 70 responsen van NMS en ruim 50 van MS worden verzameld.

Ten behoeve van de onafhankelijkheid van de validering werden verschillende maatregelen genomen.
Om te bewerkstelligen dat de validering van de toetsscores (doelstelling 4) onafhankelijk van de ontwikkeling van het scoringssysteem (doelstelling 1) zou plaatsvinden, werden ten behoeve van de analyses uit alle verzamelde data van NMS de reacties van 139 willekeurig gekozen proefpersonen apart gehouden. De reacties van deze 139 proefpersonen werden niet bij de ontwikkeling van specifieke componenten van het scoringssysteem betrokken. Na vaststelling van de procedures voor scoring werd de dataset die apart was gehouden gebruikt om een onafhankelijke analyse van de prestaties van de spraakherkenner te kunnen maken (zie hoofdstuk 6).

Om de onafhankelijkheid van doelstelling 4 van doelstelling 2 te bewerkstelligen, kregen alle proefpersonen aan het einde van de toets twee extra opgaven die niet machinaal werden gescoord en ook niet bij de toetsscore werden betrokken. Het betreft de opgaven ‘Verhalen navertellen’ (zie paragraaf 2.5.4).

De instructies werden tijdens de pretest en de daarop volgende experimenten steeds schriftelijk gegeven. Er waren instructies beschikbaar voor de proefpersonen in het Nederlands, het Turks, Marokkaans-Arabisch en het Engels (zie bijlage 6). De begeleidende docenten ontvingen eveneens een schriftelijke instructie (zie bijlage 7). De instructies voor de proefpersonen werden steeds mondeling toegelicht door een examenleider, meestal een docent die bij de proefpersoon bekend was. De mondelinge instructies werden in bijna alle gevallen in het Nederlands gegeven. De examenleiders die betrokken waren bij de dataverzameling in Nederland rapporteerden dat vooral de instructie aan anderstalige cursisten met zeer weinig onderwijservaring, meer aandacht en tijd vroeg dan CINOP vooraf had ingeschat. In sommige gevallen vonden scholen het nodig om docenten en/of onderwijsassistenten in te schakelen om proefpersonen te ondersteunen bij het kiezen van het telefoonnummer en het intikken van de TIN-code.

Wat betreft de onbekendheid met de toetsvorm en de examencondities verschilt de pretest van wat tijdens de afnames in het kader van inburgeringsexamens het geval zal zijn. De proefpersonen die mee hebben gewerkt aan de pretesten, hebben geen oefentoetsen afgelegd. De toekomstige examenkandidaten zullen de gelegenheid krijgen om vooraf te oefenen. Bovendien krijgen kandidaten in het buitenland instructies in hun eigen taal of in elk geval in een taal die zij zeggen voldoende te beheersen om de instructies te kunnen verstaan. Proefpersonen die betrokken waren bij de aanvullende onderzoeken (zie hoofdstuk 4) hebben allemaal de gelegenheid gekregen een oefentoets af te leggen. Het afleggen van een oefentoets heeft een gunstig effect op daarna behaalde toetsscores en blijkt nuttig om moeilijkheden tijdens latere afnamen te voorkomen.

Examenleiders hebben naar aanleiding van de pretest geen gevallen gemeld van MS die behoefte hadden aan méér instructie dan er schriftelijk werd gegeven. MS hebben de pretesten op verschillende locaties afgelegd: thuis, op school, op het werk. De instructie luidde dat men moest kiezen voor een rustige omgeving, zonder achtergrondgeluiden of andere storende omstandigheden, waarin men met behulp van een vaste telefoonlijn de pretest kon afleggen.

3.3 Steekproeven voor de pretest

3.3.1 Werving van proefpersonen

Bij het samenstellen van de steekproef ten behoeve van het pretesten van de items moest met een aantal punten rekening worden gehouden.

In de eerste plaats was vanaf de start van het project duidelijk dat de pretests in principe in het buitenland zouden moeten plaatsvinden om een steekproef te kunnen samenstellen die 100% representatief zou zijn voor de doelgroep van de toets in het buitenland: leerders van het Nederlands als Vreemde Taal die het inburgeringsexamen buitenland afleggen als verplicht onderdeel van de procedure voor het aanvragen van een Machtiging tot Voorlopig Verblijf in Nederland en die daarvoor examengeld moeten betalen. Mede gezien het grote aantal proefpersonen dat nodig was om het benodigde aantal opgaven te pretesten, bleek dat niet haalbaar. Zelfs via het uitgebreide netwerk van de Nederlandse Taalunie bleek het niet mogelijk in de beschikbare tijd voldoende proefpersonen voor de pretest te benaderen en - bij bereidheid tot medewerking - te toetsen. De meeste leerders van het Nederlands als Vreemde Taal die via het netwerk van de Taalunie bereikt kunnen worden, zijn bovendien hoog opgeleid en daarmee niet representatief voor de hele doelgroep van het inburgeringsexamen buitenland.

Er is daarom besloten de pretesten in Nederland te laten plaatsvinden, bij leerders van het Nederlands als tweede taal. Om de grote aantallen proefpersonen te bereiken die nodig waren, zijn Regionale Opleidingen Centra (ROC) verspreid over het land benaderd. Bij het samenstellen van de offerte werd voorzien dat er in het onderwijsveld weinig animo zou zijn om mee te werken aan een toets die onderdeel uitmaakt van de door de overheid beoogde vernieuwing van het inburgeringbeleid, dat op dat moment door veel docenten, schoolleiders en inburgeraars als onverstandig en bedreigend werd ervaren. CINOP heeft zich daarom al vóór het indienen van haar offerte verzekerd van voldoende steun in het onderwijsveld. CINOP heeft haar plannen, alvorens ze bij de opdrachtgever te offreren, gepresenteerd en voorgelegd aan de leden van de Bedrijfstakgroep Educatie, het op dat moment bestaande samenwerkingsverband van afdelingen Educatie van ROC's. De Bedrijfstakgroep reageerde positief en verklaarde zich in principe bereid medewerking te verlenen.
In de eerste maanden van 2004 zijn daarop door CINOP in totaal twaalf ROC's benaderd met het verzoek mee te werken aan de pretest. Op elk ROC werd een bijeenkomst verzorgd voor managers en docenten educatie, waarin de achtergronden van het project, de kenmerken en doelen van de te ontwikkelen toets en de opzet van de pretest werden gepresenteerd en besproken. Bovendien werden de tegenprestaties beschreven die CINOP zou leveren, te weten:

- elke pretestproefpersoon zou per afgelegde pretest (mondelinge én schriftelijke vaardigheden) een cadeaubon van 10 Euro ontvangen (in totaal 20 Euro);
- elke team van docenten zou van CINOP een training van vijf dagdelen aangeboden krijgen over achtergronden, kenmerken en gebruiksmogelijkheden van het CEF, assessment van taalvaardigheid en het beoordelen van taalvaardigheid aan de hand van het CEF.

Van de twaalf ROC's die benaderd zijn, hebben tien ROC's hun medewerking toegezegd. Twee hebben geweigerd mee te werken. Binnen de meewerkende ROC's zijn vervolgens intern afspraken gemaakt waarbij is vastgelegd dat docenten die principiële bezwaren hadden tegen deze toets, hun medewerking konden weigeren. De docenten die wel wilden meewerken, hebben vervolgens aan de hand van door CINOP verstrekte materialen hun cursisten geïnformeerd over de doelen en achtergronden van de toets waarvoor hun medewerking werd gevraagd. Het is niet bekend hoeveel docenten en potentiële proefpersonen hun medewerking hebben geweigerd op basis van argumenten die betrekking hebben op het beleid in het kader waarvan de toets werd ontwikkeld.

CINOP heeft elke school die bereid was aan het project deel te nemen een set pretestmaterialen bezorgd die uitging van een maximale respons. Elke school kreeg voor elke NT2-docent instructies en materialen ten behoeve van de afname van de pretest in elk van zijn/haar lesgroepen, waarbij de gemiddelde omvang van een lesgroep werd geschat op 25 proefpersonen.
Per docent bevatte het materialenpakket de volgende documenten: een tekst met achtergrondinformatie voor de docent (bijlage 8), een tekst met achtergrondinformatie voor de cursisten (25 exemplaren) (bijlage 9), een tekst met instructies voor de cursisten (25 exemplaren) (bijlage 6), een tekst met instructies voor de docent (bijlage 7) en een vragenlijstje met betrekking tot achtergronden van de cursisten (25 exemplaren) (bijlage 10).
Het is duidelijk dat daarmee véél meer pretestmaterialen werden verspreid dan er proefpersonen bereikt zouden worden: docenten konden weigeren mee te werken, proefpersonen konden weigeren, de omvang van lesgroepen in het volwassenenonderwijs loopt zeer uiteen en ook het aantal aanwezige cursisten kan wisselen per dag.
Daarnaast kreeg elke school een set van 75 pretestformulieren voor moedertaalsprekers. Deze formulieren werden door docenten verspreid onder autochtone studenten van de afdeling Educatie en van beroepsopleidingen in het betreffende ROC en een aantal collega's. In het kader van het streven naar een representatieve steekproef van MS, is in elk ROC nadrukkelijk gevraagd om ook autochtone deelnemers aan alfabetiseringscursussen te benaderen voor deelname aan de pretest.
De steekproef die op deze wijze tot stand is gekomen via de medewerking van ROC's is aangevuld met een aantal proefpersonen van elders om te zorgen voor voldoende representatie van NMS met hoge taalvaardigheidniveaus in het Nederlands, en voor voldoende spreiding van achtergrondkenmerken binnen de groep MS. Naast ROC's hebben in dat kader een particulier taleninstituut en een universitair talencentrum aan de pretesten meegewerkt.

3.3.2 Kenmerken van de steekproeven

Aan alle deelnemers werd een aantal achtergrondgegevens gevraagd met betrekking tot geslacht, geboortejaar, opleiding en dergelijke. De NMS werd ook gevraagd naar het land van geboorte, het aantal jaren van verblijf in Nederland, het aantal uren NT2, wel of niet gealfabetiseerd, recent behaalde scores op andere NT2 toetsen en de inschatting van het CEF-niveau door de docent voor de vaardigheden spreken en lezen (lezen vanwege de ontwikkeling van de toets geletterdheid die later werd stopgezet). De MS werd gevraagd naar de plaats van geboorte en de thuis gesproken taal (Nederlands/dialect/ andere taal). De NMS werkten over het algemeen onder begeleiding van docenten, deze vulden veelal voor hen ook de achtergrondgegevens in. De MS werkten zelfstandig en vulden derhalve zelf hun achtergrondgegevens in.

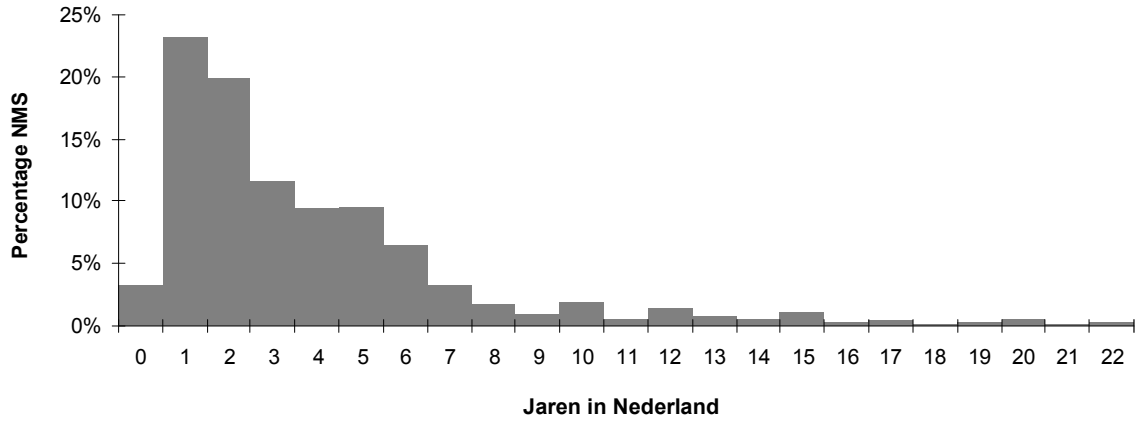
Helaas vulden niet alle docenten voor iedere deelnemer de gevraagde achtergrondgegevens in. De omvang van de verzamelde set achtergronddata verschilt daardoor per gegeven. Van de gegevens die in principe voor alle cursisten konden worden geleverd zijn de meeste gegevens beschikbaar voor geslacht (n=1574) en de minste voor het door de docent ingeschatte CEF-niveau voor leesvaardigheid (n=1226). Uiteraard namen slechts beperkte aantallen cursisten deel aan NT2 examens of toetsen in de gevraagde periode van januari tot en met maart en zijn de aantallen daarvoor nog veel lager dan het aantal verzamelde gegevens over het leesvaardigheidsniveau. In het navolgende zal daarom per achtergrondgegeven het aantal worden vermeld waarop het gegeven betrekking heeft. Tabel 3.2 geeft een overzicht van de aantallen deelnemers en de over hen verzamelde data. In het navolgende zullen we in principe uitsluitend achtergrondgegevens vermelden van personen waarover geldige testdata beschikbaar zijn.

Tabel 3.2: overzicht aantallen deelnemers en verzamelde data

Populatie	Geldige testdata	Achtergronddata (Max voor m/v)	Overlap testdata en achtergronddata
MS	821	813	768
NMS	1522	1574	1341

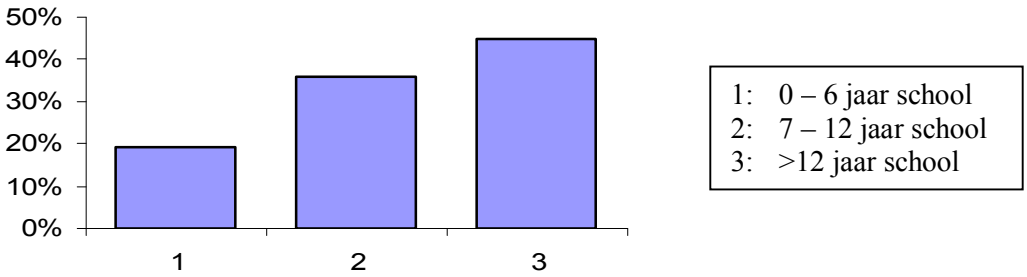
3.3.3 Enkele achtergrondgegevens van NMS

De gemiddelde leeftijd van NMS-deelnemers was 31 jaar (± 10.2, n= 1325) en liep uiteen van 8 tot 71. De verdeling man: vrouw was 36:64 (n=1341). De NMS deelnemers waren afkomstig uit 121 verschillende landen. Ongeveer 50% verbleef twee jaar of minder in Nederland (n=1222). De overigen waren verdeeld over een groot aantal verschillende jaren van verblijf. Minder dan 2.5% verbleef al langer dan 22 jaar in Nederland. De gemiddelde verblijfsduur in Nederland van de NMS was 4.5 jaar. Figuur 3.1 toont de verdeling van de NMS pretestdeelnemers tot maximaal 22 jaren verblijf in Nederland (97.5% van de NMS deelnemers).



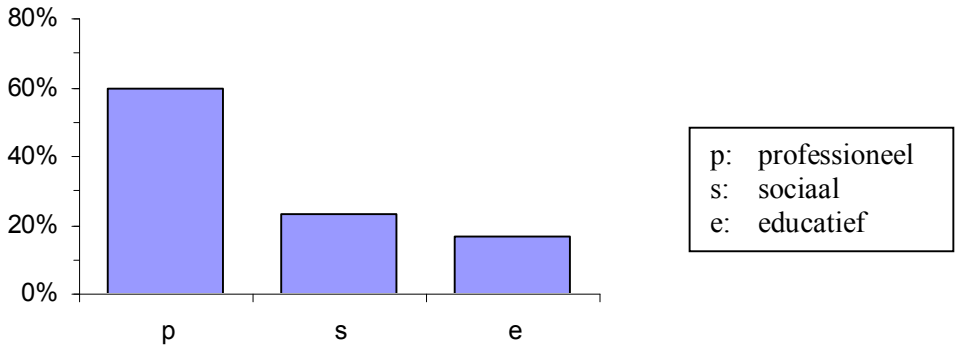
Figuur 3.1: Verdeling NMS naar jaren verblijf in Nederland (n=1222).

Circa 19% van NMS deelnemers (n=1353) had niet meer dan lagere school. De overigen hadden een opleiding op middelbaar niveau (36%) of hoger (45%). Figuur 3.2 toont de verdeling van de NMS deelnemers naar opleidingsniveau. Onafhankelijk van het opleidingsniveau was de verdeling naar plaatsing in een alfabetiseringtraject "nee" : "ja" als 92% : 8% (n=1117**).**



Figuur 3.2: Opleidingsniveau NMS deelnemers pretesten (*n=1353)*

Meer dan de helft (60%) van de deelnemers waarvan opgave (n=1097) heeft een professioneel uitstroomperspectief, de overigen hebben een sociaal (23%) of educatief (17%) uitstroomperspectief. Figuur 3.3 toont de verdeling van de NMS deelnemers naar uitstroomperspectief.



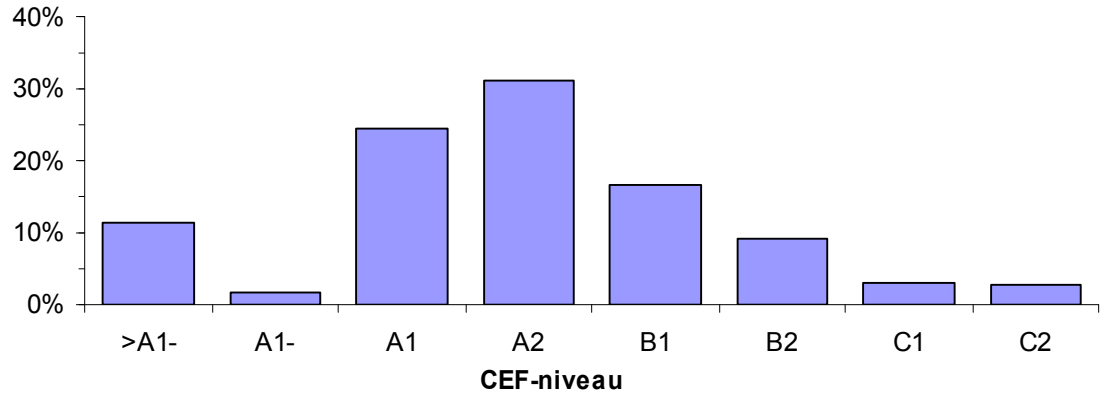
Figuur 3.3: Uitstroomperspectief NMS deelnemers pretesten (n=1097)

Docenten werd ook gevraagd naar beschikbare gegevens voor de NMS deelnemers met betrekking tot uitslagen van recent afgenomen toetsen en examens. Tabel 3.3 toont een overzicht van de gerapporteerde gegevens. Per toets is aangegeven voor welke deelvaardigheden het hoogste en het laagste aantal deelnemers werd gerapporteerd. Deze gegevens bleken echter niet bruikbaar doordat docenten zeer verschillende methoden hanteren om de toetsuitslagen te rapporteren en in veel gevallen niet kon worden achterhaald waar de notatie van de docent aan refereerde.

Tabel 3.3: Opgegeven aantallen toetsuitslagen voor NMS pretestdeelnemers

Examen / Toets	**Hoogste aantal (deelvaardigheid)**		**Laagste aantal (deelvaardigheid)**	
Staatsexamen NT2-I	15	(lezen)	13	(spreken)
Staatsexamen NT2-II	22	(lezen)	17	(schrijven)
Profieltoets	85	(schrijven)	62	(spreken)
NIVOR	477	(luisteren)	411	(spreken)
Trajecttoetsen	390	(schrijven)	351	(spreken)
Itemdito	2	(schrijven)	1	(lezen, spreken)
NT2-CAT	15	(luisteren)	12	(lezen)

Tenslotte werd docenten ook gevraagd een inschatting te geven van het niveau van de deelnemende proefpersonen uitgedrukt in een niveau op de schaal van de Raad van Europa, de zogenaamde CEF-niveaus. Figuur 3.4 toont de verdeling over de niveaus zoals door de eigen docenten van de deelnemers geschat.

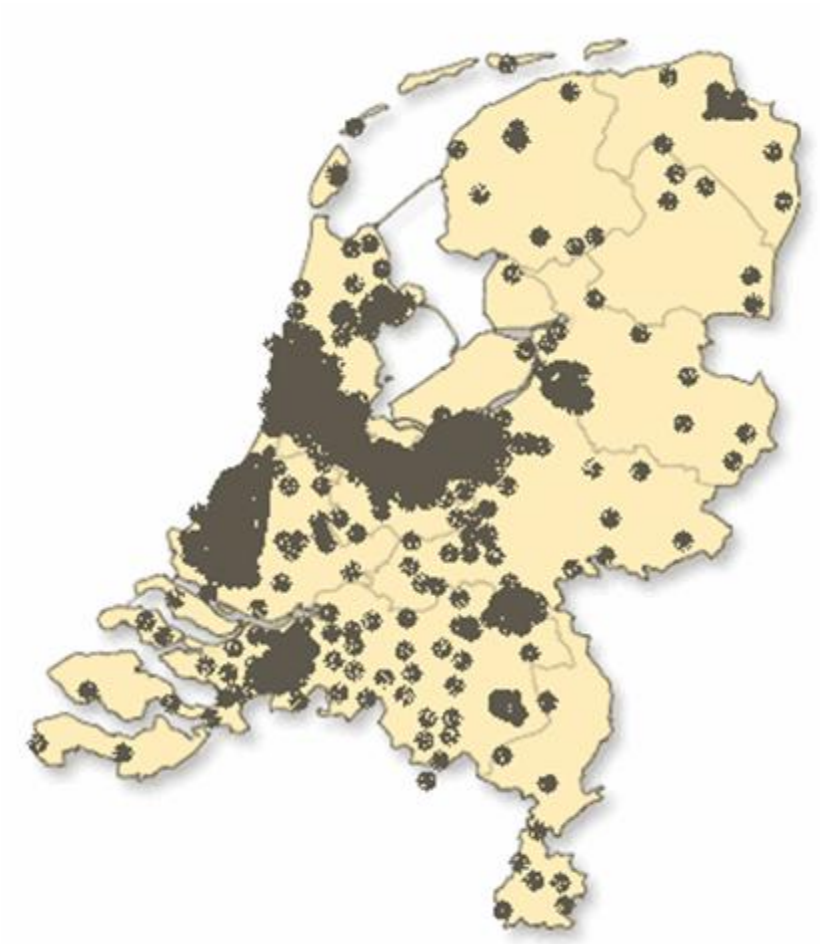


Figuur 3.4: Schatting door de eigen docenten van de vaardigheid van NMS deelnemers op CEF niveau

De docenten waren allen wel bekend met het CEF, maar hadden niet allemaal een training in beoordeling op CEF-niveaus ontvangen. Er kan daarom enige onzekerheid bestaan over deze door docenten geschatte niveaus. Toch geven zij wel een globale indruk van de verdeling van de vaardigheid van de NMS deelnemers. Ondanks ons streven om vooral veel personen met lage taalvaardigheidsniveaus in het Nederlands in de steekproef op te nemen, bleek bij de evaluatie van de enquêteformulieren na afloop van de pretest dat - althans naar de mening van de docenten - er slechts een zeer geringe vertegenwoordiging was van personen op het A1-min niveau. Met name vanwege het beperkte aantal getrainde beoordelaars onder deze docenten kan niet worden uitgesloten dat dit deels wordt veroorzaakt door onbekendheid bij de docenten met het A1-min niveau. Dit niveau is immers een toevoeging van de Commissie Franssen aan het CEF. Voor deze interpretatie spreekt ook de vorm van de verdeling in Figuur 3.4: het is niet echt aannemelijk dat toevalligerwijs juist dit niveau veel minder is vertegenwoordigd dan de aan weerszijden aanliggende niveaus. Wij komen in Hoofdstuk 6 terug op deze beoordeling.

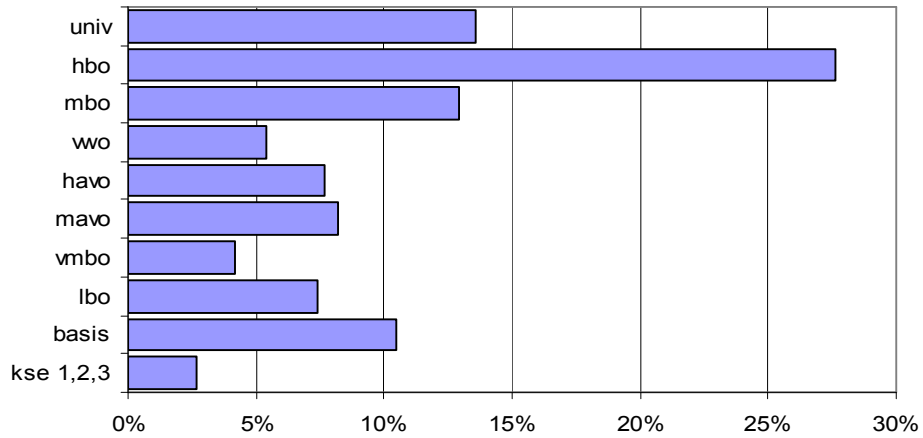
3.3.4 Enkele achtergrondgegevens van MS

De verhouding man:vrouw voor de MS was 37:63 (n=768) en de gemiddelde leeftijd bedroeg 37 jaar (± 16, n=761). MS gaven 221 (n=760) verschillende steden en dorpen verspreid over Nederland als geboorteplaats op. Daarbij waren 131 plaatsen waaruit slechts één MS afkomstig was en vier steden met meer dan dertig vertegenwoordigers (Alkmaar, Amsterdam, Delft, Den Haag). Het grootste aantal uit één stad was afkomstig uit Den Haag (47). Figuur 3.5 brengt de spreiding van de geboorteplaatsen van de MS deelnemers in beeld. Figuur 3.6 toont de opleidingsniveaus als opgegeven door de MS.



Figuur 3.5: Geboorteplaats MS deelnemers pretesten (n=760)

De grote vertegenwoordiging van proefpersonen met een hbo-opleiding wordt verklaard door de deelname van docenten.



Figuur 3.6: Opleidingsniveau MS deelnemers (n=745)

3.4 Resultaten

3.4.1 Resultaten van de proefpersonen

In deze paragraaf worden de resultaten voor MS en NMS gepresenteerd. In Hoofdstuk 2 is beschreven hoe op basis van verzamelde reacties van MS en NMS het model voor automatische scoring wordt getraind. In deze paragraaf wordt ingegaan op de analyses van de resultaten van de scoring volgens deze procedures. De evaluatie van de validiteit van het scoringssysteem en de daarmee opgeleverde toetscores wordt besproken in Hoofdstuk 6.

Wij presenteren hier alleen de deelscores omdat de totaalscore pas na de schaling (zie Hoofdstuk 5) wordt vastgesteld. Eerst moet de kwaliteit van de deelscores worden geëvalueerd. De deelscores zijn intervalscores en worden uitgedrukt op een logistische schaal. Bij de uiteindelijke rapportage van de scores aan gebruikers zullen deze waarden door middel van een lineaire regressiefunctie worden omgezet naar een schaal van 10 tot 80 omdat een logistische schaal weinig gebruikersvriendelijk is. Tabel 3.4 presenteert de resultaten naar deelscore voor de NMS.

Tabel 3.4 Resultaten pretest analyses voor NMS (n=1376)

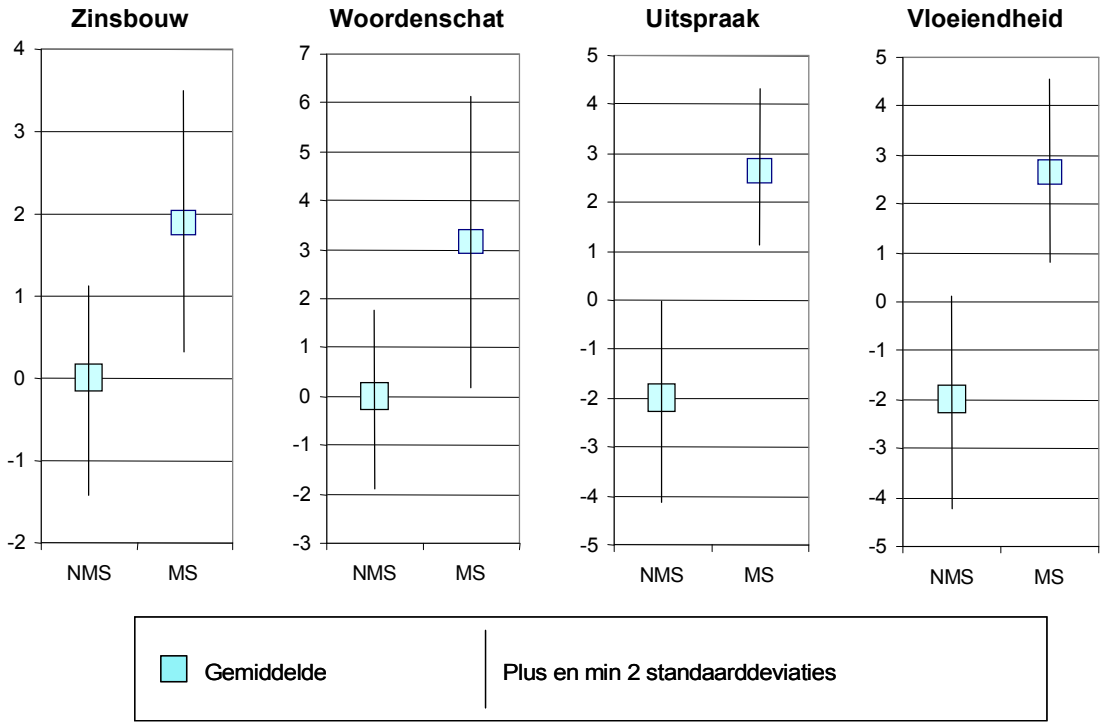
	Zinsbouw	Woordenschat	Uitspraak	Vloeiendheid
Ruwe scores				
Gemiddelde score	64.7	12.2	*	*
Standaarddeviatie score	34.2	4.0	*	*
Gemiddelde p-waarde	0.30	0.56	*	*
Geobserveerde maximumscore	166	29	*	*
Geobserveerde minimum score	1	0	*	*
Vaardigheidsparameters				
Gemiddelde parameter	0.00	0.00	-2.01	-0.12
Standaarddeviatie parameter	0.66	1.21	1.06	1.12
Maximum parameter	3.91	4.45	3.36	4.34
Minimum parameter	-5.13	-6.47	-3.14	-2.28
Gemiddelde meetfout parameter	0.14	0.58	0.35	0.33

* Voor uitspraak en vloeiendheid bestaan geen ruwe scores

In Tabel 3.4 verwijzen de ruwe deelscores voor zinsbouw en woordenschat direct naar het aantal fouten dat proefpersonen gemaakt hebben bij het beantwoorden van de opgaven. Op grond van deze ruwe scores zijn moeilijkheidsparameters voor de items en vaardigheidsparameters voor de personen geschat met gebruikmaking van het computerprogramma FACETS (Linacre, 1988,2005). Voor uitspraak en vloeiendheid worden geen fouten geteld, er is daarom geen ruwe score beschikbaar. De vaardigheidsparameter wordt direct geschat op basis van respectievelijk akoestische en temporele maten als uiteengezet bij de beschrijving van de automatische scoring in paragraaf 2.10). De meetfout van de vaardigheidsparameters is niet over de gehele schaal gelijk. De gemiddelde meetfout in Tabel 3.4 geeft slechts een indicatie van de relatieve onnauwkeurigheid van de deelscores. In Hoofdstuk 6 komen we terug op de meetnauwkeurigheid van de TGN totaalscore op de relevante punten van de scoreschaal (de cesuren).

Zoals gezegd hebben tijdens de pretest naast sprekers van het Nederlands als tweede taal ook moedertaalsprekers van het Nederlands de toets afgelegd. Een eerste eis die men mag stellen aan een toets die beoogt te meten of personen voldoende vaardigheid hebben om deel te nemen aan gesprekken in een voor hen vreemde of tweede taal, is dat de toets daadwerkelijk onderscheid maakt tussen personen die deze vaardigheid beheersen en personen die deze vaardigheid nog niet of in beperkte mate beheersen.

Figuur 3.7 toont per deelscore het gemiddelde van de parameterschattingen voor NMS en voor MS evenals de bijbehorende standaarddeviaties. Afgebeeld zijn twee standaarddeviaties aan weerszijde van het gemiddelde waarmee bij een normaalverdeling 95% van de verdeling is weergegeven. Aangezien de parameters voor ieder van de vier deelscores apart zijn geschat, zijn de schalen nog niet onderling vergelijkbaar. Pas na de schaling (Hoofdstuk 5) worden de schalen vergelijkbaar. Binnen ieder van de vier schalen is wel duidelijk dat de gemiddelde parameterschattingen voor de twee groepen (NMS en MS) zover uit elkaar liggen dat er sprake is van een duidelijk onderscheid.



Figuur 3.7: Parameter schattingen: gemiddelden en twee standaarddeviaties per deelscore voor NMS (n=1376) en MS (n= 821).

3.4.2 Resultaten aangaande de items

3.4.2.1 Initiële p-waarden

Bij de ontwikkeling van de itemtypen 'Kort-antwoordvragen' en 'Tegenstellingen' zijn door de itemconstructeurs mogelijke goede antwoorden gesuggereerd die in de itembank werden opgenomen. Na de verzameling van de reacties in de pretest werd nagegaan in hoeverre antwoorden van MS en NMS overeenstemmen met deze veronderstelde antwoorden.

Een groot deel van de responsen van de proefpersonen werd getranscribeerd door getrainde transcribeurs (zie Hoofdstuk 2). Op basis van de transcripten werden p-waarden berekend als initiële maat voor correctheid. 'Correctheid' werd gedefinieerd als 'Is precies dat gezegd wat er gezegd moest worden en wel alles en niets meer?'. Responsen kregen ofwel een score 0 (er ontbreekt iets) ofwel een score 1 (alles werd gezegd). Een herhaalopdracht werd als 1 gescoord als gold: 'alle woorden zijn herkenbaar gezegd'. Bij de 'Kort- antwoordvragen' en de 'Tegenstellingen': 'Is het antwoord van de proefpersoon gelijk aan één van de door de itemconstructeurs gesuggereerde goede antwoorden?'. Voor de 'Herhaalopdrachten' wordt deze maat later omgezet in een continue score 'hoeveel van wat gezegd had moeten worden is er gezegd?'.

3.4.2.2 Correctie van antwoordmodellen 'Korte vragen' en 'Tegenstellingen'

Voor de 'Korte vragen' en de 'Tegenstellingen' wordt nagegaan of door proefpersonen gegeven antwoorden die afwijken van het antwoordmodel, mogelijk toch als goed antwoord kunnen worden geïnterpreteerd. Op basis van de transcripties van de reacties van de proefpersonen werden de oorspronkelijk aangegeven antwoordmogelijkheden geëvalueerd en, waar nodig, aangevuld. Tabel 3.5 illustreert het effect van de op transcriptie gebaseerde bijstelling van het antwoordmodel. Score1 in de Tabel is gebaseerd op het oorspronkelijke antwoordmodel. Het criterium voor inspectie was dat meer dan 5% van de MS of meer dan 10% van de NMS een bepaalde respons geeft. Indien een respons niet aan die minimum waarden voldeed, maar toch acceptabel bleek, werd deze in de database aan het record van goede antwoorden toegevoegd. Score2 is gebaseerd op het bijgestelde model. (N.B. Niet alle verzamelde reacties werden getranscribeerd. Het totale aantal responsen in de kolom "Total n" verwijst slechts naar het totale aantal transcripties in deze analyse).

Tabel 3.5 Correctie antwoordmodellen naar aanleiding van de pretest. Een voorbeeld uit de pretest

7004.487	**Wat maak je met een fototoestel?**	**0.808**					
Correct AC:	**Foto's**	**Aantal**	**Totaal**	**Score1**	**Score2**	**%**	**Cumulatief**
3.4.2.3 Geobserveerde antwoorden							
	Een foto foto's	1	73	1	1	0.014	0.014
	Een foto's	1	73	1	1	0.014	0.027
	Foto's	55	73	1	1	0.753	0.781
	Foto's mooie	1	73	1	1	0.014	0.795
	Met een fototoestel maak je foto's	1	73	1	1	0.014	0.808
	-(f)oto's	1	73	0	1	0.014	0.822
	Foto maken	1	73	0	1	0.014	0.836
	Een foto	9	73	0	1	0.123	0.959
	Foto	2	73	0	1	0.027	0.986
	[stilte]	1	73	0	0	0.014	1.000

Tabel 3.5 illustreert de gehanteerde werkwijze aan de hand van item 7004.487: "Wat maak je met een fototoestel?". Bij de constructie van dat item werd vastgesteld dat het juiste antwoord moest luiden 'foto's'. Na inspectie van de reacties werden ook antwoorden met het enkelvoud 'foto' goed gerekend.

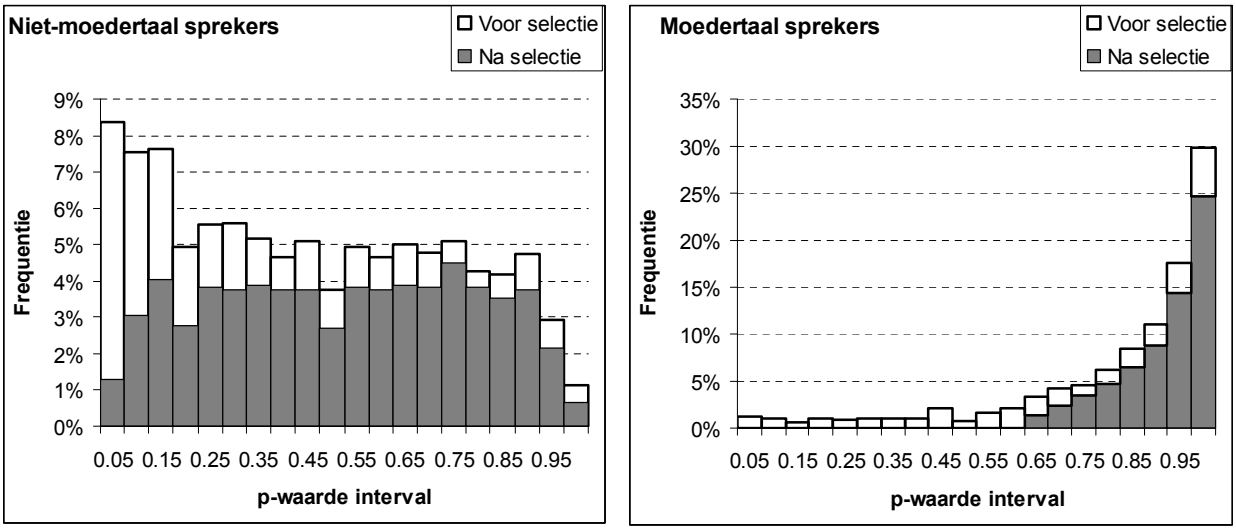
3.4.2.4 Itemselectie en parameterschattingen

Maten voor de mate waarin de opgaven bijdragen aan de meting van één onderliggende variabele zijn itemcorrelaties (in FACETS wordt de punt-biseriële correlatie gerapporteerd) en passing van de data bij het gekozen model, i.c. het Rasch model. Gegeven de omvang van de dataset is toetsing op formele passing op het niveau van de gehele bank altijd significant en is een nader onderzoek naar passing niet relevant. Op het niveau van individuele personen en afzonderlijke items is een onderzoek naar passing echter wel relevant. In de output van het FACETS programma wordt het gekwadrateerde gemiddelde van de modelafwijkingen als maat voor passing gegeven. Deze maat heeft een verwachte waarde van 1. Waarden groter dan 1 duiden op het voorkomen van onverwacht lage scores voor personen met een hoge vaardigheid als gemeten en onverwacht hoge scores voor personen met een lage vaardigheidschatting. Veelal hangt dit samen met matige discriminatie.

Waarden kleiner dan 1 duiden op het omgekeerde: meer goede antwoorden dan verwacht bij hoge vaardigheid en minder bij lage vaardigheid.
Verschillen in discriminatie kunnen onder andere worden veroorzaakt door verschillende vormen van bias, het is daarom zaak om de itemverzameling te zuiveren van items met te grote afwijkingsmaten. In een iteratieve procedure werden items met sterke afwijkingsmaten (fit >1.5) en/of met zeer lage punt-biseriële correlatie (<0.15) verwijderd.

Door de willekeurige toewijzing van items aan personen waren er ook een aantal items waarvoor een te gering aantal observaties was verzameld. Deze items zijn voorlopig op 'non-actief" gesteld en kunnen wanneer later meer data zijn verzameld en wanneer de statistische maten aan de criteria voldoen, weer worden geactiveerd.
Na verwijdering van ca 30% van de items op grond van bovengenoemde criteria was de gemiddelde punt-biseriële correlatie van de voor opname in de itembank geselecteerde items 0.59 (± 0.15), de gemiddelde passing 0.99 (±0.23).

Figuur 3.8 toont de verdeling van de geobserveerde p-waarden vóór en na de selectie. Op de x-as de p-waarden, op de y-as de procentuele frequentie van items met een p-waarde van 'x'.



Figuur 3.8 Frequentie p-waarden Zinsbouw en Woordenschat vóór en na selectie MS en NMS als percentage van pretest items

Vóór de selectie hadden 8% van de items bij de NMS een p-waarde van 0.05 of minder. Ook bij p-waarden van 0.10 en 0.15 werden ongeveer 8% van de items aangetroffen. Bij de overige intervallen lag het percentage tussen de 4% en de 5% behalve bij de hoogste twee, waar een lager percentage werd aangetroffen. Voor de MS had ongeveer 75% van de items vóór de selectie een p-waarde van 0.70 of hoger. Er waren echter ook een groot aantal items met lagere p-waarden. Door het selectieproces is er nog slechts 1% van de items met een lagere p-waarde dan 0.70 bij de MS. Voor de NMS is de verdeling van de p-waarden over de gehele schaal met uitzondering van de uiteinden uniform. De gepresenteerde data maken duidelijk dat met deze verzameling opgaven onder NMS over een brede range kan worden gemeten: er zijn voor ieder niveau passende opgaven. Ook wordt duidelijk dat de verzameling een scherp onderscheid maakt tussen NMS en MS: de opgaven kunnen heel gemakkelijk worden beantwoord door personen die Nederlands spreken.

3.5 Conclusies en onderwerpen voor nader onderzoek

Uit de analyse van de pretestgegevens blijkt dat met de geselecteerde opgaven een goed onderscheid gemaakt kan worden tussen enerzijds moedertaalsprekers van het Nederlands en anderzijds personen die het Nederlands van huis uit niet beheersen.

Dit is een indicatie van validiteit. Voorts blijkt dat met de verzameling opgaven bij NMS over een brede range kan worden gemeten. Gemiddelde punt-biseriële correlaties en gemiddelde passingsmaten vormen indicaties dat een dergelijke meting ook betrouwbare scores zal opleveren.

Er zijn echter ook een aantal problemen gesignaleerd. Ten eerste kan vanwege de willekeurige trekking van opgaven voor de samenstelling van een toets uit de opgavenverzameling geen goede evaluatie plaatsvinden van de homogeniteit en de schaalbaarheid van uit de verzameling te trekken toetsen. Voor dergelijke schattingen is een groot aantal observaties van een vaste set opgaven nodig.

Ten tweede bleek bij de inspectie van de verzamelde data dat de laagste vaardigheidsniveaus minder goed in de pretest steekproef vertegenwoordigd zijn dan was bedoeld. Hierdoor wordt de zekerheid waarmee de schaling kan worden geschat juist aan deze kant van de schaal negatief beïnvloed.

Ten derde bleek na afloop van de pretest dat de Nederlandse ambassades in het buitenland niet allemaal zonder meer over vaste telefoonlijnen kunnen beschikken, maar dat zij voor telefonie gebruik maken van een eigen beschermd netwerk. De specificaties van dit netwerk wijken echter af van de normale telefoonlijnen. Het was niet bekend welke invloed deze afwijking kan hebben op het spraakgeluid en daarmee op de waardering van de responsen van de proefpersoon.

De drie bovenstaande conclusies geven de noodzaak weer van aanvullend onderzoek. Dit onderzoek zal gericht moeten zijn op het verzamelen van nadere gegevens over:

- de invloed van het afwijkende telefoonnet dat gebruikt wordt op Nederlandse ambassades in het buitenland;
- de geschiktheid van de opgaven voor de allerlaagste niveaus van taalvaardigheid;
- de dimensionaliteit van de opgavenverzameling en de schaalbaarheid van de resultaten.

Hoewel op grond van de pretestgegevens een rapportageschaal met verwijzing naar de CEF niveaus is ontwikkeld, kan de schaling en de kwaliteit van de opgeleverde schaal pas definitief worden vastgelegd wanneer de resultaten van deze aanvullende onderzoeken daar in zijn betrokken. In Hoofdstuk 4 behandelen we de opzet en de resultaten van deze onderzoeken.

4 Aanvullende onderzoeken

In dit hoofdstuk bespreken wij de aanvullende experimenten waartoe op basis van de resultaten van de pretest (Hoofdstuk 3) is besloten. Er zijn drie in de tijd opeenvolgende experimenten opgezet:

- *het experiment 'Den Haag*' gericht op het achterhalen van de invloed van het afwijkende telefoonnet dat gebruikt wordt op Nederlandse ambassades in het buitenland;
- *het experiment 'Amsterdam'* gericht op het verzamelen van nadere gegevens over de empirische betrouwbaarheid van de toets door middel van een test-hertest bij proefpersonen op de allerlaagste niveaus van taalvaardigheid volgens het CEF;
- *het experiment 'MFA-Fit'* gericht op het verzamelen van aanvullende data - vooral op de laagste niveaus - over de invloed van het afwijkende telefoonnet van de Nederlandse ambassades in het buitenland (MFA) en van evidentie over de dimensionaliteit en schaalbaarheid van de opgavenverzameling (FIT).

In dit hoofdstuk zetten wij eerst de opzet uiteen van deze aanvullende experimenten. Vervolgens presenteren wij de resultaten en sluiten af met de conclusies.

4.1 Materiaal en methode

4.1.1 Experiment Den Haag

4.1.1.1 Probleemstelling

In de zomer van 2004, nadat de dataverzameling in het kader van de pretest was afgerond, werd duidelijk dat niet alle ambassades en buitenlandse posten van het Ministerie van Buitenlandse Zaken beschikken over een betrouwbare telefoonverbinding via het vaste publieke netwerk (hierna: 'PSTN'). Het Ministerie van Buitenlandse Zaken besloot daarom in overleg met het Ministerie van Justitie om op alle buitenlandse posten en ambassades gebruik te maken van een Voice over Internet Protocol (VoIP) telefoonsysteem via het eigen, beveiligde netwerk (hierna 'MFA-net'). Binnen het MFA-net heeft het Ministerie van Buitenlandse Zaken twee verschillende versies in gebruik:

- MFA-net-T: internet telefonie via landlijnen;
- MFA-net-S: internet telefonie via satellietverbindingen.

Beide versies maken gebruik van compressie/decompressie techniek volgens aanbeveling G.729 van de ITU (Internationale Telecommunicatie Unie, Genève, Zwitserland, maart 1996). Het pakketverlies op het netwerk is volgens de leverancier gegarandeerd niet hoger dan 0.4%. In de praktijk is de prestatie van het netwerk beter (zie ook Rosenberg, 1997). Het Ministerie van Buitenlandse Zaken heeft een service niveau overeenkomst op protocol 3 niveau. De 'Bit error rate' is van een lager protocol niveau (niveau 2) en is daarom niet apart gespecificeerd. Het maakt onderdeel uit van de garantie aangaande het pakketverlies. Naast het pakketverlies treedt bij internet telefonie ook enige vertraging op. Deze is volgens opgave van het Ministerie van Buitenlandse Zaken 200 milliseconden bij landlijnen en bij satellietverbinding 640 milliseconden.

Doel van het experiment is na te gaan of, en zo ja in hoeverre, verschillen tussen MFA-net en PSTN van invloed zijn op de scores van de proefpersonen.

4.1.1.2 Materiaal

Om na te gaan of er een invloed was van het MFA-net op de scores van de proefpersonen werden uit de op basis van de pretestgegevens 'gezuiverde' itembank per proefpersoon vier toetsen samengesteld op de standaardlengte en in de standaardsamenstelling voor het examen. Dat gebeurde door willekeurige naar itemtype gestratificeerde selecties, waardoor elke toets in principe een unieke deelverzameling items bevatte.

Per itemtype werd gestratificeerd naar tijdens de pretest gebleken moeilijkheidsgraad. Er werd gewerkt met drie strata: moeilijk, gemiddeld, gemakkelijk.

Er werd gebruik gemaakt van twee types telefoontoestellen.
- Voor MFA-net-T en MFA-net-S: *KPN, Bari 11*.
- Voor PSTN: KPN, *dialog 3212*.

Het experiment werd uitgevoerd in het gebouw van het Ministerie van Buitenlandse Zaken in Den Haag, waar gebruik kon worden gemaakt van standaard PSTN telefoonlijnen en beide versies van het MFA-netwerk: MFA-net-T en MFA-net-S.

4.1.1.3 Design en procedure

Om bekend te raken met de toets en de toetscondities werd de eerste van de vier toetsen gebruikt als 'oefentoets'. Proefpersonen deden deze oefentoets thuis of op school. Elke proefpersoon legde vervolgens de toets in drie verschillende condities af:
- via PSTN
- via MFA-net-S
- via MFA-net-T

Om volgorde effecten te neutraliseren werd er een gebalanceerd design gebruikt, waarin de proefpersonen de drie verschillende toetsen in een van de zes mogelijke volgordes aflegden.

De conditie waarin de proefpersonen de toetsen aflegden was niet bekend bij de toetsleverancier (Ordinate). Twee medewerkers van CINOP traden op als toetsleidsters. Zij deelden de proefpersonen in, bepaalden de volgorde waarin proefpersonen de toetsen maakten en hielden hierover de administratie bij. Pas na afloop van het experiment werd de conditie aan de toetsleverancier en de onderzoeker (LTS)bekend gemaakt.

In het gebouw van het Ministerie van Buitenlandse Zaken in Den Haag werden drie testruimten ingericht: kamers met een raam, een bureau, een stoel en een telefoontoestel. De drie testruimtes verschilden wat betreft de kwaliteit van de telefoonverbinding: één was uitgerust met een PSTN-verbinding, één met MFA-net-S en één met MFA-net-T. De proefpersonen werden in groepen van 6 personen uitgenodigd om naar het Ministerie van Buitenlandse Zaken te komen. Na aankomst werd gecontroleerd of de proefpersonen de oefentoets hadden afgelegd. Proefpersonen die niet geoefend hadden, kregen de gelegenheid dat ter plekke te doen. Na een korte introductie startten de eerste drie proefpersonen met het afleggen van een toets. De andere drie wachtten of deden - indien nodig - een oefentoets. Zodra de eerste groep proefpersonen klaar was met de toets, startte de volgende groep. Op die manier konden zes proefpersonen in 90 minuten ieder drie keer de toets afleggen, met tussen de afnames steeds een pauze van circa 15 minuten. Tijdens de pauzes was er koffie en thee.

De proefpersonen werden benaderd via hun docenten aan het ROC waar ze Nederlandse lessen volgden. Via de docenten ontvingen zij, wanneer zij hun medewerking hadden toegezegd, instructies bij de oefentoets en de uitnodiging om naar het Ministerie van Buitenlandse Zaken te komen. Toen de dataverzameling eenmaal liep, meldden zich bij de toetsleiders óók personen die via andere proefpersonen op de hoogte waren gesteld van het experiment. Bij het Ministerie van Buitenlandse Zaken werden de proefpersonen ontvangen door medewerkers van het Ministerie en doorverwezen naar de afdeling waar de testen plaatsvonden. Daar werden de proefpersonen ontvangen en begeleid door de twee toetsleidsters.
De dataverzameling vond plaats tussen 21 december 2004 en 28 januari 2005.

4.1.1.4 Proefpersonen

Proefpersonen werden geworven via Regionale Opleidingencentra in Den Haag en omgeving. Docenten werden benaderd met het verzoek om onder hun cursisten proefpersonen te werven. Daarbij werd gevraagd vooral te zoeken naar beginnende leerders van het Nederlands. Voor hun medewerking aan het experiment ontvingen proefpersonen VVV-bonnen ter waarde van 40 Euro. In totaal hebben 216 proefpersonen meegewerkt. Achtergrondgegevens zijn beperkt tot het geslacht van de proefpersonen en hun mondelinge beheersingsniveau in het Nederlands volgens hun docent. Van 147 proefpersonen is een docentenoordeel over hun gespreksvaardigheid bekend. Deze worden in Tabel 4.1 gepresenteerd.

Tabel 4.1: Overzicht van docentenoordelen over gespreksvaardigheid (Den Haag)

Beheersingsniveau Volgens docent	**Aantal proefpersonen**
A1	42
A2	50
B	48
C	7
Totaal	**147**

4.1.2 Experiment Amsterdam

4.1.2.1 Probleemstellingen

Na inspectie van de resultaten van de pretesten en na overleg met geraadpleegde externe experts op relevante terreinen - taal- en spraaktechnologie, methodologie, taaltoetsing - werd in overleg met de opdrachtgever besloten om de gegevens uit de pretest aan te vullen met nieuwe gegevens over de kwaliteit van de ontwikkelde toets. In dit aanvullende experiment staat een specifiek deel van de beoogde doelgroep van het inburgeringsexamen centraal: de proefpersonen die de toets in het buitenland moeten afleggen als onderdeel van de aanvraagprocedure om in aanmerking te komen voor een Machtiging tot Voorlopig Verblijf.

Twee vragen staan centraal:
1. de betrouwbaarheid van de toets;
2. de juistheid van de cesuur (A1-min).

De betrouwbaarheidsanalyses betreffen het bepalen van de test-hertestbetrouwbaarheid en de betrouwbaarheid van de zak/slaagbeslissing. Daarnaast wordt nagegaan of de toetsscores samenhangen met het gemiddelde oordeel van twee getrainde beoordelaars over de taalvaardigheid (wel of niet A1-) van de respondent. Deze laatste analyse betreft een controle op de criteriumvaliditeit van de toets of, beter, van de zak-slaaggrens.

Er werd besloten over deze twee cruciale aspecten van de ontwikkelde toets aanvullende gegevens te verzamelen bij een steekproef die méér representatief is voor de doelgroep van het inburgeringsexamen buitenland dan de groep proefpersonen die deelnamen aan de pretest. De steekproef die deelnam aan de pretests is getrokken uit inburgeraars die reeds in Nederland verblijven en die hier onderwijs volgen, en mag derhalve verondersteld worden vaardiger te zijn in het Nederlands dan de doelgroep van het examen in het buitenland. De toets zal echter óók valide moeten zijn en differentiëren in de populatie in het buitenland die waarschijnlijk veel minder spreiding in de scores zal vertonen dan de populatie in Nederland. Daardoor zullen de gegevens omtrent de samenhang tussen de scores onderling en de samenhang met andere indicatoren (waaronder het CEF) in een steekproef die representatief is voor de populatie in het buitenland lager uitvallen dan in de pretests.

Met betrekking tot de bepaling van de predictieve waarde van de toetsscores ten opzichte van het CEF hebben we in de pretests gebruik gemaakt van gegevens die verzameld zijn door middel van de opgaven 'Verhalen navertellen'. Uit de pretest bleek dat die opgaven te moeilijk waren voor proefpersonen met een beheersingsniveau A1-min. Besloten is daarom om ten behoeve van de controle van de cesuur die op basis van de pretesten is geschat, in een nieuw experiment aanvullende gegevens te verzamelen door middel van mondelinge interviews met de proefpersonen.

4.1.2.2 Materiaal

Uit de op basis van de pretestgegevens 'gezuiverde' itembank werden 900 verschillende toetsen samengesteld door willekeurige naar itemtype gestratificeerde selectie, waardoor elke toets in principe een unieke deelverzameling items bevatte. De TINcodes van de 900 toetsen werden in drie sets van 300 toetsen aan Bureau Inburgering Amsterdam gezonden. Bureau Inburgering Amsterdam verzorgde zonder verdere tussenkomst van CINOP, LTS of Ordinate de verdeling van de toetsen over de proefpersonen.
Ten behoeve van het valideringsonderzoek werd een interviewprotocol ontwikkeld (zie bijlage 11).

Om de verzamelde gegevens direct vergelijkbaar te maken met de gegevens die in de pretesten zijn verzameld, werd voor de afname van de toetsen gebruik gemaakt van telefoonverbindingen via PSTN.

4.1.2.3 Design en procedure

De dataverzameling werd na een training verzorgd door CINOP en LTS volledig zelfstandig uitgevoerd door Bureau Inburgering Amsterdam. Binnen Bureau Inburgering Amsterdam bestaat veel ervaring met de afname van toetsen en de registratie van (intake)gegevens. Bureau Inburgering Amsterdam richtte ten behoeve van het experiment twee toetsruimtes in: kleine lokalen, voorzien van een raam, een bureau, een stoel en een telefoontoestel. Ten behoeve van de organisatie van het experiment en de begeleiding van de proefpersonen werd een psychologe aangesteld als toetsleider.

Per proefpersoon waren drie toetsen beschikbaar. Alle proefpersonen kregen de opdracht om voor de eerste toetsafname op school, thuis of elders een toets af te leggen om te wennen aan de procedure en de opgaven. Elke proefpersoon legde vervolgens bij Bureau Inburgering Amsterdam twee keer een toets af, op twee verschillende dagen met een tussenpauze van gemiddeld een week. Toen bleek dat een groot deel van de proefpersonen zich meldde voor de toets zonder een oefentoets te hebben afgelegd, heeft Bureau Inburgering Amsterdam de mogelijkheid gegeven om op school te oefenen. Helaas konden ook op die manier niet alle proefpersonen worden bereikt om een oefentoets af te leggen. De belangrijkste reden waarom proefpersonen thuis geen oefentoets af hebben gelegd, is waarschijnlijk het feit dat de meeste leden van de doelgroep thuis niet beschikken over een vaste telefoonlijn. Naar schatting 80 procent van de doelgroep maakt gebruik van mobiele telefoons, waarmee de Toets Gesproken Nederlands niet afgelegd kan worden.

Met elke proefpersoon werden daarnaast twee gesprekken gevoerd.
Het eerste gesprek werd gevoerd door docenten van Bureau Inburgering Amsterdam, die daartoe een training verzorgd door CINOP en LTS hadden gevolgd. De training werd in company uitgevoerd en besloeg twee dagdelen. De inhoud bestond uit een korte introductie op het CEF, het interviewprotocol en het bijbehorende beoordelingsmodel (zie bijlage 12). Vervolgens werden véél praktijkoefeningen gedaan aan de hand van audio- en video-opnames van taalleerders.

De betrokken docenten werden na de training op twee manieren bij het onderzoek betrokken:
- zij fungeerden als interviewer, volgens het interviewprotocol;
- zij fungeerden als beoordelaar. In die hoedanigheid waren zij onopvallend aanwezig bij de interviews.

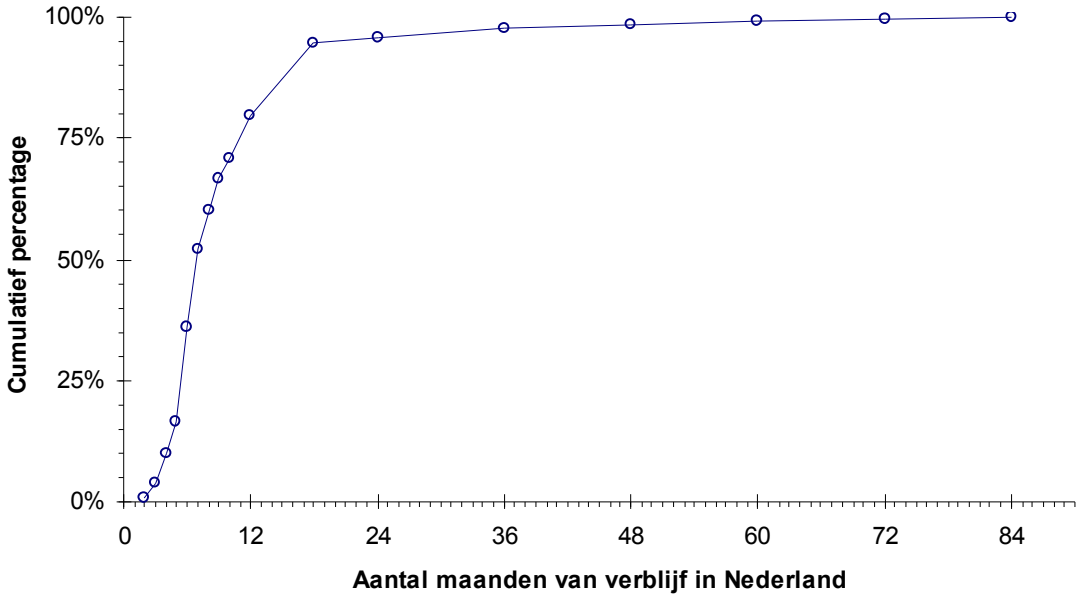
Interviewer en beoordelaar gaven na afloop van het interview zonder onderling overleg te voeren beiden een oordeel over de gespreksvaardigheid van de proefpersoon.
Het tweede gesprek was het 'loopbaangesprek' dat Bureau Inburgering Amsterdam standaard met elke ingeschreven inburgeraar voert. De mondelinge vaardigheden van de proefpersonen tijdens de loopbaangesprekken die gevoerd werden door docenten die aan de training van CINOP en LTS hadden deelgenomen, werden na afloop door de interviewer beoordeeld aan de hand van het CEF (zie bijlage 12).
Bureau Inburgering Amsterdam beschikt over een groot aantal gegevens over de achtergronden van de inburgeraars die zich melden voor een inburgeringsonderzoek. Bureau Inburgering Amsterdam heeft die gegevens, geanonimiseerd, aan CINOP ter beschikking gesteld.

Alle proefpersonen die meewerkten kregen als tegenprestatie VVV-bonnen ter waarde van 20 Euro.

4.1.2.4 Proefpersonen

Bureau Inburgering Amsterdam verzorgt de intake van alle nieuwkomers die bij het Register van de Gemeente Amsterdam en bij Vluchtelingenwerk een verblijfspas hebben gekregen. Bureau Inburgering gaat daarbij na of de nieuwkomers die zich melden in aanmerking komen voor een diagnoseprogramma. Dat diagnoseprogramma is ontwikkeld door Bureau Inburgering Amsterdam en resulteert in een gefundeerd advies omtrent een inburgeringstraject in Nederland. In de periode van 21 maart 2005 tot en met 22 april 2005 hebben alle nieuwkomers die zich bij Bureau Inburgering Amsterdam hebben gemeld en die in aanmerking kwamen voor een diagnoseprogramma meegewerkt aan het onderzoek. Dit betrof in totaal 353 personen. Het was niet mogelijk de gegevens te verzamelen bij nieuwkomers die niet in aanmerking kwamen voor een diagnoseprogramma. Om mee te werken aan het experiment moesten de proefpersonen op minimaal drie verschillende dagen naar Bureau Inburgering Amsterdam komen: twee keer voor een toets en een keer voor een interview. Het was niet haalbaar dat te organiseren met proefpersonen die geen andere redenen hadden om naar Bureau Inburgering te komen. Tot de groep die NIET heeft meegewerkt omdat ze niet in aanmerking komen voor een diagnoseprogramma, behoren: analfabeten, Surinamers, Antillianen en personen die een ontheffing of een vrijstelling hebben gekregen van hun inburgeringsverplichting.

De verhouding man:vrouw was circa 44 : 56 en de gemiddelde leeftijd bedroeg 30 jaar. Het opleidingsniveau van de proefpersonen was erg divers, ongeveer een derde van de proefpersonen was laagopgeleid. Figuur 4.1 laat zien dat de meeste proefpersonen op het moment van de toets minder dan twaalf maanden in Nederland verbleven en dat 95% van de proefpersonen minder dan anderhalf jaar in Nederland verbleef.



Figuur 4.1: Overzicht verblijfsduur in Nederland proefpersonen Experiment Amsterdam (n=328, ontbrekende gegevens: 9)

Wat de herkomst van de proefpersonen betreft kan er gesteld worden dat er een grote variatie aan herkomstlanden was (57 landen), met concentraties uit Turkije (21%) en Marokko (20%).

4.1.3 Experiment MFA-Fit

4.1.3.1 Probleemstellingen

Het derde experiment betreft het 'MFA-Fit experiment'. In dit experiment werden verschillende doelstellingen gecombineerd, die voor een deel in de naamgeving ervan naar voren komen.

De onderzoeksvragen die in dit experiment centraal staan, betreffen in de eerste plaats de kwaliteit van de verzamelde data. Bij de analyse van de pretestgegevens wordt gebruik gemaakt van Rasch-analyses. Rasch-analyses vereisen unidimensionele data. Het design van de pretest biedt onvoldoende mogelijkheid om na te gaan of aan deze voorwaarde van unidimensionaliteit wordt voldaan. Alleen de over proefpersonen gemiddelde split-half betrouwbaarheid van de toets en inter-item correlaties bieden een indicatie. Hoewel de gegevens die de pretest daarover oplevert op unidimensionaliteit wijzen, is besloten aanvullend onderzoek te doen naar de kwaliteit van de dataset.

De tweede onderzoeksvraag betreft de invloed van het telefoonnetwerk van het Ministerie van Buitenlandse Zaken op de scores van de proefpersonen. De resultaten van het Experiment Den Haag (zie par. 4.2) toonden aan dat er significante verschillen bestaan tussen de scores die proefpersonen in verschillende condities behaalden. Er werden geen significante verschillen waargenomen tussen de twee condities binnen het netwerk van Buitenlandse Zaken (MFA-net-T en MFA-net-S), maar de verschillen tussen de twee MFA-net condities enerzijds en PSTN anderzijds bleken significant. Nadere analyses toonden verder aan dat MFA-net een verschillend effect had op de in de Toets Gesproken Nederlands onderscheiden subscores. Omdat op de Nederlandse ambassades in het buitenland afname via PSTN niet in alle landen mogelijk is, werd besloten te zoeken naar een procedure met behulp waarvan de scores op via MFA-net afgelegd toetsen, gecorrigeerd kunnen worden.

De in Den Haag verzamelde gegevens boden onvoldoende mogelijkheden om te komen tot een verantwoorde procedure voor score correctie vanwege een te geringe representatie van lage vaardigheidsniveaus.

4.1.3.2 Materiaal

Volgens de standaardprocedure van Ordinate's toetssysteem worden er per proefpersoon toetsen samengesteld door willekeurige naar itemtype gestratificeerde selectie uit de itembank. Door deze procedure bevat elke toets in principe een unieke deelverzameling items en is de kans dat twee toetsen meer dan 17% overlappen, erg klein. Het direct vaststellen van de unidimensionaliteit van de toetsen wordt door deze procedure echter bemoeilijkt. Om die reden is besloten om data te verzamelen met betrekking tot een beperkt aantal 'vaste toetsen'. Door willekeurige naar itemtype gestratificeerde selectie zijn drie verschillende toetsen samengesteld. Deze drie toetsen, die elk afzonderlijk 48 verschillende opgaven bevatten die willekeurig uit de itembank zijn geselecteerd, bevatten samen 144 opgaven en worden beschouwd als representatief voor de hele itembank en de toetsen die daaruit samengesteld kunnen worden. De drie toetsen worden hierna aangeduid als 'oefentoets', 'toets A' en 'toets B'.

Ten behoeve van de onderzoeksvraag met betrekking tot de controle van de cesuur, werden in dit experiment naast de toetsscores twee onafhankelijke gegevens over de taalvaardigheid van de proefpersonen verzameld.

- Met elke proefpersoon werd een 'adaptief' interview gehouden door een daartoe getrainde interviewer. De gespreksvaardigheid van de proefpersonen tijdens het interview werd drie keer beoordeeld: door de interviewer, door een daartoe getrainde beoordelaar die bij het interview aanwezig was en door een derde beoordelaar aan de hand van een opname op audiocassette van het interview.
- De toets werd gevolgd door drie open vragen. Deze vragen vervingen de twee opgaven 'Verhalen navertellen' die gebruikt werden in de pretest en bij de twee andere experimenten. Het is gebleken dat de opgaven 'Verhalen navertellen' voor beginnende leerders van het Nederlands te moeilijk zijn. Voor elk van de drie vaste toetsen werd een apart setje van drie vragen samengesteld. De drie vragen per toets lopen op in het minimale beheersingsniveau dat ze van proefpersonen veronderstellen. De eerste vraag zou door een proefpersoon met een mondeling beheersingsniveau op A1-min-niveau begrepen en beantwoord moeten kunnen worden. Deze vraag biedt in principe ook geen ruimte om in het antwoord een hoger beheersingsniveau te demonstreren. De tweede vraag betreft een vraag naar feitelijke informatie over persoonlijke ervaringen. Volgens de CEF-normen zou deze vraag door proefpersonen met een beheersingsniveau A1 adequaat beantwoord moeten kunnen worden. In het antwoord kunnen proefpersonen echter ook hogere niveaus van vaardigheid tonen. In de derde vraag worden proefpersonen niet alleen om een beschrijving van objectieve gegevens gevraagd, maar ook om een mening en argumenten daarbij. De vraag sluit wat de vereiste luistervaardigheid betreft aan bij de CEF-beschrijving van niveau A2 en is ook in minimale vorm op dit niveau te beantwoorden. De vraag biedt echter ook ruimte om een antwoord tot zelfs C2-niveau te geven. De vragen werden ingesproken door de vrouwelijke stemacteur die ook de instructies bij de andere onderdelen van de toets heeft ingesproken. De proefpersonen hoorden alle vragen twee keer: één keer in een normaal, rustig spreektempo, en één keer in een nadrukkelijk aan het niveau van beginnende taalleerders aangepast tempo. In Tabel 4.2 wordt een overzicht van de vragen gegeven.

Tabel 4.2 Overzicht van de vragen.

Oefentoets
1 In welk jaar bent u geboren?
2 Hoe leert u Nederlands? Vertel wat u doet om Nederlands te leren. Heeft u taalles? Leest u boeken? Of kranten? Spreekt u wel eens met mensen in het Nederlands?
3 Kijkt u wel eens naar de tv? Wanneer kijkt u televisie? Naar welke programma's kijkt u graag? Vertel wat over een programma dat u mooi vindt.
Set A
1 In welk land bent u geboren?
2 Waar woont u nu? Vertel iets over de plaats waar u woont. Woont u in een stad of in een dorp? Woont u in een flat? Of woont u in een huis? Vertel iets over uw huis. Hoe ziet uw huis eruit?
3 Welk soort weer vindt u prettig? Houdt u van heet weer of van koud weer? Houdt u van regen of van zonneschijn? Waarom vindt u dit weer fijn?
Set B
1 Hoe lang woont u in Nederland?
2 Vertel iets over uw familie. Heeft u een grote familie? Vertel iets over uw broers en uw zussen. Heeft u ook kinderen? Vertel iets over uw kinderen.
3 Woont u het liefst in een grote stad of in een klein dorp? Wat is er goed aan het leven in de stad? En wat is er goed aan het leven in een dorp?

4.1.3.3 De simulator

Omdat het niet haalbaar was dit experiment met de benodigde aantallen proefpersonen uit te voeren met gebruikmaking van het 'echte' MFA-net, omdat dat telefoonnet in Nederland alleen bij het Ministerie van Buitenlandse Zaken beschikbaar is, werd een simulator ontwikkeld met behulp waarvan de effecten van het MFA-net werden gesimuleerd.

De MFA-net simulator past compressie en decompressie volgens aanbeveling G.729 van de Internationale Telecommunicatie Unie in Genève toe op alle gesproken opgaven en op de antwoorden van de proefpersonen. Daarnaast wordt vertraging toegevoegd zoals deze ook optreedt bij het MFA-net. Uit analyses van het experiment Den Haag was gebleken dat de vertraging groter was dan de waarden opgegeven door het Ministerie van Buitenlandse Zaken: voor landlijnen was de empirische vertraging van MFA-net via landlijnen 397 milliseconden groter dan via PSTN en voor satellietlijnen liep het verschil op tot 731 milliseconden. Door toepassing van compressie en decompressie volgens aanbeveling G.729 en de vertraging volgens de bevindingen in Den Haag komen de data verzameld in het MFA-Fit experiment technisch overeen met de data verzameld in het experiment Den Haag.

4.1.3.4 Procedure en design

Ten behoeve van de onderzoeksvraag met betrekking tot de correctie van scores die via MFA-net zijn verkregen werden via tussenpersonen verbonden aan scholen en asielzoekerscentra proefpersonen geworven die volgens hun docenten een beheersingsniveau in het Nederlands hadden op en onder niveau A-1min. De proefpersonen werden door de tussenpersonen willekeurig toegewezen aan één van vier experimentele groepen: G1, G2, G3 en G4.

Ten behoeve van het onderzoek naar de unidimensionaliteit van de data werden via dezelfde tussenpersonen proefpersonen geworven die volgens hun docenten een mondeling beheersingsniveau hadden van minimaal A1-min (dus ook van hogere niveaus). Zij werden door de tussenpersonen willekeurig toegewezen aan twee andere groepen, G5 en G6.

De proefpersonen in de groepen G1 tot en met G4 legden in een voor volgorde-effecten gebalanceerd design de toetsen A en B af in de twee onderscheiden condities: één keer via de MFA-net simulator en één keer via PSTN. Aan de reacties verzameld via de MFA-net simulator werd zowel de empirisch gevonden vertraging van landlijnen (397 ms) als die van de satellietverbinding (731 ms) toegevoegd. Op deze wijze werden voor deze proefpersonen drie scores gegenereerd: PSTN, MFA-net-T en MFA-net-S. De proefpersonen in G5 en G6 legden volgens een gebalanceerd design toets A en toets B beide keren via PSTN af. Tabel 4.3 geeft een overzicht van het design, uitgaande van een minimum van 128 complete casussen voor G1 tot en met G4 en 192 complete casussen voor G5 en G6.

Tabel 4.3: Design experiment MFA-Fit

Groep	CEF- niveau	Proefpersonen	Oefentoets	Afname 1	Afname 2	Minimaal aantal complete casussen
G1	<A1-min tot A1	50	PSTN-Oefen.	PSTN-ToetsA	MFA-ToetsB	32
G2	<A1-min tot A1	50	PSTN-Oefen.	MFA-ToetsB	PSTN-ToetsA	32
G3	<A1-min tot A1	50	PSTN-Oefen.	MFA- ToetsA	PSTN-ToetsB	32
G4	<A1-min tot A1	50	PSTN-Oefen.	PSTN-ToetsB	MFA-ToetsA	32
G5	A1-min en hoger	150	PSTN-Oefen.	PSTN-ToetsA	PSTN-ToetsB	96
G6	A1-min en hoger	150	PSTN-Oefen.	PSTN-ToetsB	PSTN-ToetsA	96
Totaal		**500**				**320**

De dataverzameling werd op de verschillende locaties georganiseerd door de tussenpersonen werkzaam bij die locaties. CINOP leverde elke tussenpersoon voor het door hem of haar opgegeven aantal proefpersonen TINcodes, geordend per toets en per conditie. De tussenpersonen verdeelden de TINcodes over de proefpersonen. Alle proefpersonen legden alle toetsen in gecontroleerde omstandigheden af op school of in het asielzoekerscentrum waar zij woonachtig zijn. Tussen de afnames van de verschillende toetsen hadden proefpersonen steeds minimaal een pauze van een half uur. De toetsleiding was in handen van medewerkers van de betreffende instellingen. De interviews werden afgenomen door daartoe getrainde medewerkers van de instellingen aan de hand van het door CINOP geleverd interviewprotocol dat ook in het expertiment Amsterdam was gebruikt (zie bijlage 11). Betrokken medewerkers wisselden elkaar af als interviewer en als beoordelaar. Beiden gaven een van elkaar onafhankelijk oordeel van de taalvaardigheid van de proefpersoon uitgedrukt op de CEF-schaal.

Van alle interviews werden bovendien audio-opnames gemaakt. Deze opnames maakten het mogelijk een aselecte steekproef aan een derde beoordeling te onderwerpen. De opnames zijn beoordeeld door daartoe getrainde medewerkers van de deelnemende instellingen, echter steeds van een andere instelling dan die waar de opnames waren gemaakt.

Alle proefpersonen kregen als blijk van waardering VVV-bonnen ter waarde van 20 Euro.

4.1.3.5 Proefpersonen

De proefpersonen werden geworven bij Regionale Opleidingen Centra, via het Centraal Orgaan Asielzoekers en via het netwerk van een particulier taleninstituut. Via tussenpersonen werden de proefpersonen geworven en werd informatie over hun mondeling beheersingsniveau in het Nederlands verzameld bij betrokken docenten/begeleiders.
Omdat er maar weinig tijd beschikbaar was voor de dataverzameling - een week - werd bij de werving van proefpersonen rekening gehouden met een mogelijk zeer hoge mortaliteit, dat wil zeggen proefpersonen die niet zouden komen opdagen dan wel die niet alle toetsen, inclusief oefentoets en interview, zouden afleggen. Doordat de uitval mee bleek te vallen, zijn meer gegevens verzameld dan gepland.

De verhouding man:vrouw was 40 : 60 (n = 498) en gemiddeld verbleven de proefpersonen 3.2 jaar in Nederland. Van de proefpersonen was 24% laagopgeleid, 33% middenopgeleid en 39% hoogopgeleid. Wat de herkomst van de proefpersonen betreft kan er gesteld worden dat er ook hier een grote variatie aan herkomstlanden (82) is. Er werden concentraties vastgesteld van deelnemers uit Afghanistan (7%), Irak (10%) Marokko (7%), Turkije (13%) en diverse Afrikaanse landen (16%).

Tabel 4.4 toont de gerealiseerde aantallen per groep met een volledige set van drie toetsen. Bijna alle groepen hebben de verwachte aantallen proefpersonen opgeleverd (zie Tabel 4.3). Aangezien voor het onderzoek naar dimensionaliteit alleen de PSTN-versie is betrokken en er tussen toetsafnames geen vergelijking op individueel niveau noodzakelijk is, zijn voor dit deel van het onderzoek per proefpersoon alle toetsafnamen nà de eerste als geldige toets opgevat en de eerste (welke dat ook was) als oefentoets. Op deze wijze kon uit alle groepen waarvoor PSTN-Toets A de tweede toets was een groep van 434 proefpersonen worden samengesteld. Op gelijke wijze kon een groep van 406 proefpersonen worden samengesteld voor wie PSTN-Toets B de tweede toets was. Bij het samenstellen van deze groepen werden ook reacties meegenomen van proefpersonen die slechts twee toetsen hebben gemaakt, waarvan de tweede PSTN-Toets A of PSTN-Toets B.

Tabel 4.4: Gerealiseerde aantallen per groep met volledige toetsset.

Groep	CEF- niveau	Oefentoets	Afname 1	Afname 2	Minimaal aantal complete casussen
G1	<A1-min tot A1	PSTN-Oefen.	PSTN-ToetsA	MFA-ToetsB	36
G2	<A1-min tot A1	PSTN-Oefen.	MFA-ToetsB	PSTN-ToetsA	53
G3	<A1-min tot A1	PSTN-Oefen.	MFA- ToetsA	PSTN-ToetsB	44
G4	<A1-min tot A1	PSTN-Oefen.	PSTN-ToetsB	MFA-ToetsA	29
G5	A1-min en hoger	PSTN-Oefen.	PSTN-ToetsA	PSTN-ToetsB	122
G6	A1-min en hoger	PSTN-Oefen.	PSTN-ToetsB	PSTN-ToetsA	177
Totaal					**461**

4.2 Resultaten

4.2.1 De invloed van telefoonlijnen op de toetsscores

4.2.1.1 Evaluatie van de invloed van Telefoonlijnen

Tabel 4.5 presenteert de resultaten door proefpersonen behaald op de drie versies (PSTN, MFA-net-T enMFA-net-S) bij de toetsafnamen op het Ministerie van Buitenlandse Zaken in Den Haag.

Tabel 4.5: Resultaten Experiment Den Haag

	PSTN	MFA-net-T	MFA-net-S
Gemiddelde score	52.37	46.48	46.29
Standaard Deviatie	11.96	9.45	9.63
Betrouwbaarheid	0.915	0.897	0.891
Standaard meetfout	3.49	3.03	3.18
Minimum	25	24	23
Maximum	80	75	75
N (proefpersonen)	168	168	168

Tabel 4.5 toont dat, afgezien van het verschil tussen de gemiddelde toetsscores, de distributies voor de drie experimentele condities, PSTN, MFA-net-S en MFA-net-T, overeenkomen. Door middel van t-tests werden de verschillen tussen de condities getoetst. Tabel 4.6 vat de uitkomsten samen.

Tabel 4.6: Verschillen tussen gemiddelde scores per conditie (Den Haag)

	Kolom 1 PSTN – MFA-net-T	**Kolom 2** PSTN – MFA-net-S	**Kolom 3** MFA-net-T – MFA-net-S	**Kolom 4** gemiddelde van kolom 1 & 2
Totaalscore	5.89*	6.07*	-0.18	5.98
Uitspraak	7.11*	6.87*	0.25	6.99
Vloeiendheid	9.03*	9.63*	-0.60	9.33
Woordenschat	3.87*	3.79*	0.08	3.83
Zinsbouw	2.71*	2.91*	-0.20	2.81

* significantie p<0.05

De t-tests tonen aan dat de condities significante effecten hebben. Een twee-weg variantieanalyse werd uitgevoerd om na te gaan hoeveel van de variantie in de dataset van 168 proefpersonen wordt veroorzaakt door verschillen tussen de zes groepen die aan de afzonderlijke mogelijke volgordes waren toegekend, en hoeveel door verschillen tussen de verschillende condities. Tabel 4.7 laat de uitkomsten zien.

Tabel 4.7: Twee-weg variantieanalyse volgorde en condities (Den Haag)

Bron variantie	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Volgorde	2192.2	5	438.44	3.89	0.00	2.24
Condities	2915.4	1	2915.44	25.84	0.00	3.87
Interactie	62.2	5	12.45	0.11	0.99	2.24
	36553.9	324	112.82			
Totaal	41723.8	335				

Uit Tabel 4.7 blijkt dat de zes subgroepen, hoewel de proefpersonen willekeurig aan ieder van de zes verschillende volgordes waren toegewezen, duidelijk significant ($p<0.01$) verschillen. Er is echter géén significante interactie tussen de condities en de volgorde van condities waarin de toetsen werden afgenomen. Deze bevinding, gecombineerd met de vergelijkbare verdelingen voor de drie telefooncondities zoals weergegeven in Tabel 4.5, leidt tot de conclusie dat de verschillende telefooncondities weliswaar leiden tot significante scoreverschillen op groeps- en individueel niveau, maar niet tot verschillen in de ordening van de proefpersonen en de onderlinge afstanden. Deze conclusie wordt verder ondersteund door de gegevens over de correlaties tussen de verschillende condities. Deze correlaties worden in Tabel 4.8 weergegeven boven de diagonaal. In de vetgedrukte diagonaal van Tabel 4.8 worden de betrouwbaarheden van de toetsen in de verschillende condities weergegeven. Onder de diagonaal staan cursief weer de correlaties tussen de verschillende condities, maar nu gecorrigeerd voor attenuatie.

Tabel 4.8: Correlaties tussen condities en betrouwbaarheden (Den Haag)

	PSTN	MFA-net-T	MFA-net-S
PSTN	**0.92**	0.86	0.87
MFA-net-T	*0.950*	**0.90**	0.87
MFA-net-S	*0.96*	*0.97*	**0.89**

De gegevens in Tabel 4.8 laten zien dat er behalve de grote hoeveelheid gemeenschappelijke variantie tussen de drie condities en de errorvariantie voor elke afzonderlijke conditie, slechts weinig variantie onverklaard blijft, wat erop wijst dat in de drie condities in essentie dezelfde variabele wordt getoetst.

Samenvattend kunnen we concluderen dat het gebruik van MFA-net een significant negatief effect heeft op de scores van de proefpersonen.

4.2.1.2 Correctie van de invloed van de telefoonlijnen

De Toets Gesproken Nederlands is in de eerste plaats ontwikkeld met het oog op gebruik via PSTN-telefoonlijnen. Alle gegevens die zijn verzameld over de kwaliteit van de toets, zijn dan ook gebaseerd op gegevens over proefpersonen die de toets via PSTN hebben afgelegd. Nu de resultaten van het onderzoek in Den Haag hebben aangetoond dat gebruik van het MFA-net significante invloed heeft op de scores van proefpersonen, zal gezocht moeten worden naar een mogelijkheid om de onderliggende schaal van de MFA-netversie van de toets te equivaleren naar de PSTN-schaal.

Aangezien de deelscores niet in gelijke mate door het MFA-net worden beïnvloed, is een oplossing voor ieder van de deelscores afzonderlijk noodzakelijk. Voor ieder van de deelscores zijn regressiefuncties berekend voor de vaardigheidschattingen. Deze worden op de theta's toegepast voordat transformatie naar de rapportageschalen (zie 2.10) plaatsvindt. Tabel 4.9 geeft de correcties weer waarmee de deelscores zijn getransformeerd.

Tabel 4.9: Transformatie naar PSTN schaalscores:

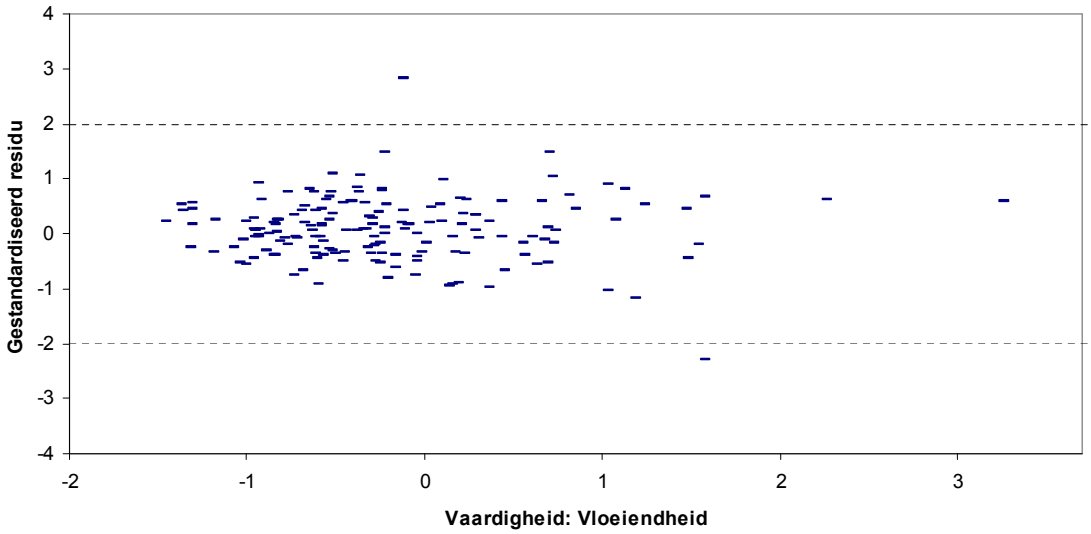
Vloeiendheid:	1.0227*MFA-net Vloeiendheid score + 0.3794
Uitspraak:	0.8955 *MFA-net Uitspraak score + 0.2165
Woordenschat:	0.7556 *MFA-net Woordenschat score + 0.2174
Zinsbouw:	0.9409 *MFA-net Zinsbouw score - 0.0724

Na correctie was geen van de t-toetsen voor verschillen tussen de condities meer significant. Op individueel niveau treden uiteraard nog verschillen tussen toetsscores op. Uit inspectie van deze verschillen bleken deze gemiddeld in minder dan 5% significant op individueel niveau, hetgeen de uitkomst van de t-toets bevestigt. Ook de samenhang tussen de verschillen en de gemeten vaardigheid treedt na correctie niet meer op.

Figuur 4.2 toont voor Vloeiendheid – de deelscore waar de grootste verschillen optraden – een inspectie van de resterende verschillen na toepassing van de correctiefactor. Op de horizontale as is de gemeten vaardigheid via MFA-net (na correctie) afgebeeld en op de verticale as de gestandaardiseerde verschilscore met de via PSTN geschatte vaardigheid. De 5% overschrijdingskans is weergegeven met twee horizontale stippellijnen. De figuur maakt duidelijk dat er nog slechts enkele proefpersonen zijn voor wie de scores significant verschillen en dat verschillen niet preferent in één richting optreden. Het grootste aantal overschrijdingen (3.7%) werd aangetroffen bij de scores voor Zinsbouw.

De gevonden transformatiefuncties werden vervolgens toegepast op de MFA-Fit data. Zoals te verwachten, was de passing minder goed (maximum aantal overschrijdingen 6.2%). Aangezien de MFA-Fit data waren verzameld via een simulator van het MFA-net en de Haagse data via het werkelijke net, werd besloten de gevonden transformatiefuncties te handhaven.

Geconcludeerd kan worden dat in theoretische zin correctie op de negatieve impact van het gebruik van MFA-net mogelijk is. Ook in praktische zin is correctie uitvoerbaar. Toetsen die zijn afgelegd op diplomatieke posten waar via het MFA-net zal worden getoetst, zullen door het automatische scoringssysteem aan de hand van hun telefoonnummer worden herkend. Via het correctiealgoritme zal de output van de automatische scoring worden gecorrigeerd voor de nadelige invloed van het MFA-net.



Figuur 4.2: Inspectie op individueel niveau van de resterende effecten van MFA-net na correctie.

4.2.2 Dimensionaliteit

Evaluatie van de dimensionaliteit kon voor de vier onderscheiden subschalen niet op gelijke wijze worden verricht. De beide kwaliteitsmaten - vloeiendheid en uitspraak – worden, zoals beschreven in Hoofdstuk 2, gegenereerd door een voor overeenstemming met menselijke oordelen geoptimaliseerde combinatie van respectievelijk een aantal temporele en een aantal fonologische maten. De resulterende continue itemscores kunnen niet met de gebruikelijke toets- en itemanalyse programma's worden geanalyseerd. De beide subscores hebben daarnaast gemeenschappelijk dat bij het uitblijven van een respons geen score kan worden toegekend. Immers wanneer een proefpersoon geen antwoord geeft, is er geen observatie en kan daarom geen oordeel worden gevormd over de vlotheid en de uitspraak. Dit brengt met zich mee dat de datasets veel 'missing values' bevatten. Een aanduiding voor unidimensionaliteit kan echter gevonden worden in een uniform patroon van hoge correlaties van de itemscores met de totaalscore.

De beide inhoudsmaten worden daarentegen wel discreet gescoord. Voor de bepaling van de score op woordenschat worden dichotome beslissingen gegenereerd: of de proefpersoon het beoogde goede antwoord al dan niet heeft gegeven. De scores voor zinsbouw worden gegeven op grond van de herhaalopdrachten. Zij worden negatief polytoom gescoord met een maximum itemscore die varieert met de lengte van de herhaalopdracht.
De deelscores worden berekend zoals beschreven in Hoofdstuk 2. Het is goed daarbij op te merken dat voor de herhaalopdrachten geldt dat hogere scores een groter aantal fouten voorstellen. De maximale score is per herhaalopdracht afhankelijk van het aantal woorden in de stimulus. De eigenschappen van de itemtypen waarmee de vaardigheid op de onderscheiden subschalen wordt gemeten, laten niet eenzelfde onderzoek naar dimensionaliteit toe.

Onderstaand Tabel 4.10 toont per itemtype de gekozen analyse ter achterhaling van de dimensionaliteit.

Tabel 4.10 Analyse dimensionaliteit per itemtype

Schaal	Aard van de score	Min	Max	Gekozen analyse	Software
Woordenschat	Positief dichotoom	0	1	IRT – OPLM	OPLM
Zinsbouw	Negatief polytoom	Variabel	0	IRT - FACETS	FACETS
Vlotheid	Positief continu	- ∞	+ ∞	Correlatie	MSExcel/SPSS
Uitspraak	Positief continu	- ∞	+ ∞	Correlatie	MSExcel/SPSS

Op de navolgende pagina's worden de resultaten gepresenteerd per deelvaardigheid/subschaal. Voor iedere subschaal zijn steeds resultaten van de twee gebruikte versies van de toets beschikbaar (zie par. 4.1.3.2) zodat door replicatie meer zekerheid wordt verkregen over de bevindingen. Wanneer resultaten op beide versies vergelijkbaar zijn, ondersteunt dit de aanname dat aselect getrokken toetsen uit de itembank gelijkwaardige toetsen kunnen opleveren. In de volgende paragrafen behandelen we achtereenvolgens de itemsoorten in de volgorde van bovenstaande Tabel.

4.2.2.1 Woordenschat

Analyses werden uitgevoerd met het programma OPLM (Verhelst, Glas en Verstralen, 1991). Dit programma biedt het voordeel dat itemparameters conditioneel kunnen worden geschat zodat geen aannamen, bijvoorbeeld over de verdeling van de vaardigheid, hoeven te worden gedaan. Het programma gaat uit van het één-parameter Raschmodel maar biedt de mogelijkheid enige variatie in de discriminatiewaarden van items toe te laten. Voor beide versies is gekozen is voor eenzelfde, laag geometrisch gemiddelde: 2. In Tabel 4.11 worden de belangrijkste resultaten samengevat.

Tabel 4.11: Woordenschat: dimensionaliteit (MFA-Fit)

	Versie A	***Versie B***
Aantal proefpersonen	434	406
Aantal items	22	22
Gemiddelde ruwe score	11.647	11.94
Standaard deviatie	4.058	3.941
Betrouwbaarheid (alpha)	0.778	0.746
Geometrisch gemiddelde discriminatie	2.032	1.948
R1c	77.800	77.179
Vrijheidsgraden	63	63
Overschrijdingskans	0.0993	0.1079
Gemiddelde theta	0.040	0.040
Variantie geschatte theta's	0.323	0.276
Variantie ware theta's	0.249	0.205
Betrouwbaarheid thetaschattingen	0.769	0.741

De analyse toont aan dat de data passen bij het gekozen meetmodel. Het meetmodel laat de discriminatieparameter weliswaar vrij maar de vrijheid is beperkt door het lage gemiddelde: in de versies A en B hebben 60% respectievelijk 72% van de items eenzelfde discriminatieparameter van 2. Een mogelijk bezwaar tegen passing met een gewogen model is dat de somscore niet langer een sufficiënte statistiek voor de schatting van de vaardigheid vormt. De correlatie tussen somscore en de latente vaardigheid werd echter door de beperkte vrijheid van de discriminatieparameter slechts in zeer geringe mate beïnvloed door de keuze voor het gewogen model. Voor beide versies was deze correlatie 0.990 voor het ongewogen model en 0.98 en daalde naar 0.980 voor Versie A en naar 0.981 voor versie B. De twee aselect getrokken itemsets vertonen een hoge mate van onderlinge overeenkomst. Deze bevindingen duiden erop dat de tijdens de pretest verzamelde gegevens voldoende informatie bevatten om homogene sets van items te kunnen samenstellen.

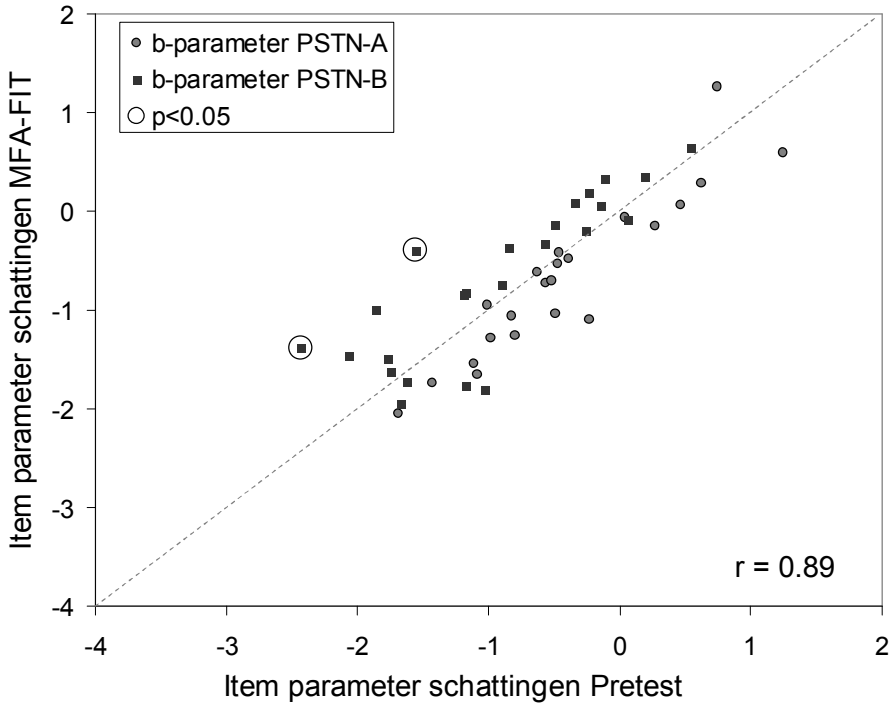
4.2.2.2 Zinsbouw

Analyses werden uitgevoerd met het programma FACETS (Linacre ,1988/2005). Voor dit programma is gekozen omdat het de mogelijkheid biedt items met uiteenlopende scoringsvoorschriften in één analyse onder te brengen. Het heeft echter als nadeel dat geen conditionele schatters kunnen worden toegepast. Parameterschattingen in dit programma worden uitgevoerd met de Joint Maximum likelihood Procedure (JML). In Tabel 4.12 worden de belangrijkste resultaten samengevat.

Tabel 4.12: Zinsbouw: dimensionaliteit (MFA-Fit)

	Versie A	***Versie B***
Aantal proefpersonen	434	406
Aantal items	24	24
Gemiddelde ruwe score	98.4	64.80
Standaard deviatie	9.858	6.369
Gemiddelde Punt-biseriële correlatie	0.67	0.63
Standaarddeviatie van de ptbis	0.06	0.06
Gemiddelde itemfit	0.99	0.99
Standaarddeviatie van de itemfit	0.40	0.39
Chi-kwadraat over hele toetsonderdeel	5940.5	4632.1
Vrijheidsgraden	432	405
Overschrijdingskans	0.00	0.00
Gemiddelde theta	0.00	0.040
Variantie geschatte theta's	0.348	0.700
Variantie ware theta's	0.331	0.456
Betrouwbaarheid thetaschattingen	0.95	0.93

Formeel gezien dient de aanname van theoretische passing van het model vanwege de hoge chi-kwadraat te worden verworpen. Anderzijds wijzen de hoge punt-biseriële correlaties en de individuele itemfitwaarden (verwachte waarde 1.0) erop dat praktische consequenties gering zijn. Dit blijkt ook uit inspectie van enkele aannamen die men bij voldoende passing theoretisch zonder meer zou kunnen maken: consistentie van de schattingen van de moeilijkheid van items en van de vaardigheid van personen. Dit zijn juist de redenen waarom men modelpassing nastreeft: om er vanuit te kunnen gaan dat itemschattingen niet afhankelijk zijn van de steekproef van proefpersonen. Figuur 4.3 geeft de moeilijkheidschattingen weer van de sets items in toetsversie A en B in het MFA-Fit experiment afgezet tegen de schattingen op basis van de pretest . De itemschattingen gebaseerd op heel verschillende soorten van afnamen op andere groepen proefpersonen stemmen in hoge mate met elkaar overeen.



Figuur 4.3: Schattingen van moeilijkheidsgraad van herhaalopdrachten op basis van pretesten en MFA-Fitexperiment (MFA-Fit)

Een tweede reden waarom unidimensionaliteit wordt nagestreefd, is dat dient te worden gewaarborgd dat de meting van de vaardigheid van proefpersonen geschiedt onafhankelijk van de verzameling items die in hun toets voorkomen. In de MFA-Fit dataset zitten in groep 5 en 6 in totaal 296 proefpersonen die aan drie toetsen hebben deelgenomen. De correlatie tussen de resultaten op de toetsen afgenomen op T1 en T2 bedraagt 0.80.

4.2.2.3 Vloeiendheid

Wanneer een proefpersoon geen respons geeft op een herhaalopdracht kan over vloeiendheid van spreken geen uitspraak worden gedaan, er is immers geen observatie. Dit leidt tot een groot aantal missende waarnemingen, waardoor geen standaard toets- en itemanalyse mogelijk is. Het is wel mogelijk om gemiddelde waarden per item te berekenen over de proefpersonen die de betreffende items wel hebben beantwoord. In onderstaande Tabellen wordt een samenvatting gegeven over deze waarden zoals verzameld over de herhaalopdrachten in PSTN toetsA en PSTN toetsB.

De Tabellen laten zich als volgt lezen. We nemen Tabel 4.13a als voorbeeld. In de kolom N lezen we per regel achtereenvolgens wat over de items het gemiddelde aantal proefpersonen is (bijna 248), wat daarvan de standaard deviatie is, dat er minstens één item is dat door 292 proefpersonen is beantwoord, dat er minstens één item is waarop slechts 195 proefpersonen hebben geantwoord en dat het verschil tussen het item met het hoogste aantal antwoorden en dat met het laagste aantal woorden uitkomt op 97.

Tabel 4.13a: Vloeiendheid PSTN toets A: dimensionaliteit (MFA-Fit)

.	N	Missing	Gemidd.	StDev	Max	Min	Range	r_{it}	$r_{i\#miss}$
Gemidd.	247.92	29%	-0.200	0.605	2.195	-0.944	3.139	0.505	-0.192
StDev	23.09	7%	0.191	0.116	0.390	0.062	0.395	0.078	0.079
Max	292	44%	0.389	0.939	2.981	-0.857	3.917	0.634	0.004
Min	195	16%	-0.475	0.479	1.482	-1.087	2.475	0.282	-0.326
Range	97	28%	0.864	0.459	1.499	0.231	1.441	0.352	0.331

Betrouwbaarheidschatting (Cronbach's alpha): 0.883 (n=349)

Tabel 4.13b: Vloeiendheid PSTN toets B: dimensionaliteit (MFA-Fit)

.	N	Missing	Gemidd.	StDev	Max	Min	Range	r_{it}	$r_{i\#miss}$
Gemidd.	252.17	25%	-0.103	0.648	2.145	-0.937	3.081	0.487	-0.206
StDev	23.67	7%	0.258	0.124	0.373	0.098	0.395	0.065	0.062
Max	290	37%	0.450	0.875	2.817	-0.849	3.811	0.603	-0.089
Min	213	14%	-0.505	0.446	1.490	-1.312	2.374	0.367	-0.322
Range	77	23%	0.955	0.430	1.328	0.463	1.437	0.236	0.234

Betrouwbaarheidschatting (Cronbach's alpha): 0.868 (n=336)

Beide versies van de toets vertonen een zeer vergelijkbaar patroon. Er is zoals men zou verwachten een geringe negatieve correlatie tussen het aantal missing en de geschatte vaardigheid op basis van de wel beoordeelde items. De correlatie van de itemscore met de totaalscore is uniform hoog. De betrouwbaarheid duidt op een consistente meting, de vergelijkbaarheid van beide versies duidt erop dat willekeurige trekkingen uit de opgavenbank vergelijkbare toetsen opleveren.

4.2.2.4 Uitspraak

Evenals vloeiendheid kan uitspraak niet beoordeeld worden wanneer een proefpersoon geen respons geeft op een vraag. Ook bij uitspraak treffen we daarom een groot aantal missende waarnemingen, waardoor geen standaard toets- en itemanalyse mogelijk is. Onderstaande Tabellen geven een samenvatting op toetsonderdeelniveau (herhaalopdrachten) voor waarden die met betrekking tot deze items kunnen worden berekend. De indeling van de Tabellen is gelijk aan die bij Vloeiendheid.

Tabel 4.14a: Uitspraak PSTN toets A: dimensionaliteit (MFA-Fit)

.	N	Missing	Gemidd.	StDev	Max	Min	Range	r_{it}	$r_{i\#miss}$
Gemidd.	260.71	26%	-0.338	0.610	1.329	-1.659	2.987	0.414	-0.108
StDev	23.26	7%	0.186	0.056	0.266	0.509	0.548	0.099	0.112
Max	304	39%	0.200	0.746	1.953	-1.371	4.690	0.589	0.147
Min	212	13%	-0.593	0.531	0.935	-3.601	2.426	0.169	-0.277
Range	92	26%	0.794	0.215	1.017	2.230	2.263	0.419	0.425

Betrouwbaarheidschatting (Cronbach's alpha): 0.844 (n=350)

Tabel 4.14b: Uitspraak PSTN toets B: dimensionaliteit (MFA-Fit)

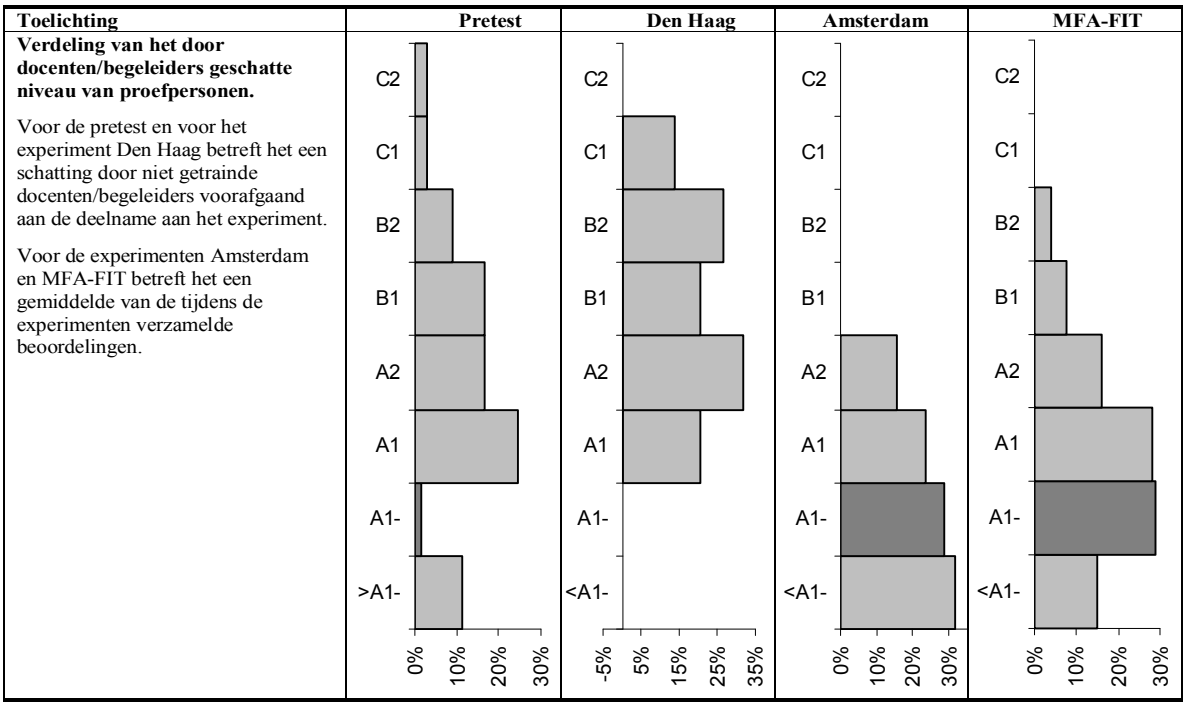
.	N	Missing	Gemidd.	StDev	Max	Min	Range	r_{it}	$r_{i\#miss}$
Gemidd.	261.92	22%	-0.277	0.621	1.449	-1.654	3.103	0.420	-0.108
StDev	21.73	6%	0.231	0.054	0.268	0.528	0.634	0.072	0.106
Max	294	34%	0.226	0.716	1.899	-1.354	5.554	0.556	0.074
Min	223	13%	-0.655	0.546	0.976	-3.655	2.438	0.294	-0.340
Range	71	21%	0.881	0.169	0.924	2.301	3.116	0.263	0.414

Betrouwbaarheidschatting (Cronbach's alpha): 0.841 (n=337)

Evenals bij Vloeiendheid het geval was, vertonen beide versies van de toets ook bij de analyse van de deelscore Uitspraak, een vergelijkbaar patroon. We vinden weer een lichte negatieve correlatie tussen het aantal missing en de vaardigheid geschat op het aantal waargenomen antwoorden en hoge item-rest correlaties. Deze bevindingen ondersteunen de aanname dat de deelvaardigheid Uitspraak consistent wordt gemeten.

4.2.3 Geschiktheid voor lage vaardigheidsniveaus en test-hertest betrouwbaarheid

Om te onderzoeken of de toets geschikt is om ook op lagere taalvaardigheidsniveaus te meten is bij twee van de aanvullende experimenten nadrukkelijk gestreefd naar de werving van proefpersonen op en rond het A1-min niveau. Figuur 4.4 toont een overzicht van door docenten en beoordelaars opgeleverde informatie over de verdeling van de vaardigheid in de onderscheiden experimenten. Ter vergelijking is hierbij ook de verdeling van de pretesteekproef opgenomen. De oordelen in Amsterdam en bij het MFA-fit experiment zijn zijn afkomstig van getrainde beoordelaars.



Figuur 4.4: Verdeling van taalvaardigheid in de pretest en aanvullende experimenten

Figuur 4.4. toont aan dat er – in ieder geval naar het oordeel van betrokken docenten en beoordelaars - bij de experimenten Amsterdam en MFA-Fit veel meer proefpersonen op en rond het A1-min niveau betrokken waren dan bij de pretest en het experiment in Den Haag.

Alhoewel alleen het experiment Amsterdam speciaal was opgezet om gegevens te verzamelen over de test-hertest betrouwbaarheid van de toets, zijn er bij ieder van de drie experimenten proefpersonen geweest die de toets na het afleggen van een oefentoets twee of meer keer hebben afgelegd. Er is hierbij dus geen sprake van test-hertest in de letterlijke zin dat eenzelfde toets tweemaal wordt afgenomen maar van een meermalige toetsing met een parallelle vorm. Hierdoor treedt niet het nadeel van een leereffect op met betrekking tot de toets*opgaven* zelf. Door het gebalanceerde design worden voorts vermoeidheids-verschijnselen en bekendheid met de toets*vorm* gecompenseerd. Tabel 4.15 geeft een overzicht van de test-hertest betrouwbaarheid zoals deze voor de totaalscore en voor de deelscores kon worden geschat in de drie aanvullende experimenten.

Tabel 4.15: Test-hertest betrouwbaarheid

		Den Haag (n=168)	**Amsterdam** (n=94)	**MFA-Fit** (n=243)
Totaalscore T1-T2		**0.88**	**0.72**	**0.87**
Deelscores	Uitspraak	0.82	0.76	0.85
	Vloeiendheid	0.84	0.70	0.63
	Woordenschat	0.62	0.67	0.73
	Zinsbouw	0.82	0.73	0.80

Tabel 4.15 toont - behalve voor Amsterdam - een acceptabele betrouwbaarheid. De betrouwbaarheid in het MFA-Fit experiment wijkt niet af van die in het experiment Den Haag en toont aan dat een redelijke betrouwbaarheid kan worden bereikt ook wanneer nadrukkelijk proefpersonen met een laag taalvaardigheidsniveau in de steekproeven zijn opgenomen. In het laatste hoofdstuk komen we terug op de toetsbetrouwbaarheid en de validiteit bij de lagere vaardigheidsniveaus.

Hier willen we de vraag of de geschatte betrouwbaarheden stand houden bij een meer uniform laag taalvaardige doelgroep afsluiten met een analyse waarin bij één van de deeltoetsen alle proefpersonen die volgens hun docenten een taalvaardigheid hoger dan A2 hebben, uit de groep zijn verwijderd. Tabel 4.16 toont de uitkomsten. Genomen is de deeltoets woordenschat bij toetsversie A in het MFA-Fit-experiment omdat in bijna alle schattingen de betrouwbaarheid van de deeltoets woordenschat als laagste van de vier uit de bus komt. Zestien personen met geschatte CEF-niveaus hoger dan A2 zijn verwijderd. Ter vergelijking is de oorspronkelijke analyse zoals opgenomen in Tabel 4.11 ter linkerzijde afgedrukt. De verwijdering van de personen op B1 en hoger uit de data blijkt geen invloed van betekenis te hebben op de gerapporteerde waarden.

Tabel 4.16: Kwaliteit deelscore woordenschat, zonder proefpersonen boven niveau A2 (MFA-Fit)

	Versie A	Versie A[1]
Aantal proefpersonen	434	418
Aantal items	22	22
Gemiddelde ruwe score	11.647	11.41
Standaard deviatie	4.058	3.91
Betrouwbaarheid (alpha)	0.778	0.758
Geometrisch gemiddelde discriminatie	2.032	2.032
R1c	77.800	75.963
Vrijheidsgraden	63	63
Overschrijdingskans	0.0993	0.1266
Gemiddelde theta	0.040	0.001
Variantie geschatte theta's	0.323	0.276
Variantie ware theta's	0.249	0.209
Betrouwbaarheid thetaschattingen	0.769	0.759

[1] Proefpersonen met een door hun docenten geschat taalvaardigheidsniveau onder of gelijk aan A2

4.2.4 Conclusies

De drie aanvullende experimenten hadden gezamenlijk de volgende hoofddoelstellingen:

- het achterhalen van de invloed van het telefoonnet van het Ministerie van Buitenlandse zaken en het, zonodig, bepalen van een compensatiefactor;
- het achterhalen van de schaalbaarheid van deelscores en
- het achterhalen van de test-hertest betrouwbaarheid van de toets ook bij lagere niveaus.

De invloed van het telefoonnet is in de experimenten Den Haag en MFA-Fit onderzocht. Gebleken is dat het afwijkende telefoonnet van het Ministerie van Buitenlandse inderdaad de scores negatief beïnvloedt maar dat hiervoor een acceptabele correctiefactor kan worden gevonden.

De schaalbaarheid van de deelscores is onderzocht in het experiment MFA-Fit. Twee aparte willekeurige deelverzamelingen op examenlengte van opgaven uit de itembank zijn onderzocht. Voor geen van de deelscores is evidentie gevonden dat de itemscores in deze deelverzamelingen onvoldoende samenhang vertonen om een geldige totaalscore te kunnen bepalen.

In de experimenten Amsterdam en MFA-Fit is nadrukkelijk gestreefd naar het samenstellen van steekproeven met een relatief grote vertegenwoordiging van proefpersonen met lage vaardigheidsniveaus. Uit de hier gepresenteerde resultaten over het MFA-Fit experiment is niet gebleken dat het onderscheidend vermogen van de toets te gering wordt wanneer uitsluitend personen met lage vaardigheidsniveaus aan de toets deelnemen.

5 Schaling en normering

Wanneer eenmaal is vastgesteld dat een verzameling opgaven geschikt is om toetsen samen te stellen waarmee op consistente wijze verschillen tussen personen - op grond van hun kennis of andere eigenschappen - kunnen worden vastgesteld, moet vervolgens nog worden bepaald wat de betekenis is van die verschillende toetsresultaten. Aangezien de uitslag van de Toets Gesproken Nederlands moet worden gerelateerd aan het CEF, zijn er tijdens de pretest en de aanvullende onderzoeken gegevens verzameld over de oordelen van docenten over de mondelinge taalvaardigheid in het Nederlands van de deelnemende proefpersonen. De docenten gaven hun oordelen op basis van de niveaubeschrijvingen van het CEF. In dit hoofdstuk rapporteren wij over deze dataverzamelingen en de wijze waarop deze zijn gebruikt om de relatie tussen de toetsuitslag en de CEF-niveaus te bepalen.

5.1 CEF beoordelingen

Tijdens de pretest zijn op twee manieren CEF-beoordelingen over proefpersonen verzameld:

- de docenten van de proefpersonen (n = 94) gaven een schatting van hun taalvaardigheidsniveau en
- vijf getrainde beoordelaars beoordeelden de reacties van de proefpersonen op het onderdeel 'Verhalen navertellen'.

Tegen beide methoden zijn bezwaren in te brengen. Wanneer docenten onvoldoende vertrouwd zijn met het CEF zal hun onzekerheid over het CEF zich reflecteren in de scores die zij aan proefpersonen toekennen. De opdracht Verhalen navertellen bleek moeilijk voor proefpersonen op een laag taalvaardigheidsniveau. Daardoor verliest deze opdrachtsoort juist op dat deel van de scoreschaal waar bij gebruik van de toets in het buitenland de hoogste precisie is vereist, aan onderscheidend vermogen.

Vanwege deze bezwaren zijn er in de aanvullende experimenten Amsterdam en MFA-Fit voorzieningen getroffen om juist over de lage taalniveaus aanvullende informatie te verzamelen.
In het experiment Amsterdam is dat op twee manieren gedaan.

- De proefpersonen zijn geïnterviewd aan de hand van een gestructureerd interview protocol (zie Bijlage 11). Bij ieder interview gaven de interviewer en een aanwezige beoordelaar onafhankelijk van elkaar een CEF-oordeel over de gespreksvaardigheid van de proefpersoon.
- De taalvaardigheid van de proefpersonen is - met toestemming van de proefpersoon - beoordeeld tijdens een loopbaangesprek. Bij ieder loopbaangesprek gaf de medewerker die het gesprek voerde een CEF-oordeel over de gespreksvaardigheid van de proefpersoon.

Bij MFA-Fit is opnieuw op twee manieren aanvullende informatie over de mondelinge vaardigheden van de proefpersonen verzameld.

- De proefpersonen zijn geïnterviewd volgens het voor Amsterdam ontwikkelde gestructureerde interview protocol. Bij ieder interview gaven de interviewer en een aanwezige beoordelaar onafhankelijk van elkaar een oordeel. Een derde beoordelaar gaf een oordeel op grond van een audio-opname van het interview.
- In de toets is het onderdeel 'Verhalen navertellen' vervangen door een set in moeilijkheidsgraad oplopende open vragen.

Naast deze oordelen ten behoeve van de schaalontwikkeling werden er ook menselijke oordelen verzameld voor de training van de automatische scoring (zie Hoofdstuk 2) en voor de validering van de scoring en van de schaling (zie Hoofdstuk 6). Verder werd ook aan de docenten van de proefpersonen die deelnamen aan het experiment Den Haag, net zoals bij de pretest, een schatting van het niveau van hun cursisten gevraagd. In totaal waren er meer dan 147 verschillende personen op enig moment betrokken bij de beoordelingen op grond van het CEF die in het kader van het ontwikkel-, schalings- en valideringsproces van de Toets Gesproken Nederlands zijn verzameld. Tabel 5.1 geeft een overzicht van de menselijke oordelen die voor de ontwikkeling, de schaling en de validering van de Toets Gesproken Nederlands zijn verzameld.

Tabel 5.1: Beoordelingen en beoordelaars betrokken bij de ontwikkeling, schaling en validering

Onderzoek	Materiaal	Medium	Aspect	Aantal	Beoordelaars nrs
Pretests, ontwikkeling	Herhaalopdrachten	Telefoon	Uitspraak	7	1-7
		Telefoon	Vloeiendheid	7	1-7
	Verhalen navertellen	Telefoon	Globaal	5	1-4, 8
Pretests, validering	Docentervaring met cursist	Life	Globaal	93	*53-145**
	Herhaalopdrachten	Telefoon	Uitspraak	6	1-4, 8-9
		Telefoon	Vloeiendheid	6	1-4, 8-9
	Verhalen navertellen	Telefoon	Globaal	5	1-4, 10
Den Haag	Docentervaring met cursist	Life	Globaal		*147 – nn**
Amsterdam, validering	Interview	Life	Globaal, 1e oordeel	10	12, 14-15, 18, 21-23, 25, 27, 30
			Globaal, 2e oordeel	10	12, 14-15, 18, 21-23, 25, 27, 30
	Loopbaangesprek	Life	Globaal	13	11-13, 16, 18-20, 22, 26-27, 29-31
MFA-Fit, schaling	Interviews	Life	Globaal, 1e oordeel	19	2, 32-49
			Globaal, 2e oordeel	19	2, 32-49
		Cassette	Globaal, 3e oordeel	6	31, 34, 36, 39, 46-47
	Open vragen	Telefoon	Globaal, 1e oordeel	5	1-2, 50-52
			Globaal, 2e oordeel	5	1-2, 50-52

* Niet of in beperkte mate formeel getraind met het CEF

5.1.1 Kenmerken van de beoordelaars

De docenten die in het kader van de pretest en van het experiment Den Haag vooraf een globale schatting gaven van het CEF-niveau van de door hen geworven proefpersonen, hadden over het algemeen géén training genoten. Alle overige beoordelaars (beoordelaars 1 t/m 52) ontvingen een intensieve training zoals beschreven in de volgende paragraaf.

De 8 beoordelaars die betrokken waren bij de ontwikkelfase van de pretest, waren geselecteerd op basis van de volgende criteria:

- moedertaalsprekers van het Nederlands;
- opleiding op minimaal mbo-niveau;
- géén bijzondere ervaring met het communiceren met leerders van het Nederlands en
- bereid om een beoordelaarstraining te volgen.

De acht beoordelaars verschilden wat betreft opleiding en beroep: twee CINOPadviseurs, één secretaresse, drie studenten (diergeneeskunde, Noors, algemene letteren), één receptioniste en één toetsexpert. Het belangrijkste selectiecriterium was 'geen bijzondere ervaring met het communiceren met leerders van het Nederlands'. Voor alle beoordelaars gold dat ze niet méér ervaring hadden met het communiceren met leerders van het Nederlands dan een gemiddelde Nederlander. Er waren géén NT2-docenten of andere personen betrokken die beroepsmatig veel te maken hebben met taalleerders. Ook personen met partners voor wie het Nederlands niet de moedertaal was, kwamen niet als beoordelaar in aanmerking.

De beoordelaars bij het experiment in Amsterdam hadden daarentegen allemaal véél ervaring met leerders van het Nederlands als tweede taal. Alle beoordelaars waren als NT2-docent, als docent maatschappij-oriëntatie en/of als trajectbegeleider werkzaam bij Bureau Inburgering Amsterdam. Voor twee beoordelaars gold dat het Nederlands niet hun moedertaal was.

Ook de beoordelaars die betrokken waren bij de interviews in het experiment MFA-Fit hadden allemaal veel ervaring met tweede-taalleerders. Een aantal beoordelaars was als NT2-docent werkzaam bij een Regionaal Opleidingen Centrum of bij een particuliere taleninstelling. De anderen waren werkzaam bij asielzoekerscentra. Alle beoordelaars waren moedertaalsprekers van het Nederlands. De reacties op de 'open vragen' werden weer beoordeeld door vijf getrainde beoordelaars die geen van allen veel ervaring hadden met leerders van het Nederlands als tweede taal.

5.1.2 De beoordelaarstraining

Voor de beoordelaars die bij de pretest waren betrokken, werden twee trainingen van elk twee dagdelen verzorgd en twee opfristrainingen van ieder één dagdeel. De andere beoordelaars kregen ieder twee dagdelen training.
Voor de pretestontwikkeling werden 6 beoordelaars getraind in het beoordelen van uitspraak en vloeiendheid. De training werd verzorgd door de zevende beoordelaar, John de Jong. De training bestond uit een korte introductie bij het project en bij de achtergronden en de opzet van het CEF. Met behulp van diverse werkvormen, waaronder puzzels en zelfbeoordelingen, werden de beoordelaars bekend gemaakt met een aantal schalen van het CEF: de globale schalen, de schaal voor gespreksvaardigheid en de schalen voor uitspraak en vloeiendheid. Het grootste deel van de training – meer dan de helft van het eerste dagdeel, het hele tweede dagdeel, en de twee opfrisdagdelen – werd besteed aan het oefenen in het gebruik van de beoordelingsschalen voor uitspraak en vloeiendheid (zie bijlage 4 en 5). Daarbij werd gebruik gemaakt van twee soorten materiaal:

- audio en video opnames van her en der verzamelde opnames van gesprekken met taalleerders en
- reacties van proefpersonen op herhaalopdrachten.

De procedure was steeds gelijk: de beoordelaars beluisterden samen een fragment en bepaalden daarop zelfstandig een score. Op een teken van de trainer toonden vervolgens alle deelnemers een gekleurde kaart met daarop het door hen toegekende niveau. Vervolgens werd een aantal deelnemers gevraagd de criteria voor hun keuze te noemen. Desgewenst werd het te beoordelen fragment dan nogmaals ten gehore gebracht. Wanneer deelnemers onderling verschilden in het toegekende niveau werd uitgebreid gediscussieerd en werd getracht tot consensus te komen. Deelnemers waren echter vrij om bij hun oorspronkelijke keus te blijven. Naarmate er meer fragmenten beoordeeld en besproken waren, was er minder discussie nodig en nam de overeenstemming tussen beoordelaars toe.

Voor de beoordeling van de reacties op 'Verhalen navertellen' werden vijf beoordelaars getraind. Vier van hen hadden ook de training met betrekking tot uitspraak en vloeiendheid gevolgd.

Omdat de andere reeds getrainde beoordelaars niet in de gelegenheid waren óók mee te werken aan het beoordelen van de navertelde verhaaltjes, werd er één nieuwe beoordelaar (studente kunstgeschiedenis) aangesteld. De opzet van de training was vergelijkbaar met de eerste training: veel oefenen, veel uitwisselen van argumenten. De beoordelaars werkten nu aan de hand van de beoordelingsschaal die is opgenomen in Bijlage 12. Kenmerkend voor die schaal is dat het beoordelingsproces in twee of drie stappen wordt verdeeld. Tijdens de eerste stap wordt de spreker ingedeeld in één van vier niveaus:

Niveau C	Vaardige gebruikers van het Nederlands: personen waarmee nagenoeg moeiteloos gecommuniceerd kan worden.
Niveau B	Onafhankelijke gebruikers van het Nederlands: personen die zich zonder hulp kunnen redden in het Nederlands, maar hierbij wel wat fouten maken, en waarvan men kan horen dat het Nederlands niet de moedertaal is.
Niveau A	Afhankelijke (of essentiële) gebruikers van het Nederlands: personen waarmee communicatie mogelijk is, echter op voorwaarde van coöperatie van gesprekspartners.
Niveau 0	'Rudimentaire' gebruikers van het Nederlands: personen die weinig of niets lijken te begrijpen van wat er wordt gezegd en niet of nauwelijks te verstaan zijn.

Tijdens de tweede stap bepaalt de beoordelaar vervolgens de precieze score. Aan de hand van het tweede deel van het beoordelingsinstrument bepaalt de boordelaar binnen het gekozen niveau of het niveau hoog (C2, B2, A2) of laag is (C1, B1, A1). Wanneer bij de eerste stap niveau 0 is toegekend bepaalt de beoordelaar of personen helemáál geen Nederlands beheersen (0) of dat er sprake is van een rudimentaire vaardigheid op niveau A1-min.

Tijdens de training werden de beoordelaars getraind in het gebruik van dit model. Er werd weer geoefend met audio en video opnames van gesprekken met taalleerders én met voorbeelden van navertelde verhaaltjes uit de pretest die via de telefoon werden beluisterd. Elk voorbeeld werd op de hiervoor beschreven wijze besproken: "Eerst zoeken we overeenstemming over het globale niveau en daarna differentiëren we binnen dat niveau".

Voor de beoordeling in de validatiefase van de pretest werd een 'opfristraining' gegeven. Hieraan namen ook enkele nieuwe beoordelaars deel.

De beoordelaars in Amsterdam en in MFA-Fit werden op een vergelijkbare manier getraind. De twee groepen beoordelaars kregen ieder een training van twee dagdelen. De training werd uitgebreid met een onderdeel 'interviewen', waarin de beoordelaars uitleg kregen bij het interviewprotocol en ermee oefenden.

De trainingen werden verzorgd door John de Jong, Anne Kerkhoff en Petra Poelmans. De beoordelaars waren zonder uitzondering tevreden over de cursus die zij beoordeelden als interessant, leerzaam en plezierig. De docenten gaven aan het prettig te vinden om zo intensief "met het vak" bezig te zijn. De gegevens over de kwaliteit van het werk van de beoordelaars tonen aan dat de trainingen ook effectief zijn geweest.

De docenten/begeleiders die niveau-inschattingen gaven van de proefpersonen die bij de pretest en bij het experiment Den Haag betrokken waren, hebben geen specifieke training volgens het hierboven beschreven model ontvangen. Zij baseerden hun oordelen op een schriftelijke beschrijving van de CEF-niveaus en op hun ervaringen met de grotendeels parallelle indeling van de NT2-niveaus volgens het Referentiekader NT2.

5.1.3 Beoordelingsprocedures

5.1.3.1 'Verhalen navertellen' bij pretest

De getrainde beoordelaars deden hun werk via de telefoon en met gebruikmaking van het toetssysteem van Ordinate. De beoordelaars krijgen hierbij van Ordinate een unieke identificatiecode. Wanneer zij hun werk willen doen, kiezen ze een rustige omgeving met een goed werkende vaste telefoon en bellen naar het toetssysteem van Ordinate. Het toetssysteem vraagt de beoordelaar om diens identificatiecode in te voeren. Het systeem herkent aan de code dat het een beoordelaar betreft en biedt achtereenvolgens een stimulus en een serie responsen van verschillende proefpersonen bij die stimulus aan. De beoordelaar voert per respons de door hem of haar toegekende score in via de numerieke druktoetsen op de telefoon. De beoordelaar kan op ieder gewenst moment de verbinding verbreken om het beoordelen op een later moment voort te zetten. Het systeem biedt ook de mogelijkheid responsen een tweede of derde maal te beoordelen of om responsen over te slaan. Beoordelaars hebben ruim de tijd om hun oordeel te vormen: pas na circa 10 minuten van non respons van de beoordelaar wordt de verbinding verbroken.

Ten behoeve van de pretesten leverden vijf getrainde beoordelaars ieder circa 480 oordelen. Samen zorgden zij voor een dataset met in totaal 2401 oordelen, verdeeld over 500 proefpersonen. Van elke proefpersoon werden twee responsen beoordeeld (beide 'Verhalen navertellen'). Elke respons werd door tenminste twee beoordelaars beoordeeld, met 10 procent willekeurige herhaling en 10 procent toekenning aan een derde beoordelaar.

5.1.3.2 Interviews bij experimenten Amsterdam en MFA-Fit

In het experiment Amsterdam en in MFA-Fit voerden medewerkers van de betrokken instellingen volgens een vast protocol gesprekken met de deelnemende proefpersonen. De gesprekken waren opgebouwd in opklimmende moeilijkheid. De betrokken medewerkers van de instellingen fungeerden afwisselend als interviewer en als beoordelaar. Bij de training was duidelijk gemaakt dat het interview erop gericht moest zijn de proefpersoon in de gelegenheid te stellen diens hoogst mogelijke vaardigheid te demonstreren. De interviewer moest gaande van laag naar hoog telkens een volgend niveau proberen totdat duidelijk was dat de proefpersoon niet hoger kon. Interviewers hadden een formulier tot hun beschikking waarop zij 'net echt' de antwoorden van de proefpersonen konden noteren. De beoordelaars hadden een meer gedetailleerd formulier met de aspecten die geacht werden in het interview aan de orde te komen. De formulieren konden als handleiding bij het voeren van het gesprek en het geven van een beoordeling gebruikt worden. Men werd echter niet verplicht deze formulieren in te vullen. Alle betrokken medewerkers waren nadrukkelijk geïnstrueerd over het belang van onafhankelijke oordelen.

In Amsterdam werden behalve oordelen over de gespreksvaardigheid tijdens een gestructureerd interview ook taalvaardigheidsoordelen verzameld tijdens het op die instelling standaard met elke deelnemer gevoerde loopbaangesprek. Het oordeel werd gegeven door de medewerker die het loopbaangesprek volgens de reguliere procedures uitvoerde. De betrokken medewerkers behoorden tot de groep die de training in interviewen en beoordelen had gevolgd.

In Amsterdam en bij het experiment MFA-Fit werden de interviews opgenomen op cassette. Alleen de cassettes van het MFA-Fit experiment zijn in het kader van de onderhavige rapportage aan een derde beoordeling onderworpen. De cassettes werden onder instellingen onderling 'geruild' zodat een derde beoordeling ter beschikking kwam van een getrainde beoordelaar die niet bij het interview aanwezig was geweest.

5.1.3.3 Open vragen bij MFA-Fit

De drie open vragen die bij het experiment MFA-Fit aan het eind van de toets werden gesteld - ter vervanging van het onderdeel Verhalen navertellen - werden beoordeeld door 5 getrainde beoordelaars. Zij voerden de beoordelingen uit op dezelfde wijze als de beoordelingen van het onderdeel Verhalen navertellen bij de pretest: via een telefoonverbinding met het scoringssysteem van Ordinate werden de reacties beluisterd en de oordelen ingevoerd.

In het kader van het MFA-Fit experiment werden twee aparte sets oordelen verzameld: de eerste set ten behoeve van onderzoek naar de betrouwbaarheid van de beoordelingen (betrouwbaarheidsonderzoek) en de tweede set ten behoeve van de onderbouwing van de relatie tussen CEF en toetsscores (hoofdonderzoek).

Ten behoeve van het betrouwbaarheidsonderzoek werd aselect een steekproef van 49 proefpersonen getrokken uit de totale groep proefpersonen die deelnamen aan het experiment MFA-Fit. Vijf getrainde beoordelaars beoordeelden de reacties van alle 49 subjecten op elk van de drie open vragen. Op die manier werden 3 x 5 x 49 = 735 beoordelingen verzameld. Bij de beoordeling werden de responsen per vraag aan de beoordelaars voorgelegd. De beoordelaars hoorden telkens eerst een vraag gevolgd door een serie aselect zonder terugplaatsing getrokken antwoorden van proefpersonen. Uitgangspunt was dat er voldoende vertrouwen zou bestaan om de beoordelaars voor het hoofdonderzoek in te zetten wanneer betrouwbaarheid kon worden aangetoond in een volledig gevulde matrix met de vijf beoordelaars. Dat bleek het geval (zie paragraaf 5.1.4).

Dezelfde vijf beoordelaars beoordeelden vervolgens in het hoofdonderzoek in totaal 300 proefpersonen (anderen dan die betrokken waren bij het betrouwbaarheidsonderzoek) op drie open vragen. De antwoorden van de proefpersonen op de open vragen werden toegewezen aan telkens twee aselect getrokken beoordelaars uit de pool van vijf. Willekeurig werd 10% van alle responsen een tweede maal aan dezelfde beoordelaar toegewezen. Bovendien werd 10% aselect getrokken van alle responsen en aan een willekeurige derde beoordelaar toegewezen. Het verwachte aantal beoordelingen in het kader van deze analyse was 3 x 1.1 x 2.2 x 300 = 2.178.

5.1.4 Kwaliteit van de menselijke oordelen

De verzameling oordelen over de responsen op het onderdeel Verhalen navertellen werd geanalyseerd met versie 3.54.1 van het programma FACETS (Linacre, 2004). De betrouwbaarheid van de vaardigheidsoordelen op de CEF schaal bedroeg 0.97. Er waren 1.743 cases waarbij het oordeel van twee beoordelaars kon worden vergeleken. In 68.2% van deze gevallen kwamen de oordelen van beide beoordelaars exact overeen.

Bij het experiment Amsterdam werden over 276 proefpersonen drie oordelen gegeven. De beoordelingen van de interviewer en de beoordelaar die ieder afzonderlijk oordeelden over hetzelfde interview stemden in 93% van de gevallen exact met elkaar overeen. Voor beiden kwam het gemiddelde oordeel uit op A1-min. Ook de beoordelingen bij het apart gehouden loopbaangesprek kwamen gemiddeld uit op A1-min. Dit oordeel stemde in 51%, respectievelijk 48% van de gevallen exact overeen met het oordeel van de interviewer en van de beoordelaar bij het interview. Bij een dichotome beslissing (ónder de ondergrens van A1-min of daarboven) stemden beide op grond van het interview gegeven beoordelingen in 98% van de gevallen exact met elkaar overeen, terwijl zij ieder afzonderlijk in 80% respectievelijk 79% overeenstemden met het oordeel gegeven op grond van het loopbaangesprek.

Bij het experiment MFA-Fit waren voor 243 proefpersonen volledige data beschikbaar over drie toetsen en drie beoordelingen. De beoordelingen van de interviewer en de beoordelaar die ieder afzonderlijk oordeelden over hetzelfde interview stemden in 68% exact met elkaar overeen. Voor beiden kwam het gemiddelde oordeel uit op A1.
De overeenstemming tussen de twee oordelen op basis van het interview en die van de apart uitgevoerde beoordeling van hetzelfde interview opgenomen op cassette, stemden respectievelijk in 40% en in 41% van de gevallen exact met elkaar overeen. Bij een dichotome beslissing (onder de ondergrens van A1-min of daarboven) stemden beide op grond van het interview gegeven beoordelingen in 95% van de gevallen exact met elkaar overeen, terwijl zij ieder afzonderlijk in 87% respectievelijk 84% overeenstemden met het oordeel gegeven op grond van de cassetteopname van het interview.

De oordelen verzameld voor het betrouwbaarheidsonderzoek naar de beoordeling van de open vragen werden geanalyseerd met het programma BEOVER (Heuvelmans, 2002). Aangezien de vragen onderling niet vergelijkbaar zijn, werd de analyse per vraag afzonderlijk uitgevoerd. Tabel 5.2 geeft een overzicht.

Tabel 5.2: Betrouwbaarheidsonderzoek menselijke oordelen over open vragen (MFA-Fit)

	Vraag 1	**Vraag 2**	**Vraag 3**
Geschatte variantie componenten			
Proefpersonen	82.8	76.3%	81.4%
Beoordelaars	0.0%	0.5%	2.6%
Residu (fout)	17.2%	23.2%	16.90%
Beoordelaarsovereenstemming	0.96	0.94	0.96
Geschatte overeenstemming bij 2 beoordelaars	0.91	0.87	0.90
Gower's coëfficiënt	0.96	0.43	0.95

Op grond van deze resultaten kon worden besloten het hoofdonderzoek met alle vijf beoordelaars uit te voeren. Voor het hoofdonderzoek werden 2.402 beoordelingen verzameld, circa 800 per vraag.

5.2 Bepaling grensscores in relatie tot CEF-schaal

5.2.1 Bepaling CEF-schaal

In de pretest waren onvoldoende observaties op zeer lage niveaus beschikbaar om de ondergrens voor A1-min met dezelfde zekerheid op de schaal te projecteren als dat voor de hoger liggende niveaus mogelijk was. Bovendien waren opgaven getoetst in samenhang met opgaven die niet voor opname in de itembank geschikt bleken. Dit zou de resultaten op de wel goedgekeurde items mogelijk hebben kunnen beïnvloeden. Er is daarom bij de evaluatie van de pretest besloten om met de definitieve bepaling van de toetsscoreschaal (en daarmee ook de grensscores) te wachten totdat er meer gegevens bekend zouden zijn over de examens in hun operationele vorm. De experimenten Amsterdam en MFA-Fit bieden deze mogelijkheid. De toetsen in deze experimenten zijn op gelijke wijze samengesteld uit dezelfde itembank als de toekomstige examens. Voor de schaalontwikkeling is een deel van de oordelen verzameld bij de pretest samengevoegd met oordelen verzameld in het experiment MFA-Fit. Een ander deel van de pretestoordelen en de oordelen verzameld in het experiment Amsterdam werden apart gehouden om een eenmaal gevonden schaal te kunnen valideren met onafhankelijke data.

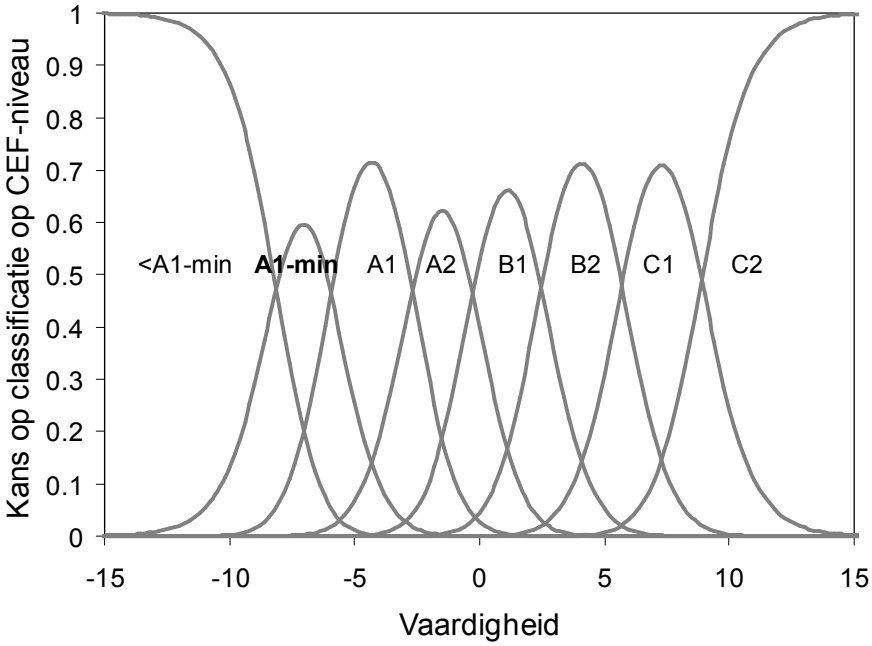
CEF-beoordelingen over 'Verhalen navertellen' uit de pretest werden samen met oordelen over de Open vragen 2 en 3 en oordelen over de interviews uit het experiment MFA-Fit samengevoegd en in één FACETS-analyse opgenomen.

Vraag 1 kon niet gelijk met andere vragen of beoordelingen worden geanalyseerd omdat bij vraag 1 alleen scores 0 (nog onder A1-min) en 1 (beheerst minimaal A1-min) kunnen voorkomen.

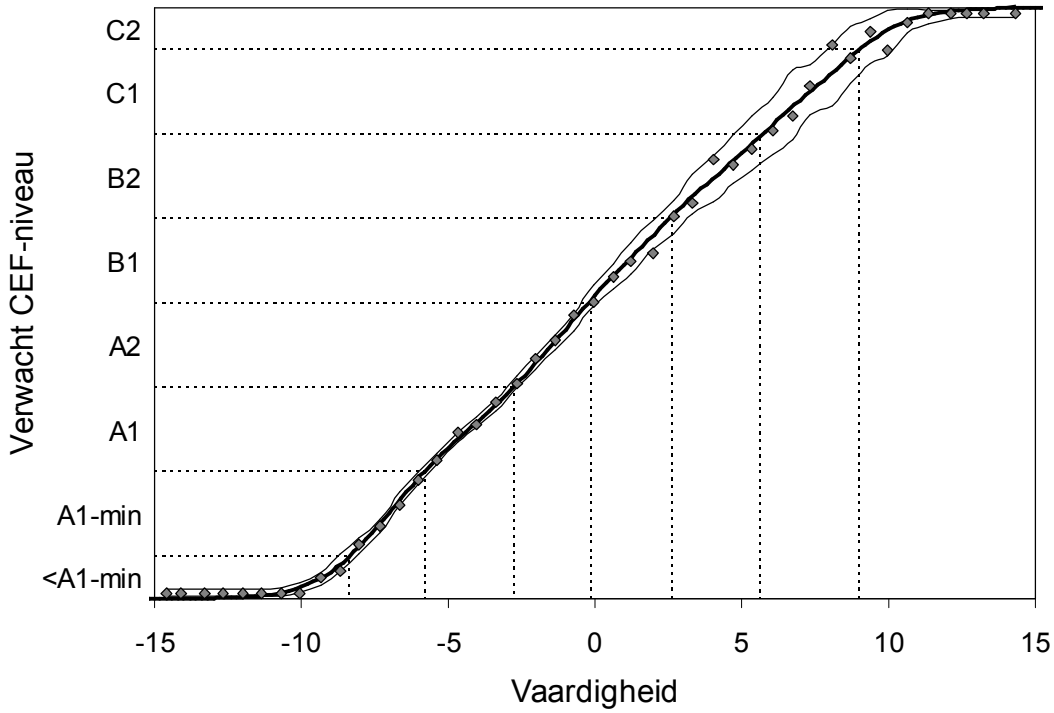
Hierdoor zou het behalen van een score van 1 moeilijker worden geschat dan het behalen van hogere scores op vragen 2 en 3. In totaal zijn in de gecombineerde dataset 5.984 oordelen beschikbaar, van 14 verschillende beoordelaars, over 1.009 proefpersonen, bij de uitvoering van 60 verschillende opgaven. De overlap tussen beide datasets werd gevormd door beoordelaars die bij beide experimenten betrokken waren.

Op grond van de FACETS analyse kan worden vastgesteld dat de beoordelingen met een hoge mate van zekerheid zijn gegeven. De geschatte betrouwbaarheid van de waardering in CEF-niveau van de responsen van proefpersonen is 0.95. De dataset bevatte 3.777 paren beslissingen (beslissingen waar twee beoordelaars dezelfde respons van dezelfde proefpersoon beoordeelden). Deze werden in 56% van de gevallen door de betrokken beoordelaars met exact hetzelfde CEF-niveau gewaardeerd. Dat is ruim 11% beter dan verwacht gegeven het aantal beoordelingscategorieën. Slechts 8 beoordelingen toonden een afwijking groter dan de 1% overschrijdingskans. Dit betekent in de praktijk van deze dataset dat een beoordelaar meer dan anderhalve categorie afweek in zijn beoordeling dan op grond van de andere data in de dataset kon worden verwacht. Figuur 5.1 geeft de probabliteitscurven per scoringscategorie van het CEF. De figuur toont dat de scoringscategorieën vrij gelijkmatig in de data voorkwamen. De snijpunten van de curven geven de grenzen tussen de CEF-niveaus weer.

Zoals te verwachten is er bij deze grote dataset geen formele passing van de data bij het meetmodel. Figuur 5.2 geeft echter een beeld van de mate waarin de data in praktische zin passen bij het model. De dikke curve lopend van linksonder naar rechtsboven representeert het ‘model’, dat wil zeggen welke vaardigheid (x-as) bij welke scoringscategorie (y-as) nodig is. De haakse stippellijnen geven de grenzen aan van de CEF-niveaus. De dunnere lijnen die min of meer parallel lopen aan de dikke lijn van linksonder naar rechtsboven stellen de meetfout voor (5% overschrijdingskans). De meetfout is in de IRT gebruikelijk groter aan de uiteinden van een schaal dan meer naar het midden waar zich de meeste data bevinden en dus de meeste zekerheid bestaat over de schattingen. In dit geval is duidelijk dat de onzekerheid over het verwachte CEF-niveau met name groter wordt naar het boveneinde van de schaal. De grijze stippen geven de data weer: hoe dichter deze bij de dikke lijn vallen hoe beter de data passen bij het model. De figuur vertoont te weinig precisie om de bovengenoemde acht afwijkingen duidelijk te zien: de afwijkingen zijn zo gering dat zij op de meetfoutlijnen vallen.



Figuur 5.1: Probabiliteitscurven per scoringscategorie



Figuur 5.2: Geschatte vaardigheid (model) en geobserveerde waarden (data)

De waarden van de schattingen van de CEF-grenzen op de onderliggende vaardigheidsschaal met de bijbehorende meetfout staan afgedrukt in Tabel 5.3. Op grond van deze CEF-grenzen kan bepaald worden welk niveau op de CEF-schaal proefpersonen hebben volgens de menselijke oordelen. Gelet op de kwaliteit van deze oordelen kan toekenning van CEF-niveaus aan proefpersonen op grond van de menselijke oordelen worden aangewend om de grensscores op de TGN te bepalen.

Tabel 5.3 Geschatte grensscores van de CEF-niveaus op grond van menselijke beoordelingen

CEF-niveau	Theta-CEF bij ondergrens	Standaard Meetfout
A1-min	-8.13	0.07
A1	-5.96	0.05
A2	-2.69	0.06
B1	-0.28	0.08
B2	2.47	0.15
C1	5.70	0.25
C2	8.88	0.27

5.3 Bepaling van de grensscores op de TGN

Functies voor de transformatie van de onderliggende schalen voor de deelscores op de TGN naar de rapportageschalen van de deelscores zijn bepaald op grond van de samenhang van ieder van de deelschalen met de thetaschaal van het CEF. De eerste stap was de projectie van de ondergrenzen van de CEF-niveaus op de onderliggende schalen van ieder van de deelscores. Vervolgens zijn de onderliggende deelschalen door middel van regressiefuncties getransformeerd naar de rapportageschalen voor de deelscores. Door gebruik te maken van lineaire regressies blijft de interval-eigenschap van de thetaschaal behouden.

Bij het bepalen van een rapportageschaal is men vrij in het kiezen van de daarop voorkomende waarden. In het Nederlands onderwijs hanteert men veelal de bekende tienpuntschaal. In andere systemen worden schalen met waarden van 1 tot en met 5, van 0 tot en met 100 en tal van andere varianten gebruikt. De meeste schalen hebben het nadeel dat zij vanwege hun exacte onder- en bovenbegrenzing niet goed in staat zijn de interval-eigenschappen van IRT-schalen die in theorie lopen van min oneindig naar plus oneindig, te behouden. Voor de TGN is gekozen voor rapportageschalen lopend van 10 tot 80. De bedoeling is om daarmee uit te drukken dat er ook nog scores zijn onder de laagst gerapporteerde score en dat de hoogste score nog niet betekent dat men een 100% beheersing heeft. De laagst gerapporteerde score van 10 betekent dat men nagenoeg niet kan functioneren in het Nederlands, terwijl de hoogste score van 80 suggereert dat men in de praktijk geen enkel probleem heeft met de Nederlandse taal. De laagste ondergrens (A1-min) moet op enige afstand liggen van de laagst gerapporteerde score. Bij een rapportage schaal van 10 tot 80 betekent dit dat de CEF ondergrenzen moeten liggen in een scoregebied tussen 15 en 80.

Nadat lineaire transformatiefuncties zijn gevonden die aan deze voorwaarden voldoen, worden deze functies gebruikt om de deelscores op de thetaschalen om te zetten naar de rapportageschalen. Deelscores worden op de rapportageschalen in eerste instantie afgegrensd bij 0 en 90 ten einde te vermijden dat zeer extreme deelscores een oneigenlijk grote invloed krijgen op de totaalscore. De vier deelscores worden vervolgens ongewogen gecombineerd tot een totaalscore. Tenslotte worden de deelscores en de totaalscore bij de rapportage afgegrensd bij 10 en 80, omdat er geen reden bestaat buiten dat gebied onderscheid te maken en omdat de meetfout daar snel in grootte toeneemt.

De gevonden transformatiefuncties voor de omzetting van de onderliggende deelscores in de deelscores op de rapportageschalen zijn weergegeven in Tabel 5.4. Ter vergelijking zijn hierbij in de eerste kolom ook de transformatiefuncties gegeven zoals die op basis van uitsluitend de pretestdata zouden zijn geschat.

Tabel 5.4: Transformatiefuncties voor omzetting van de onderliggende deelscores naar de rapportageschalen

Functies op basis van analyse pretesten 2004	Nieuwe functies op basis van gecombineerde analyse: pretest en experimenten MFA-Fit 2005
Uitspr = 23.666 * thetaUitspr + 44.456	UitsprNieuw = 23.390 * thetaUitspr + 54.842
Vloei = 26.22 * thetaVloei + 43.549	VloeiNieuw = 23.154 * thetaVloei + 53.307
Zinsb = -36.39 * thetaZinsb + 36.454	ZinsbNieuw = -38.566 * thetaZinsb + 45.921
Woord = 20.285 * thetaWoord + 36.441	WoordNieuw = 19.740 * thetaWoord + 45.921

De plaats van CEF-ondergrenzen op de TGN rapportageschalen en de bijbehorende gemiddelde meetfout wordt gegeven in Tabel 5.5.

Tabel 5.5: Relatie CEF-Niveaus en TGN score

CEF-Niveaus	Rapportage schaal	Meetfout
A1-min	16	2.92
A1	26	3.01
A2	37	3.20
B1	47	3.40
B2	57	3.62
C1	68	3.84
C2	80	4.04

5.4 Het bepalen van zak-slaaggrenzen op de TGN

De Toets Gesproken Nederlands is ontwikkeld in opdracht van het Ministerie van Justitie. Het is aan de opdrachtgever om te bepalen welke eisen er precies aan toekomstige proefpersonen gesteld zullen worden en waar de grenzen tussen 'zakken' en 'slagen' dienen te liggen. De rol van de toetsconstructeurs is beperkt tot het leveren van de informatie die daarvoor nodig is. Deze paragraaf moet in dat kader worden gelezen.

Men noemt de door een proefpersoon op een toets behaalde score de *geobserveerde* score. Voor elke toets geldt dat de geobserveerde score niet hetzelfde is als de *ware* score van de proefpersoon. Immers door allerlei omstandigheden is het mogelijk dat een proefpersoon bij een tweede of derde toetsing een enigszins afwijkende score zal halen. De proefpersoon kan minder vermoeid zijn en daardoor beter presteren, of juist vanwege een hogere temperatuur in het examenlokaal een mindere prestatie leveren. De geobserveerde score geeft ons dus slechts een *schatting* van de ware score. Deze schatting kan zowel naar boven als naar beneden afwijken van de ware score. De mate van afwijking wordt de meetfout genoemd en staat in directe relatie met de betrouwbaarheid van de toets.

De betrouwbaarheid van de TGN is .94. De TGN doet wat dat betreft niet onder voor examens met een voor veel proefpersonen vergelijkbaar civiel effect, zoals bijvoorbeeld het Staatsexamen Nederlands als Tweede Taal. Tabel 5.6 geeft de betrouwbaarheden van de verschillende onderdelen van het staatsexamen weer zoals gerapporteerd door de staatsexamencommissie NT2 over het eerste afnamemoment in 2005.

Tabel 5.6: Betrouwbaarheid Staatsexamen NT2, Examenprogramma I en II

Examenonderdeel	Examenprogramma I afname januari 2005	Examenprogramma II afname maart 2005
Lezen	.82	.87
Schrijven	.93	.91
Luisteren	.84	.86
Spreken	.94	.95

Hoewel de betrouwbaarheid van de TGN met een waarde van 0.94 dus goed genoemd kan worden, is de meetfout van de schatting bij de ondergrens voor niveau A1-min toch nog 2.9 scorepunten en voor niveau A2 zelfs 3.2 scorepunten. De betekenis van die meetfouten illustreren we aan de hand van niveau A1-min. De ondergrens voor het beoogde niveau is ingeschaald op 16 scorepunten. Dit wil zeggen dat een ware score van 16 op de toets aangeeft dat de kans dat men zal kunnen voldoen aan de beschrijving van de vaardigheid op het niveau A1-min groter is dan de kans dat men er niet aan kan voldoen. De ware score wordt echter niet geobserveerd. Het examen levert immers een geobserveerde score. De ware score ligt met 68% zekerheid tussen de geobserveerde score min de meetfout en de geobserveerde score plus de meetfout. Bij een meetfout van 2.9 scorepunten en een geobserveerde score van 16, ligt de ware score dus met 68% zekerheid tussen de 13.1 en de 18.9. Wil men een grotere zekerheid, bijvoorbeeld 95%, dan moet men een scoregebied nemen van de geobserveerde score min en plus tweemaal de standaard meetfout. Bij een score van 16 is dat dus van 10.2 tot 21.8. Met driemaal de meetfout bereikt men 99% zekerheid.

Bij gebruikmaking van de grensscores als aangegeven in Tabel 5.5 dient men zich te realiseren dat de keuze voor het toekennen van het predikaat ‘geslaagd voor een bepaald niveau’ afhankelijk is van de overwegingen met betrekking tot de meetfout. Wil men het risico van een *on*terechte toekenning van het predikaat ‘geslaagd’ verkleinen, dan zal men de score waarbij dit predikaat wordt toegekend voor een examen op niveau A1-min kiezen op een punt boven de waarde van 16. Kiest men bijvoorbeeld voor een waarde die hier meer dan driemaal de meetfout boven ligt dan kan praktisch worden uitgesloten dat de ware vaardigheid van proefpersonen zou kunnen liggen onder de grens van 16. Het probleem is echter dat men bij een dergelijk beleid een groot aantal proefpersonen waarvan de ware vaardigheid daadwerkelijk boven de 16 ligt *ten onrechte* dat predikaat ‘geslaagd’ onthoudt.

Bij de meeste toepassingen van toetsen voor het nemen van beslissingen over personen wordt een mogelijk onterechte afwijzing als even ernstig beschouwd als een mogelijk onterechte toekenning en wordt daarom de grensscore gebruikt waarbij de kans op beide soorten van foute beslissingen gelijk is. In dit voorbeeld zou dan de grensscore voor A1-min gelegd worden bij 16 en de grensscore voor A2 op 37. Voor een meer gedetailleerde uitleg over de gevolgen van de keuze van verschillende cesuren, zie bijlage 13.

6 Betrouwbaarheid en validiteit

Toetsscores moeten aan twee belangrijke voorwaarden voldoen wanneer men ze met recht wil kunnen gebruiken voor het nemen van beslissingen over mensen. De scores moeten betrouwbaar zijn, dat wil zeggen dat de gebruiker erop moet kunnen vertrouwen dat de toetsresultaten niet afhankelijk zijn van het toeval en bij verschillende afnames stabiel zullen zijn. Daarnaast moeten de scores valide zijn, dat wil zeggen dat zij daadwerkelijk gerelateerd zijn aan hetgeen de toets beoogt te meten. Betrouwbaarheid en validiteit zijn relatieve begrippen. Het is niet mogelijk om daar absolute uitspraken over te doen. Wél kunnen gegevens worden verzameld die erop duiden dat aannamen omtrent de betrouwbaarheid en de validiteit van toetsscores redelijkerwijs niet kunnen worden verworpen. In dit hoofdstuk presenteren wij deze gegevens met betrekking tot de TGN. Hierbij gaan we eerst in op evidentie met betrekking tot de betrouwbaarheid, omdat betrouwbaarheid een essentiële voorwaarde is voor validiteit. Vervolgens komen een reeks van onderwerpen aan de orde die betrekking hebben op de validiteit.
Achtereenvolgens bespreken we evidentie voor:

- de juiste werking van de spraakherkenner die bij de automatische scoring wordt gebruikt;
- het vermogen van de toetsscores onderscheid maken tussen mensen die van huis uit Nederlands spreken en mensen die met andere talen als moedertaal zijn opgegroeid;
- de afwezigheid van onterechte samenhang tussen de toetsscores en achtergrondvariabelen zoals leeftijd, geslacht, en opleiding;
- de overeenstemming van de TGN score met menselijke oordelen over beheersingsniveaus en
- de samenhang tussen menselijke oordelen over taalvaardigheid en de aspecten van taalvaardigheid waarop de deelscores van de TGN zijn gebaseerd.

6.1 Betrouwbaarheid van de TGN scores

In het kader van de pretesten en de aanvullende experimenten zijn verschillende gegevens verzameld die een indicatie vormen voor de betrouwbaarheid van de TGN. Van de data verzameld in het kader van de pretest, is een set van 139 proefpersonen apart gehouden en niet betrokken in de training van de diverse componenten van de automatische scoring, de scoreberekening en de scoretransformatieprocedures. Daarnaast zijn er de gegevens die verzameld zijn in het experiment Amsterdam. Ook deze data zijn op geen enkele wijze betrokken in de ontwikkeling van enige component van het toetssysteem. Daardoor is het mogelijk op basis van deze data onafhankelijk na te gaan of alle componenten goed zijn getraind en om een indicatie te krijgen van de generaliseerbaarheid van de gevonden eigenschappen van het systeem.

Tabel 6.1 toont de betrouwbaarheidschattingen op basis van de ‘valideringset’ van 139 proefpersonen van de pretest die apart zijn gehouden. De proefpersonen in de pretest hebben, net zoals de examenkandidaten die de toets zullen maken wanneer die eenmaal operationeel is, ieder hun eigen individuele deelverzameling opgaven gemaakt. De betrouwbaarheid van de toets is geschat door voor iedere proefpersoon apart de split-half betrouwbaarheid te berekenen en hiervan het gemiddelde te nemen. De betrouwbaarheidschatting voor gehele examens, samengesteld uit de verzameling items die gebruikt is voor de pretest dat wil zeggen de totale verzameling vóórdat daaruit op basis van de pretestresultaten opgaven zijn geschrapt) bedraagt 0.94. De (split-half schatting voor de deelscores variëren bij de pretesten van 0.73 (voor Woordenschat) tot 0.93 (voor Zinsbouw). Voor zowel Uitspraak als Vloeiendheid komt de schatting uit op 0.89.

Tabel 6.1 Geschatte split-half betrouwbaarheid Valideringset Pretest (n=139 NMS)

Deelscores	Uitspraak	0.89
	Vloeiendheid	0.89
	Woordenschat	0.73
	Zinsbouw	0.93
Totaalscore		0.94

De betrouwbaarheidscoëfficiënt voor de totaalscore van de TGN is vergelijkbaar met die van andere gezaghebbende taaltoetsen zoals de Test of Spoken English (TSE) van Educational Testing Service (ETS) en de onderdelen spreken en luisteren van het Staatsexamen Nederlands als Tweede Taal. ETS rapporteert doorgaans een betrouwbaarheid van rond 0.90. Het Staatsexamen Nederlands als Tweede Taal doet het zelfs nog beter voor het onderdeel spreken met betrouwbaarheden van rond 0.93 - 0.95. Met betrekking tot het examenonderdeel Luisteren rapporteert de Staatsexamencommissie NT2 betrouwbaarheids-coëfficiënten die per afname variëren tussen 0.80 en 0.85.

Ook op basis van de aanvullende experimenten kunnen schattingen gemaakt worden van de toetsbetrouwbaarheid. In Tabel 4.15 zijn de test-hertest betrouwbaarheden gepresenteerd die bij de drie aanvullende experimenten zijn gevonden. Voor de vier deelscores van PSTN versies A en B zijn in Hoofdstuk 4 ook Cronbach alpha betrouwbaarheidschattingen vermeld. Deze zijn echter verspreid over de Tabellen 4.11 tot en met 4.14 en worden daarom hier bijeengebracht in Tabel 6.2.Voor de totaalscores zijn split-half schattingen berekend.

Tabel 6.2 Cronbach alpha betrouwbaarheid MFA-Fit (N=336 – 434)

		PSTN-A		PSTN-B	
		alpha	split-half	alpha	split-half
Deelscores	Uitspraak	0.84		0.84	
	Vloeiendheid	0.88		0.87	
	Woordenschat	0.77		0.74	
	Zinsbouw	0.95		0.95	
Totaalscore			0.91		0.91

Een aan de betrouwbaarheid gerelateerde maat voor de bepaling van de mate waarin men staat kan maken op toetsresultaten, is de meetfout. Bij de pretest zijn voor alle deelnemers aparte schattingen van de meetfout beschikbaar. De grootte van deze meetfout is afhankelijk van het op de thetaschaal geschatte verschil tussen de vaardigheid van de proefpersoon en de moeilijkheidsgraad van de verzameling items die door de proefpersoon zijn beantwoord. Men noemt dit de voorwaardelijke meetfout. De schatting is gebaseerd op de som van de informatiefuncties van alle opgaven over alle onderdelen die aan een deelnemer aan de toets zijn voorgelegd. Een voordeel van deze schatting van de meetfout is dat - in tegenstelling tot schattingen van de betrouwbaarheid - de maat niet afhankelijk is van de verdeling van de proefpersonen, maar lokaal op de thetaschaal bepaald wordt. Daarmee biedt deze grootheid direct informatie over de meetnauwkeurigheid bij bepaalde vaardigheidsniveaus. Een complicatie bij toetsen zoals de TGN die worden samengesteld door willekeurige trekkingen uit een itembank, is dat deelnemers aan examens ieder een unieke verzameling opgaven krijgen voorgelegd.

Dat betekent dat er geen unieke informatiefunctie is maar een bij voldoende grootte van de bank nagenoeg oneindige verzameling informatiefuncties. Daarom is de schatting van de individuele meetfout voor de 1522 proefpersonen bij de pretest als uitgangspunt genomen.

Tabel 6.3 toont een samenvatting van de schattingen bij de ondergrenzen van de CEF-niveaus. Deze zijn op twee wijzen berekend: door middel van een beste passende regressiefunctie (polynomiaal van de vierde orde) en door het gemiddelde te nemen van de meetfout bij alle proefpersonen met een score van drie punten onder de betreffende cesuur tot en met drie punten daarboven.

Tabel 6.3: Schattingen van de meetfout van TGN scores bij de ondergrenzen van de CEF-niveaus volgens twee schattingsmethoden

CEF-Niveaus	Via regressie	Gemiddelden
A1-min	2.92	2.82
A1	3.01	3.21
A2	3.20	3.14
B1	3.40	3.30
B2	3.62	3.49
C1	3.84	3.69
C2	4.04	3.95

Uit Tabel 6.3 blijkt dat beide schattingen onderling zeer vergelijkbare waarden geven. Naar de hogere niveaus neemt de meetfout geleidelijk toe met een vol punt op de scoreschaal. Dit betekent dat het betrouwbaarheidsinterval rond de cesuren op de TGN kleiner is bij de lagere niveaus. Dit is in overeenstemming met het design van de toets die immers is ontworpen om voornamelijk op de laagste niveaus te meten en die volgens de opdracht een betrouwbaarheid van minimaal 0.80 moest hebben bij de meting van vaardigheid vanaf A1-min tot minimaal B2 op de CEF-schaal.

6.2 Validiteit van de TGN scores

6.2.1 Functionele precisie van de spraakherkenner

Wanneer een systeem wordt geëvalueerd, is het erg belangrijk om maatstaven te gebruiken die relevant zijn voor het betreffende systeem. Wanneer men bijvoorbeeld een voertuig zou evalueren, zou men verschillende criteria hanteren, afhankelijk van wat het beoogde gebruik van dit voertuig is. Voor autoracen is snelheid bijvoorbeeld belangrijk terwijl bij goederentransport vooral laadcapaciteit van belang is. Bij de evaluatie dient men dus de functie van het te evalueren systeem in aanmerking te nemen. Bij de evaluatie van de spraaktechnologie die is aangewend in de TGN is 'functionele precisie' het meest geschikte criterium. Functionele precisie refereert aan het *effect* dat bepaalde soorten fouten hebben op de applicatie. Om functioneel precies te zijn is het niet nodig dat een spraakherkenner elk individueel woord correct transcribeert. Neem het voorbeeld van een applicatie die dient voor het verhandelen van aandelen. Als een klant zegt: "*Verkoop asjeblieft vandaag 1.000 aandelen IBM voor 72 dollar*" en het systeem transcribeert deze uiting als "*Verkoop alstublieft vandaag 1.000 IBM voor 72*", dan zal het systeem toch tot de juiste actie overgaan (1.000 IBM aandelen verkopen). In dit geval is de functionele precisie van het systeem perfect hoewel verscheidene woorden niet of niet correct zijn herkend. Als het systeem deze uiting echter transcribeert als '*Koop asjeblieft vandaag 1.000 aandelen IBM voor 72 dollar'*, dan zijn er wel méér woorden correct herkend, maar zal het systeem een potentieel zeer kostbare vergissing maken en 1000 IBM aandelen aankopen. In dit geval is de functionele precisie van het systeem slecht hoewel het maar één enkel woord incorrect heeft getranscribeerd. Functionele precisie negeert individuele woordfouten en neemt in plaats daarvan in overweging of al dan niet de juiste actie wordt ondernomen.

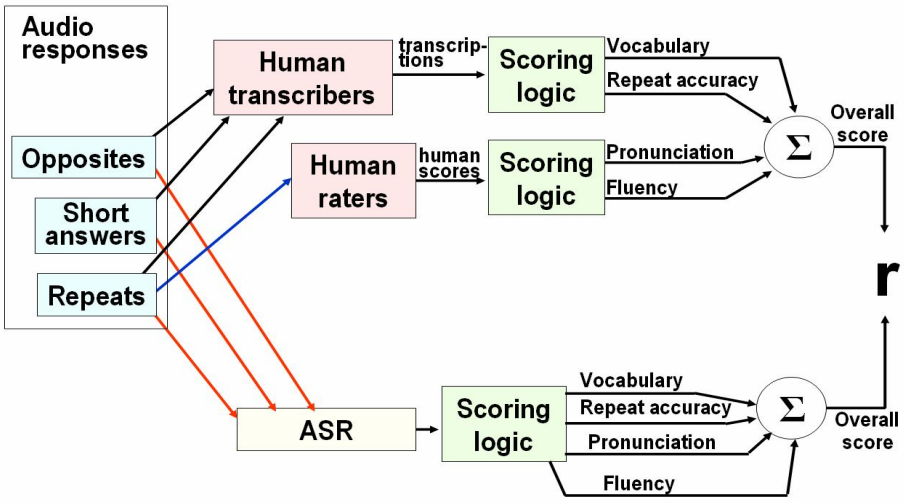
In het geval van een automatisch scoringsysteem voor een gesproken taaltoets is de actie van het systeem het toekennen van een toetsscore aan proefpersonen op basis van hun reacties op een aantal toetsopgaven. Dat betekent dat de functionele precisie van het scoringssysteem voldoende is, wanneer het scoringssysteem de juiste scores toekent aan proefpersonen.

6.2.1.1 Overeenstemming met menselijke beoordeling

Automatische spraakherkenners maken net zoals menselijke transcribeurs fouten. Automatische spraakherkenners maken zelfs méér fouten dan zorgvuldige menselijke transcribeurs (Lippman, 1996). Maar omdat de spraakherkenner in het scoringsysteem van de Toets Gesproken Nederlands onderdeel is van een complex van stochastisch geoptimaliseerde componenten (zie Figuur 2.5), moet de spraakherkenner worden geëvalueerd in het licht van het effect van fouten van de spraakherkenner op de scores die voor de Toets Gesproken Nederlands worden gerapporteerd.

Het effect van de invloed van de prestatie van de spraakherkenner op de toetsscores van proefpersonen die de Toets Gesproken Nederlands afleggen, kan worden onderzocht door de reacties van een groep NMS op de toets twee keer te scoren: één keer automatisch, en één keer op een wijze waarbij we die aspecten van de scoring die normaal via automatische spraakherkenning worden verkregen, vervangen door menselijke oordelen. Voor dit experiment hebben we opnieuw de responsen gebruikt van de 139 NMS die apart zijn gehouden en die niet in de ontwikkeling van het scoringsysteem waren betrokken.

Figuur 6.1 geeft een overzicht van de opzet van het experiment. Om het beeld te verhelderen worden een aantal onderdelen van het systeem die in Figuur 2.5 apart werden weergegeven, in Figuur 6.1 samengenomen. In Figuur 6.1 komt de component ASR (Augmented Speech Recognizer) overeen met de componenten *SR* en de *Acoustic and Language Models* van Figuur 2.5, terwijl *Scoring Logic* van Figuur 6.1 overeenkomt met de componenten *Content Scoring, Answer Model, Vocabulary IRT* en *Zinsbouw IRT* van Figuur 2.5.



Figuur 6.1: Onderzoek van de functionele kwaliteit van de spraakherkenner in de TGN

Aan de linkerkant van Figuur 6.1 zijn de responsen van proefpersonen op de drie opgavensoorten in de TGN afgebeeld *Opposites*, *Short answers* en *Repeats*.

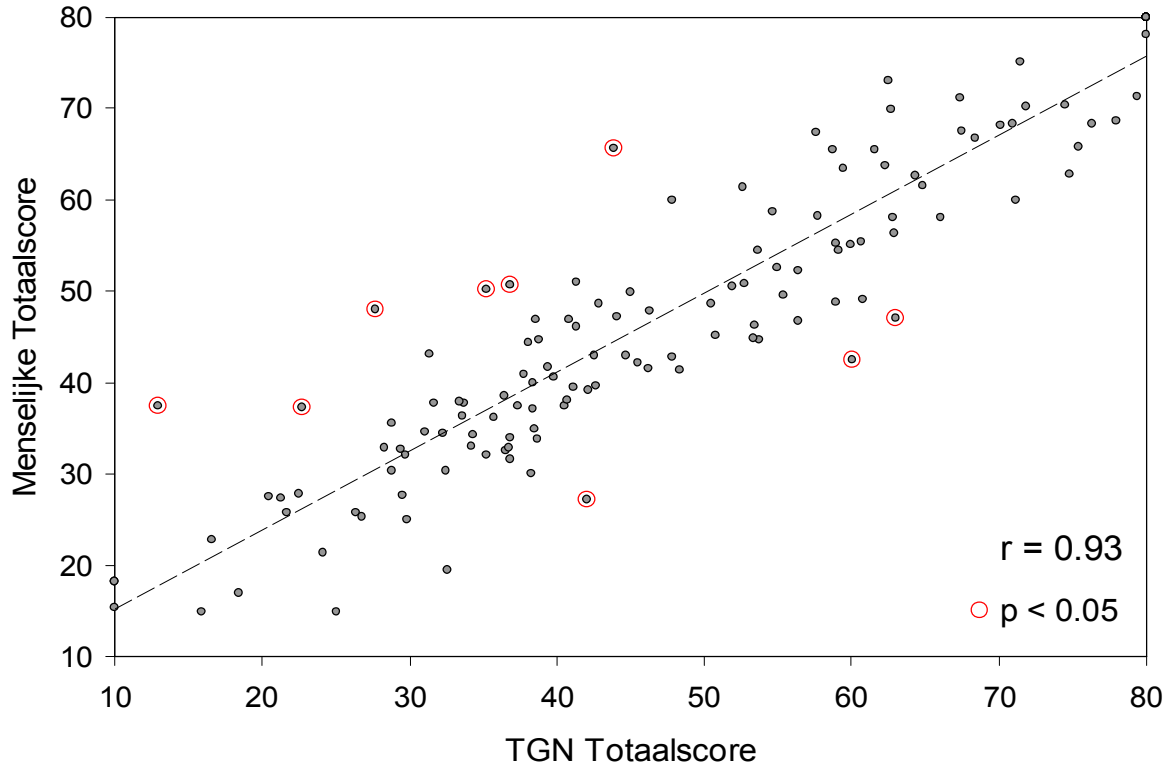
Voor de scoring gebaseerd op menselijke oordelen werden dezelfde procedures doorlopen als bij de ontwikkeling van het systeem zoals beschreven in paragraaf 2.10. Alle responsen van de 139 NMS werden handmatig getranscribeerd. De responsen op Herhaalopdrachten werden daarnaast door getrainde menselijke beoordelaars beoordeeld in twee aparte ronden: één keer op uitspraak en één keer op vloeiendheid. De beoordelaars maakten daarbij weer gebruik van Ordinate's telefonisch beoordelingssysteem. De 'menselijke oordelen' - transcripten en oordelen voor uitspraak respectievelijk vloeiendheid - werden vervolgens ingevoerd in de component *Scoring logic* waarmee op grond van deze input de scores voor de 139 NMS op normale wijze werden vastgesteld. Dezelfde responsen werden ook ingevoerd in de ASR en vervolgens gescoord in de component *Scoring logic*. Ten slotte werden de correlaties berekend tussen de deelscores en de totaalscores gebaseerd op menselijke oordelen en de scores gegenereerd via automatische scoring. De correlatie tussen totaalscores gebaseerd op menselijke oordelen en die op automatische scoring is 0.93. De correlaties voor de deelscores variëren van 0.80 voor uitspraak tot 0.94 voor zinsbouw. Tabel 6.4 geeft deze correlaties weer, alsmede de split-half betrouwbaarheidsschattingen voor beide sets van scores.

Tabel 6.4: Correlaties tussen ASR en menselijke scores en betrouwbaarheidsschattingen voor ASR en menselijke scores (n=139, pretest)

	Correlatie	Betrouwbaarheid (split-half)	
	ASR~Menselijke scores	ASR	Menselijke scores
Uitspraak	0.80	0.89	0.94
Vloeiendheid	0.84	0.89	0.92
Woordenschat	0.85	0.73	0.78
Zinsbouw	0.94	0.93	0.96
Totaalscore	0.93	0.94	0.96

De resultaten laten zien dat de testscores die tot stand komen op basis van automatische scoring nauw overeenkomen met testscores die gebaseerd zijn op het werk van menselijke beoordelaars en transcribeurs.

In het strooidiagram van beide sets van totaalscores in Figuur 6.2 zijn proefpersonen waarvan de twee soorten scores onderling significant verschillen gemarkeerd. Naast onnauwkeurigheid van één of beide wijzen van beoordeling, kunnen hier ook andere oorzaken aan ten grondslag liggen. Verschillende menselijke beoordelaars hebben bijvoorbeeld genoteerd dat proefpersonen soms duidelijk werden voorgezegd door hun docent. In dergelijke gevallen zullen menselijke beoordelaars geneigd zijn de respons negatief te scoren. Anderzijds noteerden zij soms ook achtergrondgeluiden zoals blaffende honden of huilende baby's. In die gevallen kan mogelijk de menselijke beoordelaar nog wel het antwoord van de proefpersoon van de achtergrond onderscheiden waar de machine dit niet meer kan. Dergelijke verschijnselen zullen in examensituaties uiteraard worden uitgesloten.



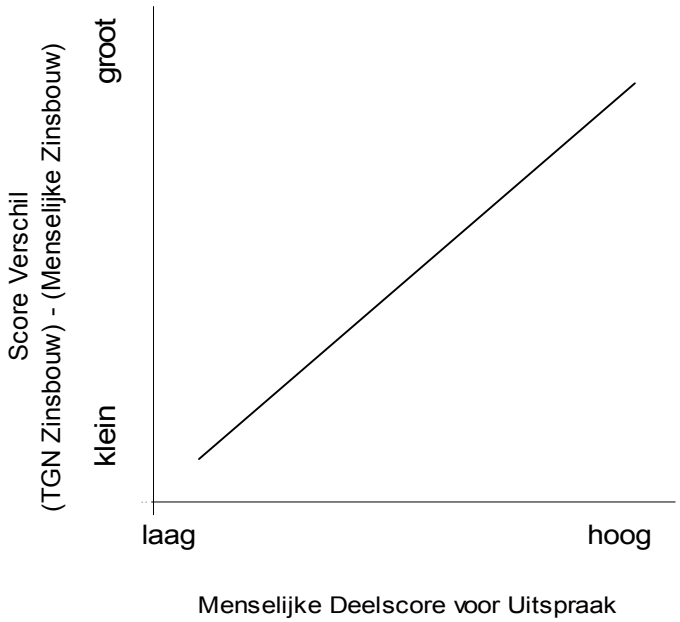
Figuur 6.2: Correlaties tussen TGN en menselijke totaalscores voor NMS (n=139, pretest)

6.2.1.2 Effecten van verschillen in uitspraak

Hoewel de gegevens in de vorige paragraaf aangeven dat de inzet van automatische spraakherkenning bij de scoring van de reacties van kandidaten op de opgaven van de TGN verantwoord is, ligt het voor de hand om te veronderstellen dat de inzet van automatische spraakherkenning nadelig is voor kandidaten met een sterk afwijkende uitspraak van het Nederlands. Om de invloed van de automatische spraakherkenner op de scores van kandidaten met uiteenlopende moedertalen nader te onderzoeken, hebben we de volgende contrahypothese onderzocht:

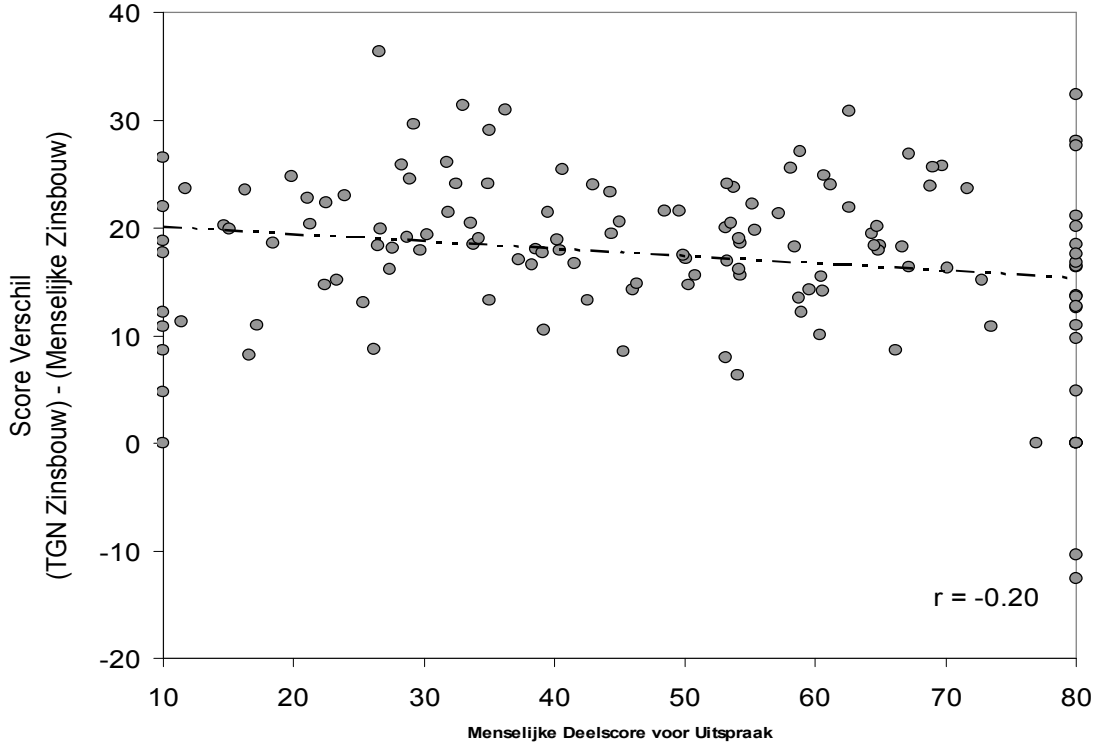
Sterke buitenlandse accenten veroorzaken extra fouten van de automatische spraakherkenner en hebben een negatief effect op de deelscores voor inhoud, te weten Woordenschat en Zinsbouw.

De contrahypothese voorspelt dat een proefpersoon meer nadeel zal hebben van de inzet van automatische spraakherkenning naarmate zijn uitspraak van het Nederlands slechter is. Het nadelige effect zou zijn scores voor Woordenschat en Zinsbouw negatief beïnvloeden doordat de spraakherkenner zijn correcte antwoorden niet goed herkent.
Om deze contrahypothese te toetsen, werd opnieuw de dataset gebruikt van de 139 NMS die niet betrokken waren bij de ontwikkeling van het scoringsmodel. Er werd gekeken naar het verschil tussen hun deelscores voor Zinsbouw gebaseerd op de output van de spraakherkenner en hun deelscores voor Zinsbouw gebaseerd op menselijke transcripties als een functie van de deelscore voor Uitspraak gebaseerd op menselijke beoordelingen. Indien de contrahypothese waar is, zouden we een verband verwachten zoals weergegeven in Figuur 6.3.



Figuur 6.3. Voorspelde trend op basis van contrahypothese

Figuur 6.4 geeft een strooidiagram van de geobserveerde scores. De contrahypothese wordt niet bevestigd. In plaats van het voorspelde beeld zien we zelfs een - weliswaar zeer zwakke - trend tegengesteld aan de voorspelde trend op grond van de contrahypothese.



Figuur 6.4: Toetsing van de contrahypothese met betrekking tot de invloed van sterk buitenlands accent (NMS, n=139)

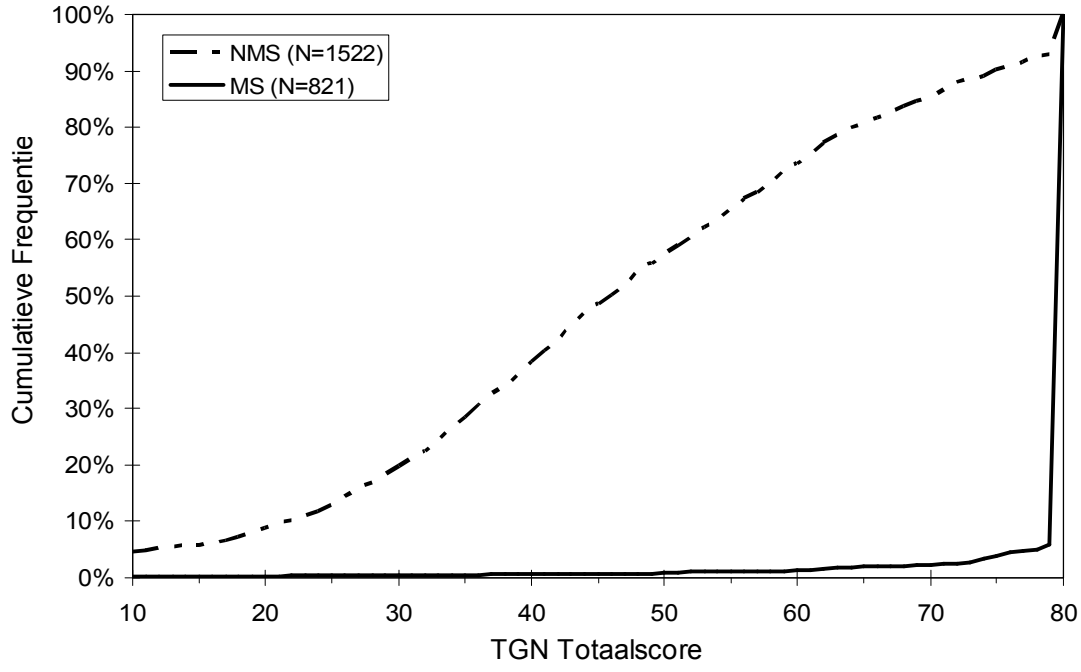
Samenvattend kan gesteld worden dat de onderzoeksresultaten laten zien dat de testscores die tot stand komen op basis van de output van de spraakherkenner nauw overeenkomen met testscores die gebaseerd zijn op het werk van menselijke beoordelaars en transcribeurs. Deze bevindingen bevestigen de functionele precisie van het automatische scoringssysteem.

6.2.2 Relatie toetsscores en beheersing van het Nederlands

6.2.2.1 Totaalscores en deelscores voor MS en NMS

Tijdens de pretests hebben zowel moedertaalsprekers van het Nederlands als sprekers van het Nederlands als tweede taal de toets afgelegd (Zie hoofdstuk 3). Een eerste eis die men mag stellen aan een toets die beoogt te meten of personen voldoende vaardigheid hebben om deel te nemen aan gesprekken in een voor hen vreemde of tweede taal, is dat de toets daadwerkelijk onderscheid maakt tussen personen die deze vaardigheid beheersen en personen die deze vaardigheid nog niet of in beperkte mate beheersen.

Figuur 6.5 toont de procentuele cumulatieve frequentie voor NMS en MS zoals verzameld tijdens de pretests. De figuur laat duidelijk zien dat MS vrijwel uitsluitend zeer hoge scores halen. Slechts 2% behaalt een score lager dan C1 (68) en 6% een score lager dan C2 (80). Van de NMS scoort 84% juist lager dan C1 (68). Circa 6% van de NMS in de steekproef van de pretest behaalt een score onder A1-min. De pretesteekproef bevatte weinig proefpersonen op de lage niveaus. In de steekproeven Amsterdam en MFA-Fit is dit percentage duidelijk hoger.



Figuur 6.5: Cumulatieve frequenties voor NMS en MS (pretests)

Tabel 6.5 toont boven de diagonaal de onderlinge correlaties van de deelscores behaald door de NMS. Op de diagonaal zijn de betrouwbaarheidsschattingen weergegeven en onder de diagonaal de correlaties gecorrigeerd voor attenuatie.

Tabel 6.5: Intercorrelaties deelscores (n=1522) (pretests)

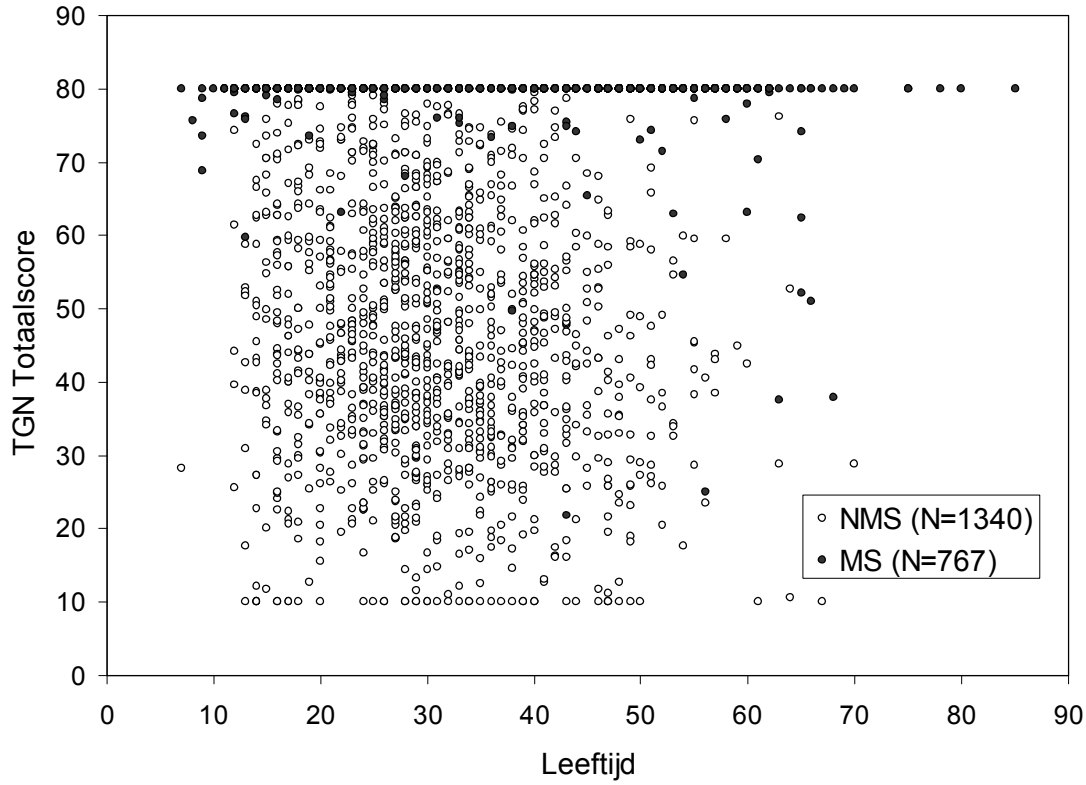
	Totaalscore	Uitspraak	Vloeiendheid	Woordenschat	Zinsbouw
Totaalscore	**0.94**	0.90	0.90	0.81	0.85
Uitspraak	*0.98*	**0.89**	0.89	0.59	0.62
Vloeiendheid	*0.99*	*1.00*	**0.89**	0.55	0.67
Woordenschat	*0.98*	*0.73*	*0.68*	**0.73**	0.68
Zinsbouw	*0.91*	*0.69*	*0.73*	*0.82*	**0.93**

De Tabel laat zien dat de afzonderlijke deelscores vanwege de partiële overlap hoog tot zeer hoog met de totaalscore correleren. De intercorrelaties tussen de deelscores onderling tonen echter aan dat ieder van de deelscores een groot deel eigen variantie heeft. De maten voor inhoudelijke correctheid hangen onderling nauwer samen dan met de afzonderlijke maten voor de kwaliteit van de spraak, zoals ook de maten voor kwaliteit van de spraak onderling nauwer samen hangen dan met de afzonderlijke maten voor correctheid.

6.2.3 Relatie toetsscores en achtergrondvariabelen

6.2.3.1 Resultaten naar leeftijd (MS en NMS)

Aangezien taaltoetsen over het algemeen teksten bevatten en teksten bepaalde onderwerpen hebben, vormt algemene kennis een van de meest bedreigende ongewenste variabelen bij taaltoetsing. Figuur 6.6 toont de toetsscore in relatie tot leeftijd. Duidelijk is te zien dat ongeacht de leeftijd MS een grote kans maken een maximum score te behalen. Net als op de definitieve scoreschaal zijn in deze figuur alle scores boven 80 als de maximale score van 80 weergegeven. Er zijn wel enkele uitzonderingen, maar voor hen kan gebrek aan inzet uiteraard niet worden uitgesloten. Bij zeer hoge leeftijd (>70 jaar) lijkt de kans op een maximumscore enigszins af te nemen. Het zijn echter maar enkele gevallen. Anderzijds maken NMS ongeacht hun leeftijd de meeste kans op scores tussen 20 en 70, met enkelen die daarboven scoren. Aangezien mag worden aangenomen dat MS en NMS over het algemeen verschillen in vaardigheid Nederlands en anderzijds personen variërend in leeftijd van 10 tot 70 aanmerkelijk in algemene kennis zullen verschillen, kan vooralsnog worden aangenomen dat vaardigheid in het Nederlands aanmerkelijk meer invloed op de scores heeft dan algemene kennis.



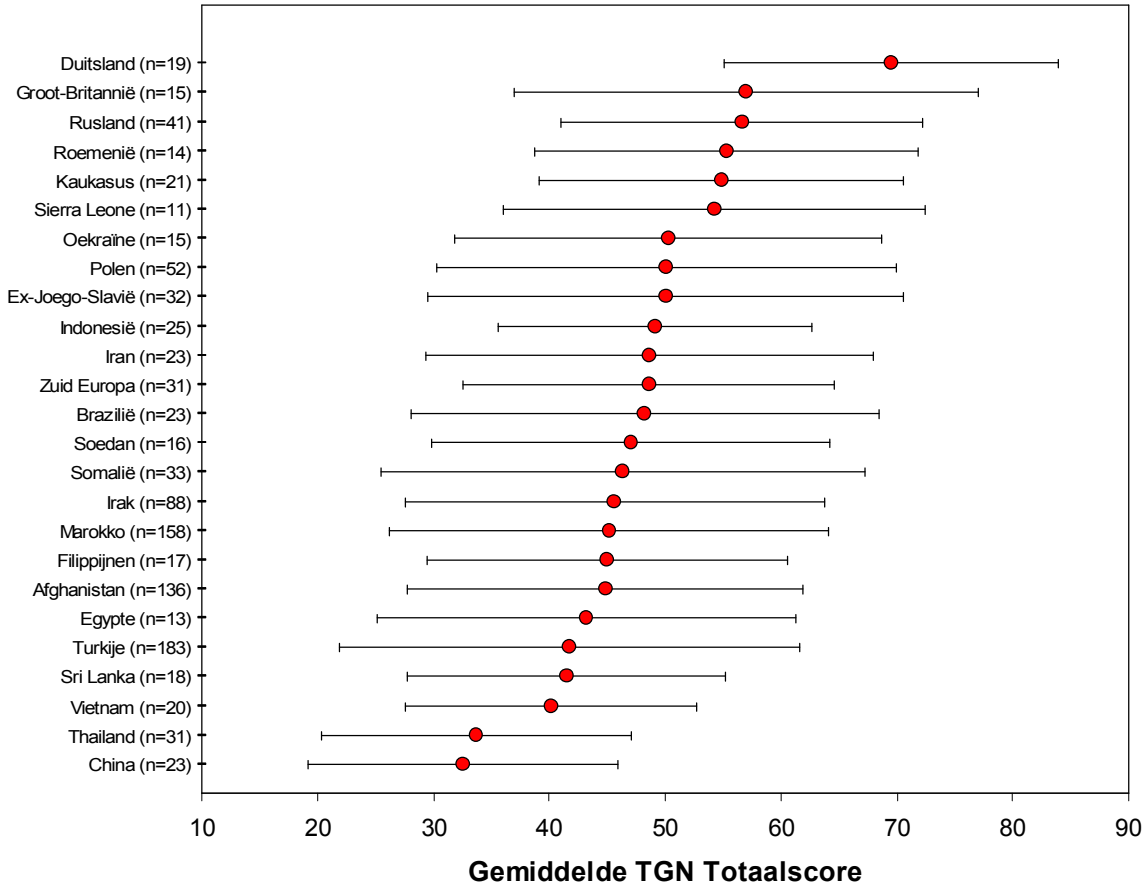
Figuur 6.6: Toetsscores in relatie tot leeftijd

6.2.3.2 Resultaten naar geslacht (NMS)

De resultaten van de pretests laten zien dat de toetsscores van mannen en vrouwen nauwelijks verschillen. De mannen (n=483) die de toets aflegden, behaalden gemiddeld een score van 44.3 (±18.7), de vrouwen (n=858)behaalden gemiddeld een score van 42.8 (±19.1).

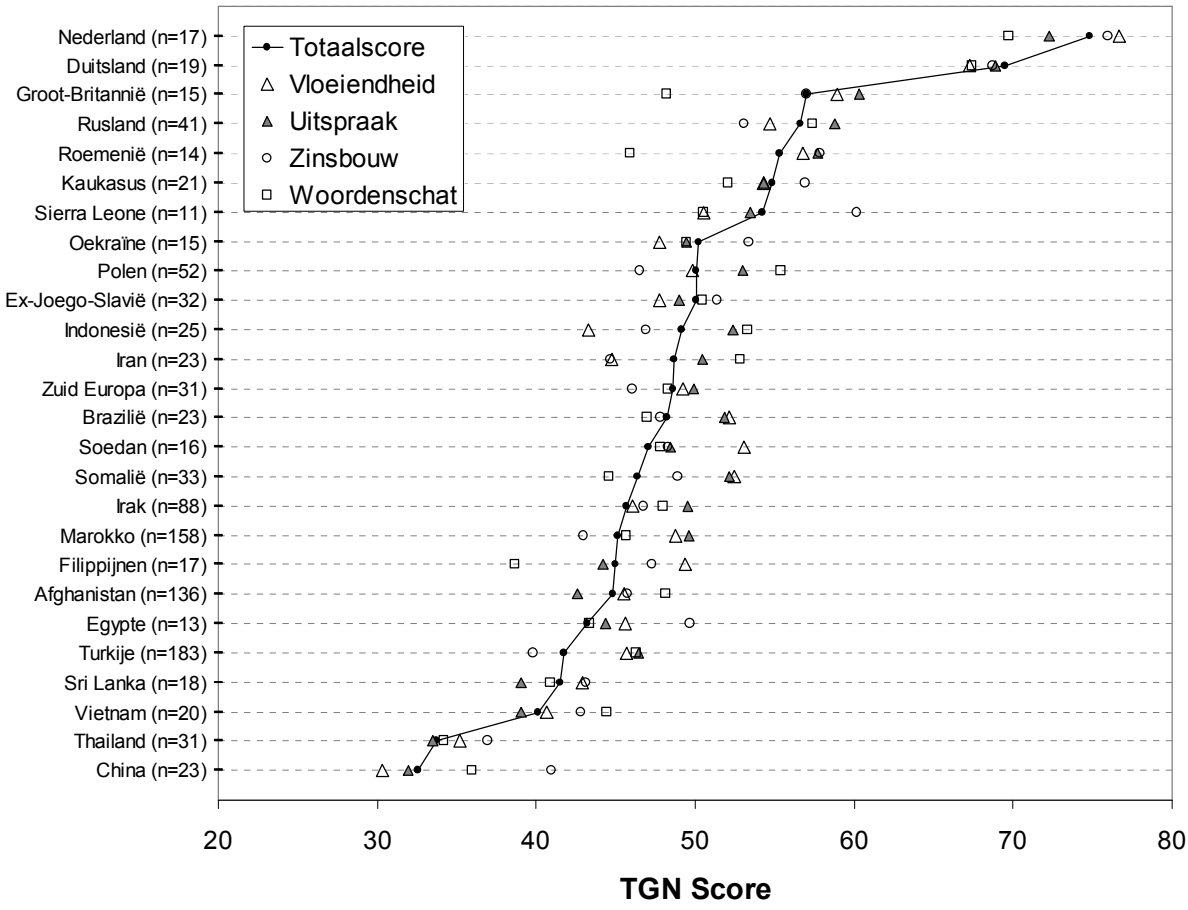
6.2.3.3 Resultaten van NMS naar land van herkomst (NMS)

Figuur 6.7 presenteert de gemiddelde score van de pretestkandidaten per land van herkomst plus en min twee standaarddeviaties, waarmee de scoreverdeling van 95% van de betrokken subgroepen is gedekt. De groepen verschillen in grootte: traditionele migrantengroepen en migranten uit landen met recente of actuele oorlogshistorie zijn meer vertegenwoordigd dan andere landen. In de figuur zijn uitsluitend landen opgenomen waaruit meer dan 10 kandidaten afkomstig waren. De landen zijn geordend naar gemiddelde score. De landen vertonen allemaal een grote spreiding. Hierdoor is er veel overlap tussen de verdeling van de verschillende landen. Het is dus niet zo dat beheersing van het Duits of het Engels zonder meer garantie op een hoge score biedt.



Figuur 6.7: TGN scores per land (pretest)

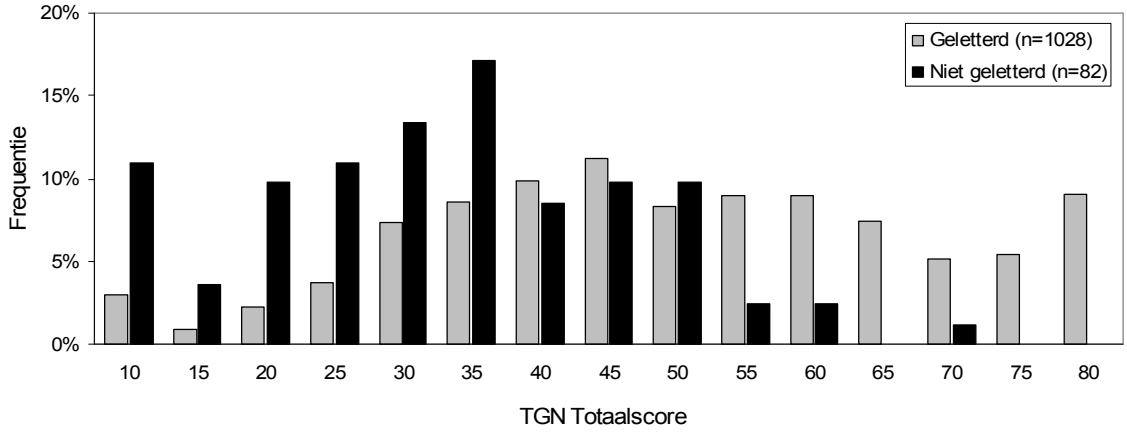
Figuur 6.8 toont de deelscores voor Uitspraak, Vloeiendheid, Woordenschat en Zinsbouw die tijdens de pretesten zijn behaald door proefpersonen uit verschillende landen. De doorgetrokken lijn geeft de gemiddeld behaalde totaalscore voor de toets. Per land van herkomst verschilt de bijdrage van de deelscores aan deze totaalscore. Een deelscore rechts van het gemiddelde duidt op een relatief gemak van de deelvaardigheid voor de betrokken groep, links van de lijn betekent juist dat de vaardigheid voor deze deelnemers moeilijk was. De groepen zijn uiteraard te klein om voor bepaalde nationaliteiten te kunnen generaliseren. Wel ondersteunen deze resultaten de onafhankelijkheid van de deelscores.



Figuur 6.8: Deelscores TGN per land (pretest)

6.2.3.4 Toetsscore in relatie tot geletterdheid

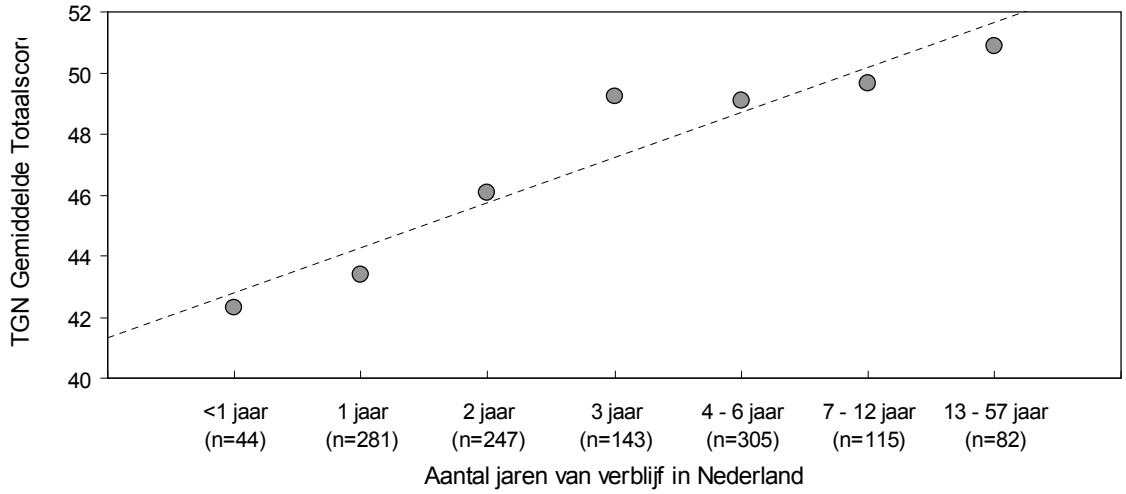
Gelet op de eerst beoogde doelgroep voor de toets (mensen die nog woonachtig zijn in het land van herkomst) is besloten dat geletterdheid geen rol mocht spelen bij de examinering Nederlandse taal. Figuur 6.9 geeft de scoreverdeling apart voor geletterden en niet geletterden die deelnamen aan de pretesten. Uit deze grafiek blijkt dat niet geletterden weliswaar gemiddeld een lagere score behalen dan geletterden, maar dat ook niet geletterden wel degelijk kans maken een hoge toetsscore te behalen.



Figuur 6.9: Scoreverdeling NMS voor geletterden en niet-geletterden (pretest)

Aan de aanvullende experimenten 'Amsterdam' en 'MFA-Fit hebben te weinig analfabeten deelgenomen om afzonderlijke analyses mogelijk te maken.

6.2.3.5 Toetsscores naar jaren van verblijf

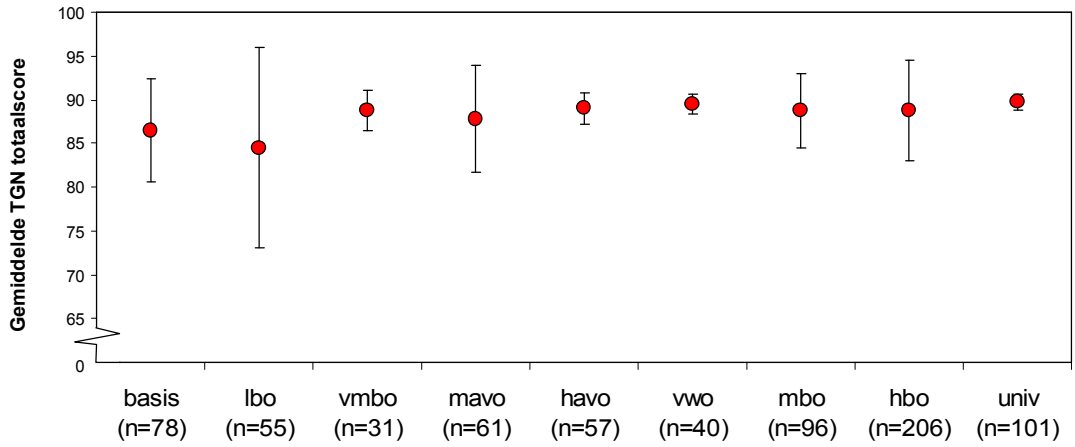


Figuur 6.10: Toetsscores NMS naar jaren van verblijf (pretest)

Figuur 6.10 geeft de relatie weer tussen toetsscore en aantal jaren van verblijf in Nederland. Uit de figuur blijkt dat de in de toets gemeten vaardigheid in de eerst drie jaar toeneemt (weliswaar in zeer geringe mate).Vanaf 3 tot en met 12 jaar laat de toetsscore nauwelijks of geen progressie zien. Pas na een zeer langdurig verblijf ontstaat een verband met verdere scoretoename.

6.2.3.6 Toetsscore in relatie tot hoogst genoten opleiding (MS)

In Figuur 6.11 worden gegevens weergegeven zoals verzameld in de pretesten over moedertaalsprekers. Per opleidingsniveau is met verticale balkjes de gemiddelde toetsscore $\pm$ 2 standaard deviaties aangegeven. De gemiddelde toetsscores liggen alle boven de maximum score en zouden in de praktijk als 80 worden gerapporteerd. De figuur toont geen verband tussen de toetsscore en het opleidingsniveau van moedertaalsprekers.



Figuur 6.11: Verdeling toetsscores MS naar opleiding (pretests)

6.2.3.7 Toetsscore in relatie tot thuis gesproken taal (MS)

Deelnemende MS is tijdens de pretesten gevraagd naar de door hen thuis gesproken taal: dialect of Algemeen Nederlands (AN). Tabel 6.6 geeft de gemiddelde scores en de standaarddeviaties voor beide groepen.

Tabel 6.6: Scores standaardsprekers versus dialectsprekers

Taal thuis	n	Gemiddeld	StDev
Dialect	42	84.89	12.66
Nederlands	713	88.04	5.83

Tabel 6.6 toont dat er een gering maar wel significant ($p<.05$) verschil in gemiddelde score is tussen beide groepen. De gemiddelde scores voor beide groepen zijn echter boven de beoogde maximale score van 80. De verschillen zijn daarom niet relevant.

6.2.4 Overeenstemming TGN-score en menselijke oordelen

Evidentie dat een toets meet wat deze beoogt of dient te meten kan ook worden verzameld door te onderzoeken in hoeverre de toetsresultaten overeenstemmen met die van andere beoordelingsprocedures gericht op het meten van dezelfde vaardigheid. Het probleem hierbij is echter dat er geen instrument bestaat zoals de platina meter in Parijs waarmee een nieuw ontwikkeld instrument kan worden vergeleken. Alle bestaande instrumenten zijn behept met een meetfout. Bij gebrek aan overeenstemming van de instrumenten die in de vergelijking worden betrokken, is het meestal niet mogelijk te bepalen aan welk instrument dit gebrek aan overeenstemming moet worden toegeschreven.

Gedurende het ontwikkeltraject van de TGN zijn vele duizenden oordelen over de spreek- en luistervaardigheid Nederlands van meer dan 2000 verschillende NMS verzameld. Pogingen om resultaten behaald op gestandaardiseerde instrumenten te verzamelen zijn gestrand op de verwarring omtrent de registratie, de codering en de interpretatie van deze resultaten bij de gebruikers van deze instrumenten. Met meer succes zijn oordelen verzameld die werden gegeven door docenten/begeleiders van proefpersonen op grond van globale of meer gestructureerde beoordelingsprocedures aan de hand van de niveaudefinities van het CEF. Veel van deze oordelen zijn ingezet voor de schaling en normering van de TGN. Nu deze schaling is afgerond, kunnen niet eerder gebruikte verzamelingen van oordelen worden ingezet voor validatie. Hierbij concentreren we ons op evaluatie van beslissingen rond het A1-min niveau omdat beoordelingen op dit zeer lage niveau het meest problematisch zijn.

Drie sets van oordelen komen hiervoor in aanmerking:

- de globale oordelen van docenten over het niveau van hun cursisten die als proefpersonen deelnamen aan de pretest;
- de globale oordelen van docenten over cursisten die deelnamen aan het experiment Den Haag en
- de op grond van gestructureerde interviews gegeven oordelen door docenten en begeleiders in Amsterdam.

De oordelen van de pretestdocenten en de docenten in het experiment Den Haag enerzijds en de oordelen uit Amsterdam anderzijds onderscheiden zich doordat het bij de eerste om ‘globale indrukken’ gaat, terwijl de laatste oordelen zijn gebaseerd op gestructureerde en geprotocolleerde interviews. Een tweede onderscheid betreft het feit dat de docenten in Amsterdam een intensieve training hadden gevolgd en vertrouwd waren met de beoordelingscriteria en de niveaudefinities van het CEF. Er mag daarom worden verondersteld dat de oordelen van de docenten in Amsterdam een hogere kwaliteit hebben dan de beide andere verzamelingen oordelen.

Wanneer nu de relatieTGN – docentenoordelen verzameld in Amsterdam duidelijk hoger is dan die in beide andere sets, dan heeft de TGN het vermogen om evidentie te leveren voor het veronderstelde verschil in kwaliteit en kan dit alleen maar het geval zijn wanneer de TGN de vaardigheid meet die in de docentenoordelen ligt besloten. Immers de TGN fungeert in deze vergelijking als standaard waarmee het verschil in kwaliteit wordt blootgelegd.

Tabel 6.7 toont de correlaties tussen de TGN en drie sets van oordelen van docenten die onafhankelijk van de TGN zijn gegeven en die ook niet zijn gebruikt bij de ontwikkeling van de TGN.

Tabel 6.7: Relatie TGN en Docentenoordelen

Dataset	n docenten	Training CEF	Beoordeling	n cursisten	correlatie met TGN
Pretest	92	Beperkt/niet	Globaal	1245	0.62
Den Haag	ca 10	Beperkt/niet	Globaal	95	0.56
Amsterdam	19	Intensief	Gestructureerd	76	0.70

Tabel 6.7 toont in alle datasets een positieve correlatie tussen oordelen van docenten gebaseerd op het CEF en de resultaten door proefpersonen behaald op de TGN. De correlatie is het hoogst met de best getrainde docenten die op gestructureerde wijze hun oordeel hebben uitgebracht. Dat wil zeggen dat de TGN beter met de oordelen van docenten overeenstemt naarmate de oordelen van de docenten beter de CEF-niveaus representeren. Het positieve resultaat in Amsterdam is met name opmerkelijk omdat de steekproef een zeer geringe spreiding had: 93% van de proefpersonen bevond zich in het vaardigheidsinterval van <A1-min t/m A2.

De TGN is in de eerste plaats ontwikkeld om op twee niveaus van het CEF de beslissing te kunnen nemen of kandidaten dat CEF-niveau al dan niet beheersen. De eerste toepassing van de TGN zal meest waarschijnlijk een beoordeling inhouden over de beheersing van het niveau A1-min. Later zal de TGN ook voor A2 en mogelijk andere niveaus worden ingezet.

We gaan eerst na hoe de kwaliteit van een oordeel over het beheersen van een zeer laag niveau zoals A1-min met gebruikmaking van de automatische procedures van de TGN, zich verhoudt tot een menselijke beoordeling omtrent het beheersen van dat niveau. Tabel 6.8 toont de mate van overeenstemming tussen een aantal gepaarde beoordelingen. De paren betreffen de confrontaties mens-mens, mens-machine en machine-machine. Alle oordelen zijn verzameld in het experiment Amsterdam en betreffen dezelfde 228 proefpersonen waarvoor al deze oordelen beschikbaar waren. In alle paren zijn de oordelen volledig onafhankelijk van elkaar tot stand gekomen en zijn de oordelen gebaseerd op binnen de betreffende paren verschillende prestaties van dezelfde proefpersonen. De data zijn ook niet betrokken bij ontwikkeling en schaling van de TGN .

De eerste twee kolommen in Tabel 6.8 vermelden de leden van het betreffende paar die de oordelen hebben gegeven. De volgende vier kolomen geven achtereenvolgens aan of beide beoordelaars oordelen dat de proefpersoon het niveau A1-min beheerst (1-1), of de eerste beoordelaar meent dat de proefpersoon A1-min beheerst en de tweede niet (1-0) of andersom (0-1) of beide beoordelaars zijn van mening dat de proefpersoon het niveau A1-min *niet* beheerst (0-0). De laatste kolom sommeert over de kolommen 1-1 en 0-0 en indiceert daarmee in welke mate beide beoordelaars tot dezelfde beslissing komen. De eerste twee paren betreffen de oordelen gegeven door één van beide bij het gestructureerde interview betrokken docenten als eerste lid en de docent die het loopbaangesprek voerde als tweede lid. Binnen de paren zijn de proefpersonen dus in een andere situatie op grond van een ander gesprek beoordeeld.

Voor het derde paar is de hoogste van de drie menselijke oordelen afgezet tegen de hoogste van de twee TGN-scores. Het laatste paar betreft twee verschillende versies van de TGN.

Tabel 6.8: Overeenstemming bij cesuur A1-min mens-mens, mens-machine en machine-machine

Oordeel 1	Oordeel 2	1 - 1	1 - 0	0 - 1	0 - 0	% Eens
Interviewer	Loopbaan	64%	10%	9%	17%	81%
Beoordelaar	Loopbaan	64%	9%	11%	16%	80%
Mens–Max (van 3)	TGN-Max. (van 2)	70%	14%	7%	8%	78%
TGN-1	TGN-2	58%	7%	12%	23%	81%

Uit Tabel 6.8 blijkt dat de verschillende paren nagenoeg hetzelfde niveau van overeenstemming bereiken, ongeacht de leden van het paar. Dit betekent dat goed getrainde menselijke beoordelaars die op grond van een gestructureerd interview of van een functioneel loopbaangesprek een oordeel vormen over de taalvaardigheid van een proefpersoon niet beter met elkaar overeenstemmen dan met een automatisch gegenereerde toetsscore en dat deze automatische score de door mensen gegeven beoordeling even goed voorspelt als twee beoordelaars elkaar kunnen voorspellen. Overigens moet opgemerkt worden dat op basis van deze gegevens niet kan worden bepaald welk lid van ieder paar 'gelijk' heeft of vaker gelijk heeft: de ware vaardigheid van de proefpersonen is immers onbekend. Wel kan worden gesteld dat de geheel verschillende operationalisaties – automatische toets en menselijke oordelen - een even grote overlap in de oordelen over taalvaardigheid vertonen als de menselijke oordelen onderling. Ondanks de verschillende operationalisaties moet er dus voor een groot deel hetzelfde worden gemeten.

In Tabel 6.9 presenteren we dezelfde gegevens, echter nu voor de cesuur bij A2. Zij geven eenzelfde beeld te zien. De zekerheid van de menselijk beoordelaars onderling neemt iets af, waardoor ook de overeenstemming tussen mens en machine wat afneemt.

Tabel 6.9: Overeenstemming bij cesuur A2 mens-mens, mens-machine en machine-machine

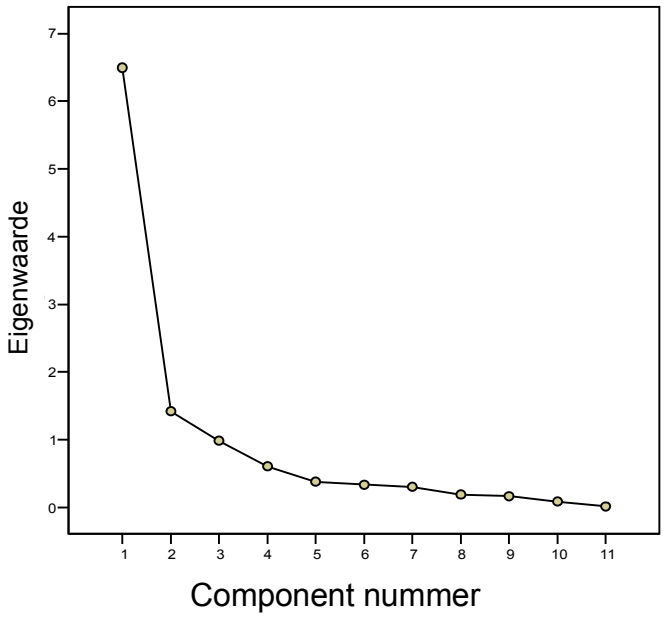
Oordeel 1	Oordeel 2	1 - 1	1 - 0	0 - 1	0 - 0	% Eens
Interviewer	Loopbaan	36%	12%	9%	44%	79%
Beoordelaar	Loopbaan	35%	11%	10%	44%	79%
Mens–Max (van 3)	TGN-Max. (van 2)	30%	18%	8%	45%	75%
TGN-1	TGN-2	24%	4%	9%	62%	86%

De overeenstemming tussen de menselijke oordelen en de automatische beoordeling met de TGN is over de gehele schaal hoog. Blijkbaar bestaat er een overlap tussen hetgeen mensen op grond van functionele gesprekken in hun oordeel betrekken en de vaardigheden die worden gemeten met de TGN.

6.2.5 Samenhang menselijke oordelen en aspecten van taalvaardigheid in de TGN

Om te achterhalen waar de in de vorige paragraaf genoemde overlap uit zou kunnen bestaan, zijn de data van het experiment Amsterdam nader onderzocht met behulp van een factoranalyse. In totaal zijn 11 metingen in de analyse ondergebracht: vier deelscores van toetsmoment 1, vier deelscores van toetsmoment 2, de twee beoordelingen bij het interview en de beoordeling bij het loopbaangesprek.

Uit de grafiek van de initiële eigenwaarden in Figuur 6.12 komen twee componenten met een eigenwaarde groter dan één naar voren.



Figuur 6.12 Initiële eigenwaarden mens- en machine oordelen

De ladingen na extractie volgens principale componenten analyse staan afgedrukt in Tabel 6.10. De ladingen op de eerste component zijn voor alle maten ongeveer gelijk. Ten opzichte van de tweede component ontstaan de verschillen. De analyse is nog hypothesevormend. Men zou kunnen veronderstellen dat de eerste component effectiviteit in communicatie representeert en dat de tweede component accuratesse in de uitspraak toevoegt. Een dergelijke interpretatie bevestigt de aanname in paragraaf 2.4 waar een onderscheid werd gemaakt in de beheersing van de linguïstische code en beheersing van de fonologie die gezamenlijk bijdragen aan succesvol communiceren.

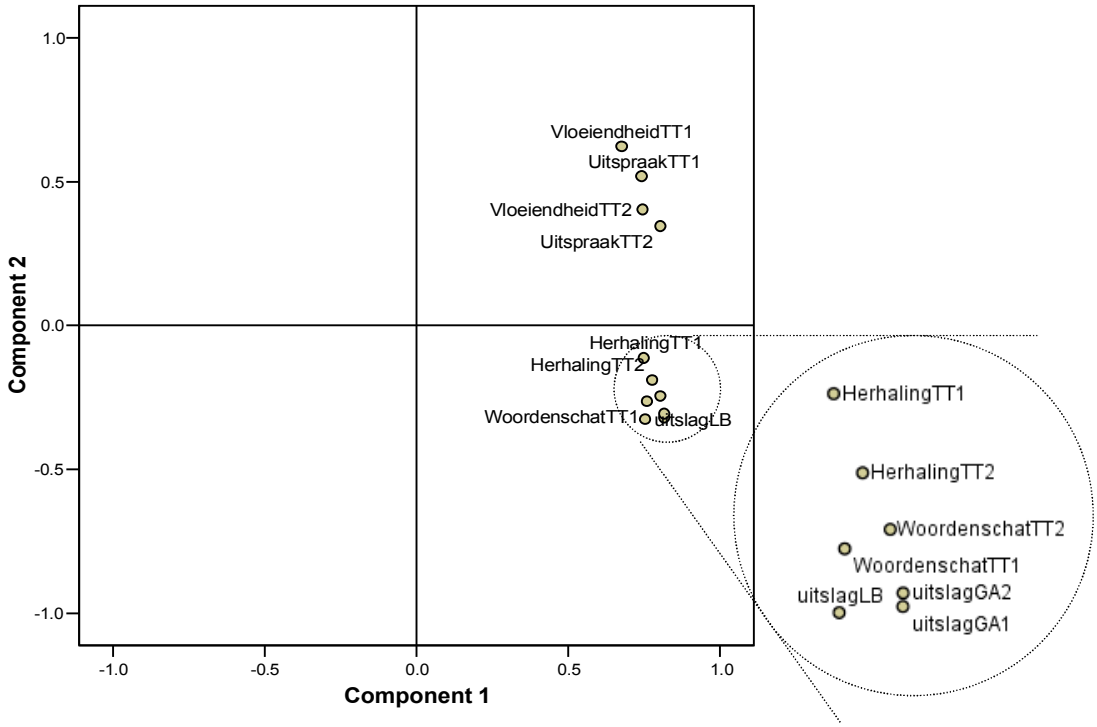
Tabel 6.10 Componentenmatrix

Score	Component 1	Component 2
Zinsbouw (Herhaling)TT1	.748	-.112
WoordenschatTT1	.759	-.263
VloeiendheidTT1	.676	.623
UitspraakTT1	.742	.519
Zinsbouw (Herhaling)TT2	.777	-.190
WoordenschatTT2	.804	-.245
VloeiendheidTT2	.745	.404
UitspraakTT2	.803	.346
GA1=gespreksassessment/interviewer	.816	-.320
GA2=gespreksassessment/beoordelaar	.816	-.307
LB=loopbaangesprek	.754	-.326

Extractie methode: Principale Componenten Analyse
2 componenten geëxtraheerd.

Een strooidiagram van de lading van de betrokken oordelen en deelscores op component 1 en 2 is afgebeeld in Figuur 6.13. De oordelen en deelscores vallen duidelijk in twee groepen uiteen: Vloeiendheid en Uitspraak vormen een aparte groep die positief op beide componenten laadt. De menselijke oordelen vertonen in hun ladingenpatroon de meeste overeenstemming met de inhoudelijke scores (Zinsbouw en Woordenschat).

Een paar componenten hebben een zo precies gelijkend ladingenpatroon dat zij in de figuur niet met het blote oog kunnen worden onderscheiden. Een deel van de figuur is daarom uitvergroot in de ingezette cirkel.



Figuur 6.13 Factorladingen op twee principale componenten

6.3 Conclusies

Diverse schattingen van de betrouwbaarheid indiceren dat met de TGN voldoende betrouwbaar kan worden gemeten: de schattingen overtreffen ruimschoots de in de opdracht gestelde minimale streefwaarde van .80 en laten zich goed vergelijken met de betrouwbaarheidscoëfficiënten van toetsen die nationaal en internationaal als betrouwbaar worden beschouwd. Het automatische spraakherkenning- en scoringssysteem functioneert voldoende precies om met mensen – mits goedgetraind - vergelijkbare oordelen te kunnen genereren. In de verzamelde gegevens is een samenhang gevonden tussen de toetsscores en relevante maten voor de beheersing van gesproken Nederlands in interactie met sprekers van het Nederlands. Er is in de verzamelde data géén aanwijzing gevonden dat de scores op de TGN worden beïnvloed door eigenschappen en kenmerken van kandidaten waarvan verondersteld kan worden dat ze niet samenhangen met taalvaardigheid.
De verzamelde gegevens geven verder aan dat docenten na een intensieve training goed in staat zijn om met gebruikmaking van het CEF betrouwbare oordelen te geven over het taalvaardigheidsniveau van leerders van het Nederlands. Deze bevinding kan beschouwd worden als een ondersteuning voor het door CINOP voorgestelde model voor een inburgeringsexamen in Nederland waarin een decentraal deel op basis van portfolio-assessment en een centraal ontwikkeld examen worden gecombineerd.

Referenties

- American Council on the Teaching of Foreign Languages (1999) *ACTFL proficiency guidelines-speaking: Revised 1999*. Hastings_on_Hudson, NY: ACTFL.
- Baaren, R. van, R. Holland, B. Steenaert en A. van Knippenberg (2003) Mimicry for Money: Behavioral Consequences of Imitation. In: *'Journal of Experimental Social Psychology'*, 39, 393-398.
- Baddely, A. (1986) *Working memory*. Oxford: Clarendon Press.
- Baddely, A. (2000) The episodic buffer: a new component of working memory? In: '*Trends in Cognitive Science'* 4(11), 417-423.
- Breiner-Sanders, K.E., Lowe, P.J. & Miles, J. & Swender, E. (2000) ACTFL Proficiency Guidelines-Speaking Revised 1999. In *'Foreign Language Annals*, Vol. 33, No. 1.
- Bull, M. & Aylett, M. (1998) An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In 'Mannell, R.H. & Robert-Ribes, J. (Eds), *Proceedings of the 5th International Conference on Spoken Language Processing.* Canberra: Australian Speech Science and Technology Association (ASSTA)'.
- Carroll, J.B. (1961) Fundamental considerations in testing for English language proficiency of foreign students. In *'Testing the English proficiency of foreign students'* (p.30-40) Washington, DC: Center for Applied Linguistics.
- Carroll, J.B. (1986) Second Language. In R.F. Dillon, & R.J. Sternberg (Eds.), *Cognition and Instruction.* Orlando, FL: Academic Press.
- CGN (2004). *Corpus Gesproken Nederlands*. Copyright (c) March 2004 Nederlandse Taalunie, Den Haag. Distrubuteur: ELDA, Paris. S0113: Spoken Dutch Corpus.
- Chomsky, N. & Miller, G.A. (1963) Introduction to the formal analysis of natural languages. In: 'Luce, R.D. & Bush, R.R. & Galanter, E. (Eds) *Handbook of Mathematical Psychology.* Vol. 2. Wiley, New York, 269-321.
- Commissie Franssen / Franssen, J. et al. (2004) *Inburgering getoetst. Advies over het niveau van het inburgeringsexamen in het buitenland.* Den Haag.
- Commissie Franssen / Franssen, J. et al. (2004) *Normering inburgeringsexamen. Advies over het niveau van het nieuwe inburgeringsexamen in het Nederland.* Den Haag.
- Council of Europe (2001) *Common European Framework of References for Languages: Learning, teaching and assessment.* Cambridge: Cambridge University Press.
- Cutler, A. (2003) Lexical Access. In 'Nadel, L. (Ed), *Encyclopedia of Cognitive Science' Vol. 2, Epilepsy-Mental imagery* (p 858-864)) London: Nature Publication.
- De Jong, J.H.A.L. & Van Ginkel, C.W. (1992) Dimensions in oral foreign language proficiency. In: Verhoeven, L.T. & De Jong, J.H.A.L. (Eds) *The Construct of Language Proficiency: Applications of Psychological Models to Language Assessment.* Amsterdam: John Benjamins.
- Gibson, E. (1991) *A computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh PA.
- Gibson, E. (1998) Linguistic Complexity: Locality of syntactic dependencies. In: '*Cognition*', 68, 1-76.
- Hest, E. van (1996), *Self-repair in L1 and L2 production.* Dissertatie. Radboud Universiteit Nijmegen.
- Heuvelmans, A. (1994*) OVERTON, A Computer Programme for Estimating Interrater Reliability*. Arnhem: CITO.
- Higgs, T.V. & Clifford, R. (1982) The push towards communication. In: Higgs, T.V. (Ed), *Curriculum, Competence, and the Foreign Language Teacher*. Lincolnwood, IL: National Textbook Company.

- Hoofdlijnenakkoord voor het kabinet CDA, VVD en D66 (2003) *Meedoen, meer werk, minder regels*.
- Jescheniak, J.D., Hahne, A. & Schriefers, H.J. (2003) Information flow in the mental lexicon during speech planning: evidence from event-related brain potentials. In '*Cognitive Brain Research'* 15(3), p 261-276.
- Kerkhoff, A. (2002) *Klaar voor de start.* De Bilt: Bve Raad.
- Lennon, P. (1990) Investigating Fluency in EFL: A Quantitative Approach. In: '*Language Learning'*, 40:3, 387-417.
- Levelt, W.J.M (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
- Levelt, W.J.M (2001) Spoken word production: A theory of lexical access. In '*PNAS*' Vol. 98, No. 23, 13464-13471.
- Linacre, J.M (1988;2005) *A Computer Program for he Analysis of Multi-Faceted Data.* Chicago, IL: Mesa Press.
- Lippman, R.P. (1996) Speech perception by humans and machines. In: '*Proceedings of the European Speech Communication Association Tutorial and Research Workshop on the Auditory Basis of Speech Perception*', Keele University, UK, July 15-19.
- Miller, G.A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. In: '*The Psychological Review'*, 63, pp. 81-97.
- Miller, G.A. & Isard, S. (1963) Some perceptual consequences of linguistic rules. In: '*Verbal Learning and Verbal Behavior'* Vol. 2, 217-228.
- Miller, G. A. & Isard, S. (1964) Free recall of self-embedded English sentences. In : '*Information & Control'* 7, 293-303.
- North, B. (2000) *The development of a Common Framework Scale of Language Proficiency.* New York, NY: Peter Lang.
- Pennington, M.C. (1989) Teaching Pronunciation from the top down. In: '*RELC Journal*' 20/1, 21-38.
- Poelmans, P. (2003) *Developing Second Language Listening Comprehension: Effects of training lower-order skills versus higher-order strategy.* Diss. Amsterdam UvA
- Rosenberg, J. D. (1997) *G.729 error recovery for internet telephony.* Technical report. Columbia University.
- Salame, P., & Baddeley, A. D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. In: *Journal of Verbal Learning and Verbal Behavior*, 21, 150–164.
- Schneider, W. & Shiffrin, R.M. (1977) Controlled and automatic human information processing: I. Detection, search and attention. In '*Psychological Review'* 84, 1-66.
- Van Turenhout, M., Hagoort, P. & Brown, C.M. (1998) Brain Activity During Speaking: From Syntax to Phonology in 40 Milliseconds. In '*Science*' 280, 572-574.
- Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1991) *OPLM: A computer program and manual*. Arnhem: Cito.

Overzicht bijlagen

Bijlage 1

Instructie Taaltoets

(Bron: Ministerie van Buitenlandse Zaken, 2005)

INSTRUCTIEBLAD

TAALTOETS

Instructieblad voor de kandidaat bij het Examen Nederlandse Taal, uit te reiken aan begin van de pauze.

Het examen Nederlandse Taal begint met een geluidstest. U hoort het volgende:
We testen eerst het geluid. Zeg alstublieft de naam van de stad en van het land waar u nu bent.

Als u niet hard genoeg spreekt hoort u: *Uw stem klink erg zacht. U moet harder spreken.*

Of als u te hard spreekt hoort u: *Uw stem klinkt erg hard. Houd de microfoon iets verder van uw mond. Kijk naar het plaatje.*

U hoort dan een stem die zegt: *Zeg nog eens de naam van de stad en van het land waar u nu bent.*

Als uw stem goed duidelijk is begint het examen. U hoort dan: *Welkom bij het examen Nederlandse Taal.*

Deel A Nazeggen:
U hoort steeds een zin. Zeg de zin precies na. Bijvoorbeeld een stem zegt: “dat is een mooi verhaal” en u zegt: "dat is een mooi verhaal”. Nu is het uw beurt. Luister naar de zin en zeg precies na wat u hoort.
U krijgt in deel A twaalf zinnen om na te spreken. Iedere zin is anders.

Aan het eind van deel A hoort u een belgeluid. Dan begint deel B. U hoort:

Deel B Vragen:
U hoort steeds een korte vraag. Geef op elke vraag een kort antwoord. Bijvoorbeeld: een stem zegt: "Is januari een dag of een maand?” En u zegt: "maand" of "een maand". Of u hoort: "Een auto, heeft die twee wielen of vier wielen?En u zegt: “vier” of “vier wielen”. Nu is het uw beurt: luister naar de vraag en geef dan antwoord.
U krijgt in deel B veertien vragen.

Aan het eind van Deel B hoort u een belgeluid. Dan begint deel C. U hoort:

Deel C Nazeggen:
U hoort weer zinnen. Zeg elke zin weer precies na. Bijvoorbeeld: een stem zegt: "dat is een mooi verhaal" en u zegt: "dat is een mooi verhaal". Nu is het uw beurt. Luister naar de zin en zeg precies na wat u hoort.

U krijgt in Deel C weer twaalf zinnen om na te spreken iedere zin is anders.

Aan het eind van deel C hoort u een belgeluid. Dan begint deel D. U hoort:

Deel D Tegenstellingen:
U hoort steeds een woord. U zegt het tegenovergestelde. Bijvoorbeeld: u hoort 'hoog' dan zegt u 'laag', of u hoort 'niet' dan zegt u: 'wel'. Nu is het uw beurt. Luister naar het woord en zeg het tegengestelde woord.
U krijgt in deel D tien woorden.

Aan het eind van deel D hoort u een belgeluid. Dan begint deel E. U hoort:

Deel E Verhalen navertellen:
U hoort korte verhalen. U moet het verhaal navertellen. U krijgt daarvoor 30 seconden. Vertel zoveel mogelijk. Denk bijvoorbeeld aan: wie deden er mee? Wat gebeurde er? Waar was het? Hoe liep het af?
U krijgt in Deel E twee verhalen te horen. Aan het eind van het verhaal hoort u een zachte pieptoon. Dan bent u aan de beurt. U moet het verhaal navertellen.

Na 30 seconden klinkt een harde pieptoon. Dan komt het tweede verhaal. Ook aan het eind van dat tweede verhaal hoort u een zachte pieptoon. Dan bent u weer aan de beurt. U moet het verhaal navertellen.

Na 30 seconden klinkt weer een harde pieptoon.

Daarna hoort u: ***Dank u voor het bellen. U kunt nu ophangen.***

Daarmee is het examen afgelopen. U mag de hoofdtelefoon neerleggen.

Bijlage 2

Werkproces inburgeringsexamen buitenland

(bron: Ministerie van Buitenlandse Zaken, versie 1.0, 22 maart 2005)

Werkproces inburgeringsexamen buitenland

Het onderstaande werkproces moet als handleiding dienen bij het afnemen van examens op de posten in het buitenland in het kader van inburgering.

1.	**Betaling voor het inburgeringsexamen in Nederland.**
2.	**Medewerker maakt afspraak met kandidaat voor het afnemen van het examen en verwerkt dit in een afsprakensysteem.**
3.	**Bevestiging van de afspraak.**
4.	**De kandidaat meldt zich op de afgesproken datum en tijd op de post.**
5.	**Medewerker controleert de kandidaat op hulpmiddelen die niet zijn toegestaan tijdens het examen te gebruiken.**
6.	**Medewerker verifieert geldigheid legitimatiebewijs kandidaat.**
7.	**Medewerker legt biometrische gegevens van de aanvrager vast en verwerkt deze in het systeem.**
8.	**De medewerker instrueert de kandidaat over de gang van zaken tijdens het examen.**
9.	**Nadat medewerker heeft vernomen dat kandidaat instructie heeft begrepen start het examen 'Kennis van de Nederlandse Samenleving'.**
10.	**Na afronding van het examen 'Kennis van de Nederlandse Samenleving' geeft medewerker aan hoelang de pauze is.**
11.	**Na pauze start het examen 'Kennis der Nederlandse Taal'.**
12.	**Afronding examen; mededelen wanneer en hoe resultaat van de toets bekend wordt gemaakt.**
13.	**Medewerker registreert gebruikte TINcodes.**
14.	**Medewerker ontvangt uitslag en registreert deze in het systeem.**
15.	**Medewerker brengt de kandidaat op de hoogte van het resultaat van het examen op de afgesproken wijze.**

Stap 1. Betaling voor het inburgeringsexamen in Nederland

De kosten voor het inburgeringsexamen worden giraal betaald op één centrale bankrekening van BZ in Nederland. Hierbij geldt de onderstaande betalingsprocedure:

- Iedere kandidaat die het inburgeringsexamen wil afleggen, heeft per definitie een referent in Nederland. De verantwoordelijkheid voor betaling wordt neergelegd bij de referent.
- Algemene betalingsinformatie wordt beschikbaar gesteld op internet en via brochures van de IND.
- De referent in Nederland geeft door het invullen van een formulier op internet aan BZ door dat een kandidaat het inburgeringsexamen wil afleggen.
- BZ verwerkt de gegevens van de referent en de partner in een computersysteem. Dit computersysteem genereert een uniek betalingskenmerk dat de referent dient te vermelden bij de overmaking op de bankrekening van het Ministerie.
- Na ontvangst van de betaling stelt BZ de post daarvan op de hoogte.
- De kandidaat kan vervolgens een afspraak maken met de betreffende ambassade voor het examen. Bij het maken van de afspraak dient de kandidaat het betalingskenmerk te vermelden.

Stap 2. Afspraak inburgeringsexamen

Het is de taak van de post om bekend te maken waarom en op welke wijze mensen een afspraak kunnen/dienen te maken voor een inburgeringsexamen. Indien de betaling niet heeft plaatsgevonden wordt er geen afspraak gemaakt. Mocht iemand aangeven medische ontheffing te willen aanvragen, dan geldt de aparte procedure voor medische ontheffingen.

Voor het afleggen van een inburgeringsexamen dient een afspraak gemaakt te worden door de kandidaat met de post. Dit om het aanbod van kandidaten te kunnen reguleren en/of om ervoor te kunnen zorgen dat de systemen die bij het examen worden gebruikt beschikbaar zijn en goed functioneren. De post hanteert hiervoor een afsprakensysteem. Het is aan te bevelen in het begin afspraken niet te dicht op elkaar te plannen om vertragingen te voorkomen die zouden kunnen ontstaan doordat deze nieuwe werkprocessen nog niet volledig worden beheerst.

Bij het plannen van de afspraak dienen posten die de biometrie-opstelling in de examenruimte hebben staan, er rekening mee te houden dat wanneer een MVV-klant voor gezinshereniging zijn MVV komt ophalen hij/zij wel gecontroleerd moet worden op vingerafdruk Hiervoor is toegang van de vingerscan in de examenruimte nodig Voorkomen moet worden dat examens hierdoor worden verstoord.

Stap 3. Bevestiging van de afspraak

De post dient bij het maken/bevestigen van de afspraak kandidaat te melden dat kandidaat zijn paspoort/identiteitsbewijs en zijn betalingskenmerk dient mee te brengen en dat naast het examen ook biometrische kenmerken zullen worden afgenomen. De post meldt hierbij hoeveel tijd dit in zijn geheel in beslag zal nemen.

De kandidaat dient ook gemeld te worden dat het niet is toegestaan apparatuur mee naar de examenruimte te nemen; zoals mobiele telefoon of opnameapparatuur. Kandidaat dient het examen alleen af te leggen; het meenemen van kinderen of andere familieleden of vrienden in de examenruimte is NIET toegestaan. In geval de kandidaat een taal spreekt die de postmedewerker niet machtig is, kan dit een probleem vormen bij het geven van de instructie. In deze gevallen dient betrokkene

er op gewezen te worden dat hijzelf voor een tolk (evt. familie) kan zorgdragen. De tolk kan aanwezig zijn bij de instructie, daarna zal betrokkene het examen alleen afleggen.

Indien mogelijk zal de afspraak schriftelijk bevestigd worden (dit staat de post vrij om zelf te bepalen). Kandidaat kan dan in genoemde brief gevraagd worden de correspondentie mee naar de post te nemen (dit kan handig zijn voor beveiliger die mensen controleert bij binnenkomst) wanneer hij/zij komt voor het examen.

Stap 4. Kandidaat arriveert op de post
Kandidaat dient zich op afgesproken datum en tijd te melden bij de post en volgt het vigerende beveiligingregime voor toelating kanselarij.

Voor de kandidaat is het inburgeringsexamen van groot belang. Indien hij hiervoor niet slaagt bestaat de kans dat hij iedere mogelijkheid aangrijpt om een klachtenprocedure te starten. Voor de behandeling van die klachten is het van belang om alle bijzonderheden die zich voordoen tijdens het examen te noteren. Voorbeelden van bijzonderheden zijn problemen die zich voordoen bij het afnemen van de vingerafdrukken of plotseling omgevingsgeluid tijdens het afleggen van het examen. Na afloop van het examen dienen de bijzonderheden geregistreerd te worden in het Inburgerings Examen Biometrie Systeem, in het veld opmerkingen, zoals toegelicht wordt bij stap 6.

Wanneer kandidaat onverhoopt niet kan arriveren op afgesproken datum en tijd dient hij voortijdig hiervan de post in kennis te stellen en een nieuwe afspraak te maken. Mocht kandidaat later zijn dan afgesproken tijdstip is het aan de post zelf om kandidaat alsnog zijn examen te laten doen of weg te sturen en te vragen een nieuwe afspraak te maken.

Stap 5. Controle van de kandidaat
Aan kandidaten wordt, bij het maken van een afspraak voor het inburgeringsexamen, reeds duidelijk meegedeeld dat er tijdens het examen niets meegenomen mag worden in de examenruimte .

Checken op verboden voorwerpen (bij afname examen buiten de beveiligde zone)

In beginsel geldt voor het controleren van de kandidaat het heersende veiligheidsregime op de post. Daarnaast zal aan de kandidaat gevraagd moeten worden om persoonlijke spullen tijdelijk in te leveren. Dit om afleiding en fraude tijdens het examen te voorkomen. Expliciet moet worden gevraagd of hij in het bezit is van elektronische apparatuur, in het bijzonder een mobiele telefoon. De mobiele telefoon moet uitgeschakeld worden voordat de telefoon wordt ingeleverd (dit met het oog op opnamefunctie dan wel afluisteren). Laat de kandidaat de voorwerpen aan zijn kant van de balie neerleggen (bijvoorbeeld in een mandje). Let op dat de postmedewerker nooit in aanraking met de spullen kan komen, om problemen en klachten achteraf te voorkomen. Zorg er voor dat zowel de kandidaat zelf als de postmedewerker zicht kunnen blijven houden op de voorwerpen.

Als de postmedewerker twijfelt of alle apparatuur of andere ongewenste voorwerpen daadwerkelijk zijn ingeleverd kan een extra controle verricht worden met behulp van een handscanner. Een dergelijke controle dient overigens te allen tijde uitgevoerd te worden door een beveiligingsmedewerker.

Indien de kandidaat medewerking weigert of indien tijdens het examen blijkt dat er toch verboden voorwerpen op de kandidaat aanwezig zijn wordt het examen direct afgebroken. De kandidaat ontvangt geen certificaat en heeft geen recht op restitutie van het examengeld.

Dit geldt ook indien de kandidaat zelf het examen afbreekt omdat er bijvoorbeeld toch zijn mobiele telefoon overgaat.

Checken op verboden voorwerpen (bij afname examen binnen de beveiligde zone)

In geval van het afnemen van examens binnen de beveiligde zone gelden naast het bovenstaande de volgende additionele regels:

1. De kandidaat dient gecontroleerd te worden alvorens hij de beveiligde zone betreedt (door middel van detectiepoortje, handscanner en/of veiligheidsfunctionaris).
2. De kandidaat dient te allen tijde binnen de beveiligde zone begeleid te worden door een veiligheidsfunctionaris. De veiligheidsfunctionaris zal gedurende het examen in de nabijheid van de kandidaat en toezichthouder blijven.
3. De ruimte waarin het examen wordt afgenomen dient zo dicht mogelijk bij de ingang van de kanselarij te bevinden, zodat de kandidaat zo min mogelijk inzicht kan krijgen in de fysieke en organisatorische aspecten van de kanselarij. De ruimte dient zich te bevinden in zone twee.

Stap 6. Medewerker verifieert legitimatiebewijs kandidaat

Medewerker verifieert het legitimatiebewijs, dat kandidaat heeft meegenomen, op echtheid. Kopie wordt hiervan gemaakt voor het dossier.

Stap 7. Registreren gegevens kandidaat

Het registreren van gegevens van de kandidaat geschiedt in het Inburgerings Examen Biometrie Systeem (IEBS). Voor meer gedetailleerde informatie over de wijze van registratie wordt u verwezen naar de gebruikershandleiding van het IEBS. De functies van het IEBS worden gestart vanuit het hoofdscherm. Dit ziet er als volgt uit:

Eerste maal examen

Hieronder wordt de procedure beschreven voor de registratie van een kandidaat die voor de eerste maal een examen komt afleggen. Deze procedure bestaat uit de volgende onderdelen:

1. Afnemen vingerafdrukken

- Laat de kandidaat de handen omhoog houden en de handpalmen open tonen. Controleer op de volgende punten:
  - Zijn de handen schoon? Zo niet laat handen schoonmaken.
  - Bevatten de handen verf en lijm resten en dergelijke. Indien aanwezig laat de handen schoonmaken.
  - Bevatten de handen silicone nep vingerafdrukken. Indien aanwezig dienen de gangbare procedures bij constatering van fraude toegepast te worden.
  - Zijn er vingers niet afneembaar in verband met pleisters of verband. Zo ja (en bij meerdere vingers, met name wijsvingers) probeer na te gaan of verwondingen echt zijn.
  - Ontbreken er vingers of vingertoppen.
- Geef de kandidaat een doekje aan en laat de vingers hiermee afnemen. Laat de kandidaat hiermee ook de glasplaat van de vingerscanner afnemen. Dit is niet voor iedere afname direct nodig, maar kan de kandidaat een vertrouwder en hygiënischer gevoel geven.
- Leg uit hoe het afnemen gaat verlopen en geef de plaat met pictogrammen over het plaatsen van de vingers aan.
- Neem via de functie 'Select by finger' eerst één vingerafdruk af (zie de aandachtspunten voor een afname).
  - Wordt de kandidaat met deze vingerafdruk gevonden controleer dan de gegevens en de foto. Indien deze niet met de persoon overeenkomen noteer dan de gegevens van de persoon die bij de vinger van de kandidaat werd gevonden als bijzonderheid

(wordt later geregistreerd bij opmerkingen). Voer nogmaals een controle uit met een andere vinger.
    - Wordt de persoon met meerdere vingers teruggevonden, of bestaat er anderszins een vermoeden van fraude, pas dan de gangbare procedures bij constatering van fraude toe.
    - Wordt de kandidaat niet gevonden ga dan verder met het afnemen van alle tien de vingers.
- Neem via de functie 'Enrollment' één voor één de vingerafdrukken af (zie de aandachtspunten voor een afname).
- Indien na 3 pogingen de vingerafdruk niet afgenomen kan worden of de vinger is niet afneembaar (door verband, pleister, of ontbreken), registreer dit bij deze vinger met de knop 'Ignore'. Kies in de selectielijst voor:
    - Temporary trauma: indien er een wondje of pleister op de vinger zit.
    - Permanent trauma: indien de vinger blijvend beschadigd is (of ontbreekt).
    - Unable to capture: indien er geen directe oorzaak aan de vinger kan worden gevonden, maar een afdruk niet kan worden afgenomen.
- Druk na de laatste vinger op de knop 'OK'. De vingerafdrukken worden nu vergeleken met de reeds geregistreerd vingerafdrukken.
    - Geeft het systeem aan dat de kandidaat al bekend is in het systeem controleer dan de gegevens en de gemaakte foto. Betreft het een andere persoon dan is er sprake van een vergissing van het systeem of is er een poging tot fraude. Bestaat er een vermoeden van fraude, pas dan de gangbare procedures bij constatering van fraude toe.
    - Indien de kandidaat nog niet door het systeem wordt herkend wordt vervolgd met het registreren van de persoonsgegevens.

Fall-back procedure: Er kunnen geen of niet voldoende vingerafdrukken afgenomen worden
Voor een latere vergelijking is het van belang om meerdere vingerafdrukken af te kunnen nemen van de kandidaat. Bij een latere vergelijking van vingerafdrukken is de pink het minst betrouwbaar. Daarom is het van belang dat er afdrukken van minimaal twee vingers (waaronder niet de pinken) afgenomen kunnen worden. Lukt dit niet dan treedt de fall-back procedure in werking. Deze bestaat uit de volgende stappen:
- Betrek een tweede postmedewerker bij de procedure en probeer nogmaals om vingerafdrukken af te nemen.
- Besteed extra aandacht aan de controle van het meegebrachte identiteitsbewijs. Vraag eventueel om extra bewijsmateriaal om de identiteit vast te stellen.
- Indien registratie door vingerafdrukken ook door een andere medewerker niet mogelijk blijkt, registreer dan de kandidaat met behulp van de functie 'Add person'. Deze functie is beveiligd met een wachtwoord en kan alleen door het hoofd visumzaken opgeroepen worden.

Aandachtspunten bij het afnemen van een vingerafdruk
- Indien de handen niet gedurende het gehele proces zichtbaar zijn of indien er twijfel bestaat laat de kandidaat dan nogmaals de handen tonen.
- Let op dat de juiste vinger wordt geplaatst.
- Let er op dat de vinger op de juiste wijze wordt geplaatst:
    - Vlak op de scanner, niet alleen de toppen of gedraaid.
    - In het midden van de scanner.
    - Tegen de bovenrand van de scanner.
- In het afnamescherm wordt een grote cirkel en een kruis getoond.
    - Indien het kruis in het midden van de cirkel staat verandert het kruis van kleur (van rood naar groen). Gebeurt dit niet vraag dan de kandidaat om de vinger in de juiste richting te verplaatsen.

  - Indien de afbeelding op het scherm er visueel slecht uit ziet, bijvoorbeeld onderbroken lijn of ontbrekende stukken in de afbeelding, instrueer de kandidaat dan de vinger harder op scanner te drukken.
  - Indien het beeld 'volloopt', de vingerafdruk lijkt dan een grote zwarte vlek, instrueer dan de kandidaat minder hard te drukken.
- Indien het niet lukt binnen de beschikbare tijd een vingerafdruk af te nemen laat dan nogmaals de vinger schoonmaken met het doekje. Afhankelijk van de lokale omstandigheden (vochtige omgeving) is het aan te bevelen om vlak voor iedere afname (dus per vinger) de vinger nog even droog te laten maken met een tissue / zakdoekje.
- Door onwennigheid van de kandidaat duurt het afnemen van een afdruk van de eerste vinger meestal langer. Om de kandidaat niet onnodig zenuwachtig te maken is het aan te raden na twee of drie pogingen te stoppen met deze vinger en door te gaan met de volgende. Nadat één vinger succesvol is afgenomen volgen de andere vingers meestal gemakkelijker. In de meeste gevallen kan dan als laatste ook nogmaals de vinger(s) geprobeerd worden die in eerste instantie overgeslagen zijn. Deze lukken dan meestal ook.

2. Registreren persoonsgegevens
- Registreer de persoonsgegevens zo volledig mogelijk.
- Gebruik als case numer het betalingskenmerk, zoals gecommuniceerd bij het maken van de afspraak.
- Voorlopig is van een kandidaat voor een inburgeringsexamen nog geen vreemdelingennummer bekend. Dit gegeven hoeft tot nader order niet ingevuld te worden.

3. Nemen foto
- Let bij het maken van een foto op de volgende punten:
  - Foto is frontaal van het gezicht.
  - Zorg dat de afstand van de kandidaat tot de camera ongeveer 1 meter is. Verplaats de camera omhoog of omlaag zodat het gezicht beeldvullend is.
  - Plaats de camera goed tegen de ruit aan, om schaduw en verkeerde lichtinval te voorkomen.
  - Let op dat bij het schuiven van de camera tegen de ruit de focus van de camera kan veranderen, dit levert wazige foto's op. Stel de camera scherp door aan de ring aan de voorzijde te draaien.
  - Zorg voor een juiste belichting. Bijvoorbeeld geen open raam of deur waar veel licht doorkomt als achtergrond.
  - Ogen moeten open zijn op de foto.
  - Gezichtsuitdrukking moet neutraal zijn op de foto.
  - Het gehele gezicht moet herkenbaar zijn, bedekking die (delen van) het gezicht verhullen zijn niet toegestaan.

4. Registreren aanvullende informatie
- Vraag de volgende aanvullende informatie aan de kandidaat:
  - Is de kandidaat analfabeet?
  - Wat is het opleidingsniveau?

Deze informatie kan alleen mondeling van de kandidaat gekregen worden, de antwoorden kunnen niet worden geverifieerd.
Tevens dient hier aangegeven te worden indien een kandidaat een ontheffing voor het examen heeft.

5. Registreren legitimatiebewijs

Registreer de gegevens van het legitimatiebewijs waarvan reeds een kopie werd gemaakt.

- Vul de velden op het scherm in met de gegevens van het legitimatiebewijs.
- Registreer de gegevens van het identiteitsbewijs en leg deze met de knop 'Add' vast.

6. Registreren opmerkingen

- Registreer eventuele bijzonderheden.
- Registreer hier ook bijzonderheden die tijdens de biometrieafname naar voren zijn gekomen.

7. Opslaan gegevens

Nadat alle gegevens geregistreerd zijn worden deze opgeslagen met de knop 'Save'. Nadat de gegevens opgeslagen zijn verschijnt het scherm met persoonlijke gegevens weer. Laat dit scherm openstaan om na afloop van het examen eventuele bijzonderheden te registreren.

Herexamen

Indien een kandidaat komt voor een herexamen hoeven niet nogmaals alle vingerafdrukken geregistreerd te worden. De volgende stappen worden dan doorlopen:

- Zoek de kandidaat via functie 'Select by name'.
  - Selecteer bij 'Search for + sort' voor Full name.
  - Vul (een deel van) de achternaam in en druk op de knop 'Search'.
- Selecteer de juiste persoon in de lijst en druk op 'OK'.
- Controleer de persoonsgegevens en de foto aan de hand van het meegebrachte paspoort / identiteitsbewijs.
- Neem een vingerafdruk af via de functie 'verify' (zie de aandachtspunten voor een afname).
  - Wordt de vinger niet herkend is er mogelijk sprake van fraude. Probeer de 'verify' nogmaals met een andere vinger. Bestaat er een vermoeden van fraude, pas dan de gangbare procedures bij constatering van fraude toe.

Kandidaat komt voor MVV aanvraag

Indien een kandidaat is geslaagd voor het examen en terugkomt voor de MVV aanvraag dient nogmaals gecontroleerd te worden of deze persoon dezelfde is als degene die het examen heeft afgelegd. Gebruik de procedure zoals beschreven bij **herexamen** om de gegevens van de kandidaat met behulp van een vingerafdruk te verifiëren. Controleer of de kandidaat geslaagd is voor het examen.
Is de kandidaat niet geslaagd moet aanvraag wel worden ingenomen en behandeld. Waarbij gewezen moet worden op de twee weken verzuimperiode.

Stap 8. Instructie aan de kandidaat

Het examen bestaat uit twee delen, die altijd in onderstaande volgorde worden afgenomen:

1. Examen 'Kennis van de Nederlandse Samenleving'.
2. Examen 'Kennis van de Nederlandse Taal'.

Voorafgaand aan de instructie wordt het instructieblad voor het examen 'Kennis van de Nederlandse Samenleving' aan de kandidaat uitgereikt (in de gewenste taal) en krijgt de kandidaat 5 minuten leestijd. Indien betrokkene analfabeet is, zal het instructieblad puntsgewijs door medewerker van de post moeten worden voorgelezen.
De examenleider zorgt dat de tien verschillende fotoboeken beschikbaar zijn.
Nadat de kandidaat de instructies heeft gelezen start de instructie voor het examen 'Kennis van de Nederlandse Samenleving'.

Examenleider:
De examenleider vraagt vooraf of de kandidaat de video en het studiemateriaal heeft gebruikt ter voorbereiding op het examen, zoniet, vermeld dit dan bij ‘opmerkingen’. De examenleider legt uit wat er in de toets gaat gebeuren. Geeft uitleg over de gebruikte apparatuur:
'Straks bel ik het toetssysteem met deze telefoon (wijst op de eigen telefoon van de examenleider) en maak de verbinding voor u klaar.
Daarna gaat u spreken door deze hoofdtelefoon.
Weet u hoe de hoofdtelefoon te gebruiken?
Zorg dat de microfoon zich straks ongeveer ter hoogte van de mond bevindt.'
Test door de hoofdtelefoon op te zetten. Hierna de hoofdtelefoon afzetten voor de duur van de instructie.

Examenleider:
'Als de toets begint schakel ik uw hoofdtelefoon in. Het examen verloopt zoals u in de instructie hebt gelezen. Eerst is er een geluidstest, om er zeker van te zijn dat u goed hoorbaar bent. Daarna volgen er twee voorbeeldvragen. Deze tellen niet mee voor de score.

De vragen worden gesteld aan de hand van het fotoboek dat u van mij krijgt. In totaal zijn er 30 vragen. U krijgt een vraag over elke foto uit het boek. U moet zelf de pagina omslaan om de volgende foto te zien. Het examen duurt ongeveer twintig minuten.

Op het instructieblad heeft u gelezen hoe het examen verloopt? Heeft u nog vragen?

Instructieblad voor de kandidaat bij het Examen Kennis van de Nederlandse Samenleving, uit te reiken aan de kandidaat, 5 minuten voor het examen begint.

Het examen Kennis van de Nederlandse Samenleving begint met een geluidstest. U hoort het volgende:
We testen eerst het geluid. Zeg alstublieft de naam van de stad en van het land waar u nu bent.

Als u niet hard genoeg spreekt hoort u:
Uw stem klink erg zacht. U moet harder spreken.

Of als u te hard spreekt hoort u:
Uw stem klinkt erg hard. Houdt de microfoon iets verder van uw mond. Kijk naar het plaatje.

U hoort dan een stem die zegt:
Zeg nog eens de naam van de stad en van het land waar u nu bent.

Als uw stem goed duidelijk is begint het examen. U hoort dan:
Welkom bij het examen Kennis van de Nederlandse Samenleving.
Volg nu de instructies bij dit examen. U krijgt dertig vragen. Bij elke vraag hoort een foto.
Kijk goed in het fotoboek.
Geef een kort antwoord. Daarna hoort u het nummer van de volgende foto.
Sla dan de pagina om.
Kijk naar de volgende foto.
Eerst hoort u twee voorbeeldvragen.
Sla de pagina om. Voorbeeldvraag.

'U ziet de Nederlandse vlag. Wat zijn de kleuren van de Nederlandse vlag?' En u zegt: 'rood, wit, blauw.'

Sla de pagina om. Voorbeeldvraag.
'U ziet een foto. Is dit Willem van Oranje of prinses Maxima?' En u zegt: 'Willem van Oranje.'

Dat waren de voorbeelden. Het eigenlijke examen begint daarna; u hoort:
En nu is het uw beurt. Geef op elke vraag een kort antwoord.
Let op: we beginnen nu met het examen.

Nu komen de vragen waar u zelf antwoord op moet geven. Spreek luid en duidelijk. Als u het antwoord niet weet zeg dan niets, of zeg: 'ik weet het niet'. Let op het is belangrijk dat u duidelijk praat. U moet ook het hele examen afmaken. Als u het examen niet afmaakt is het examen niet geldig. Het examen is afgelopen als u hoort:
Dank u voor het bellen. U kunt nu ophangen.

Heeft u alles goed begrepen? Stel uw vragen over het examen aan de postmedewerker voor het examen begint! Tijdens het examen kunnen geen vragen meer worden gesteld.

Examenleider:
Controleert nogmaals of de kandidaat de instructies goed heeft begrepen. Optioneel kan de post er voor kiezen de kandidaat te laten tekenen hiervoor, om klachten achteraf te voorkomen.

Stap 9. Afnemen examen 'Kennis van de Nederlandse Samenleving'
'Dan gaan we nu beginnen. Wilt u de hoofdtelefoon goed opzetten? Ik ga nu de toets opstarten. Wacht op mijn teken.'

Nu volgen de handelingen om de verbinding met het toetssysteem van Ordinate leggen. Dit is een interactie tussen de examenleider en het toetssysteem.

Examenleider	Belt het examennummer (voorgeprogrammeerd in het telefoontoestel, iedere examenpositie heeft een uniek inbelnummer in het toetssysteem).
Toetssysteem	Controleert of de binnenkomende oproep afkomstig is van een geldig nummer van een van de geautoriseerde BZ testcenters. Zo niet: meldt dat vanaf een niet geautoriseerde lijn wordt gebeld en breekt af.
Toetssysteem	Laat de volgende boodschap horen: '*Thank you. For a language test, press 1; for a culture test, press 2.'* (Dank u. Voor een taalexamen toets 1, voor een cultuurexamen toets 2.)
Examenleider	Toetst 2 op de telefoon.
Toetssysteem	Selecteert willekeurig een versie [x] uit de 10 versies van het Cultuurexamens, haalt het eerstvolgende TestIdentificatieNummer (TIN) bij die versie op en laat de volgende boodschap horen: '*Please use test booklet for culture test form number [x]. To repeat this information, press1; otherwise, press 2*'(Gebruik alstublieft het fotoboek behorend bij versie [x] van het cultuurexamen. Om deze boodschap nogmaals te horen, toets 1; wilt u doorgaan toets 2.)
Examenleider	Kiest het juiste fotoboek en overhandigt dit aan de kandidaat.

Examenleider	Toetst 2 op de telefoon (tenzij nog onzekerheid bestaat over het nummer van de versie van het cultuurexamen en door 1 te toetsen het versienummer nogmaals wordt gemeld).
Toetssysteem	'*Please write down the following culture test identification number in the candidates file: <xxxx xxxx>. Repeat: <xxxx xxxx>.If you would like to hear the test identification number again, press 1; otherwise press 2'.* (Noteer alstublieft het volgende toetsidentificatienummer in het dossier van de kandidaat: <xxxx xxxx>. Ik herhaal: <xxxx xxxx>. Als u het toetsidentificatienummer nogmaals wilt horen, toets dan 1, zo niet toets 2.)
Examenleider	Toetst 2 op de telefoon (tenzij nog onzekerheid bestaat over het toetsidentificatienummer van de versie van het cultuurexamen en door 1 te toetsen het toetsidentificatienummer nogmaals wordt gemeld).
Toetssysteem	'*Thank you. We are now ready to begin the test. Please press 3 to begin the test and switch the telephone to the candidate' (*Dank u. Wij zijn nu klaar om met het examen te beginnen. Toets alstublieft 3 om het examen te starten en schakel de telefoon over naar de kandidaat.)
Examenleider	Schakelt de telefoon over naar de kandidaat en zegt: *'Het examen gaat nu beginnen. Succes!'* (teken medewerker aan kandidaat)
Toetssysteem	Wacht tien seconden. Leidt de kandidaat door het examen, stelt vragen.
Kandidaat	Geeft antwoorden.
Examenleider	Ziet er nauwlettend op toe dat de kandidaat telkens slechts een bladzijde tegelijk omslaat en dat de nummering op de pagina's van het fotoboek worden gevolgd.
Toetssysteem	*'Dank u voor het bellen. U kunt nu ophangen.'*
Kandidaat	Legt hoofdtelefoon neer.
Examenleider	Luistert via de telefoon of het examen inderdaad afgerond is, en verbreekt daarna de verbinding. Indien de verbinding verbroken wordt voordat het examen is afgelopen kan er geen score worden bepaald, en dus is de toets ongeldig. Neemt het gebruikte fotoboek weer in.

<u>Stap 10. Pauze en instructie examen 'Kennis van de Nederlandse Taal'</u>

Examenleider	*'U bent nu klaar met dit deel van het examen. Wij gaan straks door met het examen Kennis van de Nederlandse Taal. U kunt even pauze nemen. Ik geef u ook alvast het Instructieblad taalexamen, dat u door moet nemen. Het examen zal net zo verlopen zoals de oefentoets die u reeds heeft gedaan'* (pauze maximaal 10 minuten, bekertje water/toilet, kandidaat niet naar buiten?)
Examenleider:	De examenleider vraagt vooraf of de kandidaat een oefentoets heeft gemaakt ter voorbereiding (zoniet, vermeld dit dan bij 'opmerkingen'). Legt uit dat het taalexamen op dezelfde wijze wordt afgenomen als de

oefentoetsen, en ook hetzelfde als het examen Kennis van de Nederlandse samenleving.

Examenleider: *Op het instructieblad heeft u gelezen hoe het taalexamen verloopt. Zijn daar nog vragen over?*

Instructieblad voor de kandidaat bij het Examen Nederlandse Taal, uit te reiken aan begin van de pauze.

Het examen Nederlandse Taal begint met een geluidstest. U hoort het volgende:

We testen eerst het geluid. Zeg alstublieft de naam van de stad en van het land waar u nu bent.

Als u niet hard genoeg spreekt hoort u:
Uw stem klink erg zacht. U moet harder spreken.

Of als u te hard spreekt hoort u:
Uw stem klinkt erg hard. Houdt de microfoon iets verder van uw mond. Kijk naar het plaatje.

U hoort dan een stem die zegt:
Zeg nog eens de naam van de stad en van het land waar u nu bent.

Als uw stem goed duidelijk is begint het examen. U hoort dan:
Welkom bij het examen Nederlandse Taal.
Deel A
Nazeggen. U hoort steeds een zin. Zeg de zin precies na. Bijvoorbeeld een stem zegt: 'dat is een mooi verhaal' en u zegt: 'dat is een mooi verhaal'. Nu is het uw beurt. Luister naar de zin en zeg precies na wat u hoort.

U krijgt in deel A twaalf zinnen om na te spreken iedere zin is anders.

Aan het eind van deel A hoort u een belgeluid. Dan begint deel B. U hoort:

Deel B
Vragen. U hoort steeds een korte vraag. Geef op elke vraag een kort antwoord. Bijvoorbeeld: een stem zegt: "Is januari een dag of een maand?" En u zegt:"maand" of " een maand" Of u hoort:
"Een auto, heeft die twee wielen of vier wielen?En u zegt: 'vier' of 'vier wielen'. Nu is het uw beurt: luister naar de vraag en geef dan antwoord.
U krijgt in deel B veertien vragen.

Aan het eind van Deel B hoort u een belgeluid. Dan begint deel C. U hoort:

Deel C
Nazeggen. U hoort weer zinnen. Zeg elke zin weer precies na. Bijvoorbeeld: een stem zegt: 'dat is een mooi verhaal' en u zegt: 'dat is een mooi verhaal'. Nu is het uw beurt. Luister naar de zin en zeg precies na wat u hoort.
U krijgt in Deel C weer twaalf zinnen om na te spreken iedere zin is anders.

Aan het eind van deel C hoort u een belgeluid. Dan begint deel D. U hoort:

Deel D
Tegenstellingen. U hoort steeds een woord. U zegt het tegenovergestelde. Bijvoorbeeld: u hoort 'hoog' dan zegt u 'laag', of u hoort 'niet' dan zegt u: 'wel'. Nu is het uw beurt. Luister naar het woord en zeg het tegengestelde woord.
U krijgt in deel D tien woorden.

Aan het eind van deel D hoort u een belgeluid. Dan begint deel E. U hoort:
Deel E
Verhalen navertellen. U hoort korte verhalen. U moet het verhaal navertellen. U krijgt daarvoor 30 seconden. Vertel zoveel mogelijk. Denk bijvoorbeeld aan: wie deden er mee? Wat gebeurde er? Waar was het? Hoe liep het af?

U krijgt in Deel E twee verhalen te horen. Aan het eind van het verhaal hoort u een zachte pieptoon. Dan bent u aan de beurt. U moet het verhaal navertellen. Na 30 seconden klinkt een harde pieptoon. Dan komt het tweede verhaal. Ook aan het eind van dat tweede verhaal hoort u een zachte pieptoon. Dan bent u weer aan de beurt. U moet het verhaal navertellen. Na 30 seconden klinkt weer een harde pieptoon.
Daarna hoort u:
Dank u voor het bellen. U kunt nu ophangen.
Daarmee is het examen afgelopen. U mag de hoofdtelefoon neerleggen.

Examenleider *Hebt u alles goed begrepen? Is de instructie duidelijk?* Optioneel kan de post er voor kiezen de kandidaat te laten tekenen hiervoor, om klachten achteraf te voorkomen.

Stap 11. Afnemen examen 'Kennis van de Nederlandse Taal'

Examenleider Dan gaan we nu beginnen. U mag de hoofdtelefoon weer opzetten. Ik ga het examen starten. Wacht op mijn teken.'

Nu volgen de handelingen om de verbinding met het toetssysteem van Ordinate leggen. Dit is een interactie tussen de examenleider en het toetssysteem.

Examenleider Belt het examennummer (voorgeprogrammeerd in het telefoontoestel, iedere examenpositie heeft een uniek inbelnummer in het toetssysteem).

Toetssysteem Controleert of de binnenkomende oproep afkomstig is van een geldig nummer van een van de geautoriseerde BZ testcenters,

Toetssysteem Na inbellen laat de volgende boodschap horen: *'Thank you. For a language test, press 1; for a culture test, press 2.'* (Dank u. Voor een taalexamen,toets 1; voor een cultuurexamen, toets 2.)

Examenleider *Toetst 1 op de telefoon.*

Toetssysteem Haalt het eerstvolgende TestIdentificatieNummer (TIN) voor het Taalexamen op en laat de volgende boodschap horen: *'Please write down the following language test identification number in the candidates file: <xxxx xxxx>. Repeat: <xxxx xxxx>.If you would like to hear the test identification number again, press 1; otherwise press 2'.* (Noteer alstublieft het volgende toetsidentificatienummer in het dossier van de kandidaat: <xxxx xxxx>. Ik herhaal: <xxxx xxxx>. Als u het toetsidentificatienummer nogmaals wilt horen, toets dan 1; zoniet toets 2.)

Examenleider *Toetst 2 op de telefoon* (tenzij nog onzekerheid bestaat over het toetsidentificatienummer van het taalexamen en door 1 te toetsen het toetsidentificatienummer nogmaals wordt gemeld).

Toetssysteem	*'Thank you. We are now ready to begin the test. Please press 3 to begin the test and switch the telephone to the candidate'* (Dank u. Wij zijn nu klaar om met het examen te beginnen. Toets alstublieft 3 om het examen te starten en schakel de telefoonlijn over naar de kandidaat.)
Examenleider	Het examen begint. De examenleider houdt toezicht.Schakelt de telefoon over naar de kandidaat en zegt: *'Het examen gaat nu beginnen. Succes!'*
Toetssysteem	Wacht tien seconden. Leidt de kandidaat door het examen, stelt vragen.
Kandidaat	Geeft antwoorden.
Toetssysteem	*'Dank u voor het bellen. U kunt nu ophangen.'*
Kandidaat	Legt hoofdtelefoon neer.
Examenleider	Luistert via de telefoon of het examen inderdaad afgerond is, en verbreekt daarna de verbinding. Indien de verbinding verbroken wordt voordat het examen is afgelopen kan er geen score voor het examen worden bepaald.
Examenleider	*'U bent nu klaar met het examen.'*

Stap 12. Afronding examen

Examenleider *'U bent nu klaar met het examen.'*

De examenleider deelt de kandidaat mee wanneer en op welke wijze de uitslag bekend gemaakt wordt. De post kan zelf bepalen wanneer deze uitslag aan de kandidaat wordt bekendgemaakt. Denk aan volgende mogelijkheden:

a. Laat de kandidaat wachten op de uitslag.
b. Laat de kandidaat later terugkomen (zelfde dag, dag later).
c. Deel de uitslag telefonisch mede.
d. Schriftelijk per post.
e. Per e-mail.
f. Op publicatiebord van de post.

Stap 13. Registreren gebuikte TINcodes

Nadat de kandidaat is vertrokken worden de door het toetssysteem doorgegeven ToetsIndentificatieNummers (TINcode) geregistreerd, evenals de bij het examen opgetreden bijzonderheden.

Registreren bijzonderheden

- Na afloop van het examen dienen alle bijzonderheden genoteerd te worden die zich hebben voorgedaan.
- In het IEBS staat het scherm voor met de persoongegevens van de kandidaat (zie stap 6). Druk in dit scherm op de knop 'Edit' en daarna op het tabblad 'Remarks'.
- Vul in dit scherm de bijzonderheden in die zich hebben voorgedaan tijdens het examen.
- Hebben zich geen bijzonderheden voorgedaan dan dient hier expliciet de tekst 'geen bijzonderheden' geregistreerd te worden.
- Sla de ingevulde gegevens op met de knop 'Save'.

Registreren gebruikte TINcodes

- Druk op de knop 'Edit' en vervolgens op het tabblad 'Examinations'.
- Druk in het scherm dat nu verschijnt op de knop 'New exam'.
- Registreer de datum en tijd van het examen en de gebruikte TIN codes.

Stap 14. Uitslag van het examen

Nadat het examen is afgerond ontvangt de post, via de inbox van de CA postbus de uitslag van het examen. Dit bericht (e-mail) is onder normale omstandigheden binnen één uur na afloop van het examen op de post aanwezig en refereert naar de TIN-code. Print dit bericht en voeg het in het dossier. Indien er binnen afzienbare tijd nog geen uitslag is ontvangen, dan moet contact worden opgenomen met de helpdesk in Den Haag.

Registreren examenuitslag in het IEBS

- Zoek het examen via functie 'Fill in results' op het hoofdscherm.
- Vul één van de genoemde TINcodes uit het email bericht in en druk op de knop 'Search'.
- Registreer de uitslagen van de toetsen NLT en KNS.
- Sla de gegevens op met de knop 'Save'.

Stap 15. Meedelen resultaat examen aan kandidaat

Medewerker brengt de kandidaat op de hoogte van het resultaat van het examen op de afgesproken wijze. Indien de kandidaat is gezakt, dient hem te worden meegedeeld welk deel/delen van het examen onvoldoende was/waren. De kandidaat wordt dan ook geïnformeerd over de mogelijkheid het gehele examen nogmaals, en weer tegen betaling, te doen waarvoor een nieuwe afspraak gemaakt moet worden.

Bijlage 3

Lijst woordvormen in de TGN[1]

met een frequentie lager dan 17 in CGN (2005)

WOORDVORM	RANK	LEMMA	LEMMAFREQ
oneerlijk	7220	oneerlijk	16
snor	7155	snor	16
tam	7146	tam	16
bliksem	7681	bliksem	15
kleedkamer	7569	kleedkamer	15
omgewaaid	7529	omwaaien	15
progressief	7492	progressief	15
achterdeur	8091	achterdeur	14
bestuurt	8062	besturen	14
dal	8028	dal	14
diamant	8021	diamant	14
figuurlijk	7992	figuurlijk	14
gekrompen	7923	krimpen	14
hamer	7976	hamer	14
inkt	7953	inkt	14
krimpen	7923	krimpen	14
rijkdom	7831	rijkdom	14
smelten	7813	smelten	14
terughalen	7787	terughalen	14
tijger	7779	tijger	14
verkleinen	7760	verkleinen	14
blaas	8454	blaas	13
maandelijks	8293	maandelijks	13
sieraad	8203	sieraad	13
aankomst	8937	aankomst	12
afwezig	8919	afwezig	12
uitgang	8570	uitgang	12
verliezer	8552	verliezer	12
waterleiding	8527	waterleiding	12
alfabet	9436	alfabet	11
dop	9995	dop	10
hoofdzaak	9891	hoofdzaak	10
lente	9805	lente	10
eb	10727	eb	9
oprapen	10428	oprapen	9
rugtas	10349	rugtas	9
kippensoep	11430	kippensoep	8
retourtje	11182	retour	8
vegetariër	11024	vegetariër	8
groentesoep	12585	groentesoep	7
kledingstuk	12460	kledingstuk	7
binnengaan	14277	binnengaan	6

[1] De overige woorden in de TGN komen voor in de lijst van de 7000 meest frequente woorden in het Corpus Gesproken Nederlands

karbonade	13815	karbonade	6
kiespijn	13795	kiespijn	6
lucifer	13683	lucifer	6
beterschap	16243	beterschap	5
mals	15426	mals	5
uitvinder	14685	uitvinder	5
veters	14598	veter	5
vloed	14581	vloed	5
natellen	17767	natellen	4
wijnglazen	16544	wijnglas	4
bijzaak	23386	bijzaak	3
ijverig	22212	ijverig	3
lichaamsdeel	21735	lichaamsdeel	3
omrijden	21255	omrijden	3
optimist	21082	optimist	3
glasbak	29565	glasbak	2
gootsteen	29533	gootsteen	2
legitimatiebewijs	28207	legitimatiebewijs	2
zonneschijn	23862	zonneschijn	2
blijdschap	54630	blijdschap	1
bloemenwinkel	54576	bloemenwinkel	1
pessimist	41307	pessimist	1
rijstrook			0
vulpen			0

Bijlage 4

Descriptoren voor Uitspraak

Definitie van Uitspraak

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

Algemene scoringsregels

- Bij TWIJFEL tussen twee scores: geef altijd de laagste.
- NUL wordt gebruikt voor STILTE of voor een IRRELEVANT of totaal ONBEGRIJPELIJK antwoord.
- NUL wordt ook gebruikt wanneer het aantal woorden waarvan het zeker is dat ze zijn gezegd, minder is dan de helft van het verwachte aantal woorden.

Toelichting

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

Scoreschaal voor Uitspraak

6 "MOEDERTAAL" Uitspraak
Alle C's en V's worden uitgesproken zoals een moedertaalspreker dat doet. De spraak is onmiddellijk en met zekerheid verstaanbaar. De spreker gebruikt reductie, assimilatie en weglating van klanken zoals gebruikelijk is in vlot taalgebruik van moedertaalsprekers. Klemtonen zijn correct.

5 GEVORDERDE Uitspraak
Uitspraak van C's en V's is helder en ondubbelzinnig. Enkele afwijkingen in klank of klemtoon zonder gevolgen voor de begrijpelijkheid. Ieder woord kan gemakkelijk worden verstaan. Klemtonen in gangbare woorden zijn correct.

4 GOEDE Uitspraak
Uitspraak van de meeste C's en V's is correct. Sommige consistente uitspraakfouten leiden bij een aantal woorden mogelijk tot onduidelijkheid. Een paar C's of V's worden mogelijk in bepaalde contexten regelmatig verkeerd uitgesproken of weggelaten. Door het uitspreken van de stomme *e* of andere klinkers die in lopende spraak horen weg te vallen, wordt het klemtoonpatroon mogelijk verstoord.

3 MATIGE Uitspraak
Bepaalde C's en V's worden consistent verkeerd uitgesproken. Spraak is over het algemeen verstaanbaar, maar de toehoorder moet wel wennen aan het accent. Bovendien worden sommige C's regelmatig vervormd of weggelaten en worden consonant clusters vereenvoudigd. Klemtoon is bij sommige woorden verkeerd geplaatst, of onduidelijk.

2 STORENDE Uitspraak
Veel C's en V's worden verkeerd uitgesproken, waardoor een sterk buitenlands accent ontstaat dat het begrip hindert. De toehoorder kan mogelijk een belangrijk deel van de woorden (≥ 33%) niet verstaan. Bovendien worden veel C's vervormd of weggelaten en worden veel consonant clusters vereenvoudigd. De plaatsing van klemtonen is onduidelijk, onbeklemtoonde V's worden niet verkort of juist weggelaten, soms wordt een hele lettergreep toegevoegd of weggelaten.

1 SLECHTE Uitspraak
De uitspraak is eigenlijk volledig die van een andere taal. Veel C's en V's worden verkeerd uitgesproken, verhaspeld of weggelaten. De toehoorder zal in het begin nauwelijks iets kunnen verstaan. Er is weinig verschil tussen beklemtoonde en onbeklemtoonde lettergrepen. Meerdere woorden hebben niet het juiste aantal lettergrepen.

0 GEEN EVIDENTIE
Stilte, of spraak die irrelevant of volledig onverstaanbaar is

Bijlage 5

Descriptoren voor Vloeiendheid

Definitie van Vloeiendheid
Vloeiend en snel kunnen spreken hetgeen blijkt uit passend ritme, frasering, pauzering en woordklemtoon in doorlopende spraak. “Het loopt en klinkt zoals het hoort”.

Begripsdefinitie
Een **langere uiting** is een uiting met negen of meer woorden (≥ 9 woorden)

Algemene scoringsregels

- Bij TWIJFEL tussen twee scores: geef altijd de laagste.
- NUL wordt gebruikt voor STILTE of voor een IRRELEVANT of totaal ONBEGRIJPELIJK antwoord.
- NUL wordt ook gebruikt voor antwoorden met minder dan de helft van het verwachte aantal woorden.
- VERMINDER een score met 1 punt als spraak erg langzaam is; met 2 punten als het extreem langzaam is.

Scoreschaal voor Vloeiendheid

6 “MOEDERTAAL” Vloeiendheid
De uiting van de kandidaat vertoont vloeiend ritme en frasering als van een moedertaalspreker, zonder aarzelingen, herhalingen, valse starts of onnatuurlijke reductie van fonemen.

5 GEVORDERDE Vloeiendheid
De uiting van de kandidaat heeft een aanvaardbaar ritme met de juiste frasering en woordklemtoon. Uitingen bevatten hooguit een enkele hapering, herhaling of valse start. Opvallend onnatuurlijke fonologische reductie komt niet voor.

4 GOEDE Vloeiendheid
De uiting van de kandidaat heeft een aanvaardbaar tempo, maar kan wat onevenwichtig klinken. Langere uitingen kunnen meer dan een enkele aarzeling bevatten, maar de meeste woorden worden in lopend zinsverband uitgesproken. Er zijn weinig herhalingen of valse starts. Er vallen geen lange pauzes en de spraak klinkt niet stoterig.

3 MATIGE Vloeiendheid
De uiting van de kandidaat kan wat onevenwichtig of stoterig klinken. Uitingen met 6 of meer woorden bevatten tenminste een opeenvolging van drie vloeiend lopende woorden en niet meer dan twee of drie aarzelingen of valse starts. Er kan een langere pauze voorkomen, maar niet twee of meer.

2 BEPERKTE Vloeiendheid
De uiting van de kandidaat vertoont onregelmatige frasering of zinsritme. Gebrekkige frasering, stoterig verloop - per lettergreep - en/of veelvuldige aarzelingen, herhalingen of valse starts maken de uiting duidelijk onevenwichtig en onderbroken. Langere uitingen kunnen een of twee lange pauze bevatten en vertonen mogelijk onjuiste woord- of zinsklemtoon.

1 NIET VLOEIEND
De spraak van de kandidaat is langzaam en verloopt moeizaam, met nauwelijks enige frasering en veelvuldig aarzelingen, herhalingen, valse starts, en/of grove fonologische reducties. De meeste woorden worden los uitgesproken en er kunnen meer lange pauzes voorkomen.

0 GEEN EVIDENTIE
Stilte, of spraak die irrelevant of volledig onverstaanbare is.

Bijlage 6

Toets Gesproken Nederlands
Instructies voor de kandidaat

Wat moet u doen?
Lees rustig deze aanwijzingen door, dan weet u hoe de toets gaat. Als u iets niet begrijpt, vraagt u dan om uitleg.
U kunt dit formulier gebruiken als u de computer gaat bellen. Dat is niet echt nodig. De computer zegt straks precies wat u moet doen.

LET OP

*Misschien vindt u de toets straks erg moeilijk. Het kan bijvoorbeeld dat U niet alles goed kunt verstaan. Of misschien kent u niet alle woorden. Dat betekent NIET dat u gezakt zou zijn. De toets is bedoeld voor **alle** leerders van het Nederlands. Niet alleen voor beginners, maar ook voor gevorderden. Mensen die al heel goed Nederlands beheersen, halen natuurlijk de hoogste scores. De scores van beginners zijn veel lager. Schrik dus niet van een moeilijke opgave. Ga gewoon door. U hoeft niet alles te weten om de toets goed te doen!*

De toets bestaat uit een Inleiding en vijf onderdelen (onderdeel A, B, C, D en E). Bij sommige opgaven zijn meer antwoorden goed.

Dit gebeurt bij de Inleiding

Stap	Dit is wat u moet doen	Dit gebeurt er	Dit is wat u hoort
1	Toets het telefoonnummer in. *(Staat **bovenaan** uw testformulier)*	De computer antwoordt in het Nederlands.	*Dank u voor het bellen met het toetssysteem van Ordinate. Toets uw Toets Identificatie Nummer in.*
2	Toets op de telefoon uw **TIN-code** in. *(Staat bovenaan uw testformulier)*	Nu gaat de computer controleren of de verbinding goed is.	*Welkom bij het inburgering-examen Nederlandse taal. We testen eerst het geluid. Zeg alstublieft de naam van de stad en van het land waar u nu bent.*
3	Zeg duidelijk de stad en het land. *U mag ook een andere stad en land noemen.* *Als u niet hard genoeg spreekt, wordt gevraagd of u het nog eens wilt doen.*	De computer controleert of uw stem duidelijk overkomt. Als dat goed gaat begint de toets.	*Volg nu de aanwijzingen voor Onderdeel A tot en met E.* *Let op we beginnen nu met de toets.*

Daarna begint de toets vanzelf.
Kijk op de volgende pagina voor voorbeelden van opgaven.

Voorbeelden van opgaven

Deel A: Nazeggen
Nazeggen. U hoort steeds een zin. Zeg de zin precies na.

Bijvoorbeeld: een stem zegt *"Dat is een mooi verhaal"*
en u zegt: *"Dat is een mooi verhaal"*.

Probeer niet alleen de woorden maar ook de manier van spreken precies na te doen. Spreek vlot, en aarzel niet.

Deel B: Vragen
U hoort steeds een korte vraag. Geef op elke vraag een kort antwoord.

Bijvoorbeeld: een stem zegt *"Als je thee zet, gebruik je dan heet water of gebruik je koud water'?*
En u zegt: *"heet water"* of *"heet"*.

Of u hoort: *"Een auto, heeft die twee wielen of vier wielen?"*
En u zegt: *"vier" of "vier wielen"*.

Deel C: Nazeggen (hetzelfde als onderdeel A)
U hoort steeds een zin. Zeg de zin precies na.

Bijvoorbeeld: een stem zegt: *"Dat is een mooi verhaal"*
en u zegt: *"Dat is een mooi verhaal"*.

Deel D: Tegenstellingen
U hoort steeds een woord. U zegt het tegenovergestelde.

Bijvoorbeeld: u hoort *"hoog"*
dan zegt u *"laag"*.

Of u hoort *"niet"*
dan zegt u *"wel"*

Onderdeel E: Verhalen navertellen
U hoort een kort verhaal. U moet het verhaal navertellen. U krijgt daarvoor 30 seconden. Luister goed. Vertel zoveel mogelijk.

U moet twee verhaaltjes navertellen.

Aan het eind hoort u: "*Dank u voor het bellen. U kunt nu ophangen".*
Dan kunt u ophangen.

LET OP:

U moet de toets helemaal afmaken. Als u de telefoon voor het einde van de toets neerlegt, telt de toets niet mee. Zorg dat de computer weet dat u er bent. Zeg dus altijd iets. Als u niet weet hoe u moet antwoorden op een vraag, zeg dan bijvoorbeeld: "ik weet het niet" of "nee".

Aan de telefoon

- **Gebruik een vaste telefoon met druktoetsen. U kunt niet met een mobiele telefoon bellen.**
- **Kies een rustige plek om te bellen. Zorg dat u niet gestoord wordt.**
- **Het is belangrijk dat uw stem goed overkomt. Spreek daarom luid en duidelijk.**
- **Houd de telefoon goed voor uw mond. Kijk naar het plaatje.**

NEE
hoorn te laag en te veraf

JA
hoorn voor de mond

JA
hoorn op juiste afstand

- **Als uw stem niet goed overkomt, zal de computer u vragen de telefoon neer te leggen. U kunt dan opnieuw gaan bellen. U moet dan eerst een nieuwe TIN-code aan uw docent vragen.**

.

SUCCES!

Bijlage 7

Ontwikkeling Inburgeringstoets NT2
Instructies voor docenten: NT2-leerders

Voor nadere informatie en vragen:
Anne Kerkhoff (akerkhoff@cinop.nl), Anne Toorenaar (atoorenaar@cinop.nl),
Miranda de Jong (mjong@cinop.nl)
Telefoon: 073 6800724

LET OP: het lijkt ingewikkelder en tijdrovender dan het zal zijn.

Voor de toetsafname: materialen voor NT2-leerders.
U ontvangt van CINOP een pakket met de noodzakelijke materialen. CINOP levert steeds pakketten met setjes materiaal voor 25 kandidaten. U hoeft niet alle setjes te gebruiken. Ongebruikte setjes kunt u aan collega's doorgeven, of aan ons terugsturen. We sturen u 25 setjes, omdat het mogelijk is dat kandidaten na een mislukte poging om de toets te doen, u om een nieuwe TIN-code vragen. U kunt dan een compleet nieuw setje materialen uitreiken. Vraag dan wel om het setje van de mislukte poging in te leveren. Zet daar dan een groot kruis doorheen voordat u het aan ons terugstuurt. Vergeet niet het registratieformulier aan te passen!

In het volgende overzicht is de inhoud van de pakketten beschreven.

Pakketten voor docenten met NT2-leerders

Materiaal voor docenten:
- Een tekst 'Algemene achtergrondinformatie voor de docent'
- De onderhavige tekst 'Instructies voor docenten: NT2-leerders'
- Een registratieformulier
- 25 vragenlijsten met ieder twee TIN-codes

Materiaal voor NT2-leerders.
Elke NT2-leerder krijgt een setje van vier documenten:
- Een tekst 'Informatie voor de cursist'
- Een blad met een telefoonnummer, twee TIN-codes en wat algemene aanwijzingen
- Een tekst 'Instructies bij de toets Gesproken Nederlands'
- Een toetsblad bij de toets Geletterdheid

LET OP:
De mondelinge toets bevat voor moedertaalsprekers meer opgaven dan voor NT2-leerders. Verwar de materialen daarom niet.

Voorbereiding
- U kiest een cursistengroep die aan de toets zal deelnemen. CINOP rekent op een gemiddelde groepsgrootte van 15 deelnemers. Natuurlijk is elke kandidaat extra zeer welkom. Indien dat niet anders kan, kunt u ook met een kleinere groep deelnemen. Of met twee groepen.

LET OP: elke NT2-leerder neemt bij voorkeur aan beide toetsen deel (en ontvangt daarvoor 2 keer 10 euro). Ook echtgenoten, buurmannen en vrienden van uw cursisten (minimum leeftijd: 15 jaar) kunnen in principe meedoen. Voorwaarde is wel dat u beschikt over de gevraagde gegevens over de achtergrond van de kandidaten en over hun taalvaardigheidsniveau in het Nederlands.

- U kent elke NT2-leerder twee TIN-codes toe. U doet dat door de naam van de cursist in te vullen op het blad met het telefoonnummer en de twee TIN-codes. Eventueel kunt u ook de andere documenten van het setje voorzien van de naam van de cursist. U kunt dat uiteraard ook door de cursisten zelf laten doen.
- Registreer wie welk setje krijgt. Dat kan via het registratieformulier. Het kan ook door erop toe te zien dat de kandidaten hun naam op alle formulieren met de TIN-codes zetten. U kunt dan achteraf, door de stroken met naam en TIN-code van de bladen te knippen en aan elkaar te nieten, alsnog administreren wie er heeft meegewerkt onder welke TIN-code. *LET OP: Het enige dat telt is dat u zelf weet welke kandidaat welke TIN-codes zal gebruiken. U moet immers bij elke TIN-code de juiste achtergrondgegevens kunnen invullen. Bovendien wilt u straks weten wie er recht heeft op een vergoeding. Verwijder de namen van cursisten voor inzending van alle formulieren om zo de privacy van de kandidaten te garanderen. CINOP heeft alleen de TIN-codes nodig om de toetsresultaten van de cursisten te kunnen koppelen aan hun achtergrondgegevens. CINOP heeft geen namen nodig.*

Instructie aan de kandidaten

- Geef elke cursist zijn eigen setje. Als u nog geen namen van cursisten hebt ingevuld, kunt u de cursisten nu vragen om dat zelf te doen. Maak duidelijk dat die namen NIET aan CINOP of anderen worden doorgegeven.
- Behandel met de cursisten het formulier 'Informatie voor de kandidaten'. Benadruk het feit dat deelname aan de toetsen geheel anoniem is. Indien gewenst kunnen kandidaten eind september 2004 via CINOP een uitslag krijgen. Zij moeten daartoe hun TIN-code bewaren.
  Benadruk dat kandidaten de toetsen zelf moeten afleggen, en dat ze zich niet laten bijstaan door anderen. Als de kandidaten zich te goed voorbereiden op de toets, zal de toets gemakkelijker lijken dan hij is. Daardoor zullen er straks hogere eisen gesteld worden aan kandidaten.
- Bespreek met de cursisten het formulier 'Instructies voor de kandidaten bij de toets Gesproken Nederlands'.
- Bespreek met de cursisten het toetsblad bij de toets 'Geschreven Nederlands'. Benadruk vooral bij minder goede lezers dat het van belang is dat de cursisten de toets ZELF doen. Als zij zich nu laten helpen, zal de toets makkelijker lijken dan hij is. Dat zal ertoe leiden dat er straks hogere eisen aan kandidaten worden gesteld.
- Bespreek tot slot het blad met het telefoonnummer en de TIN-codes.
- Benadruk dat de cursisten als ze de toetsen hebben afgelegd, alle materialen die ze hebben ontvangen moeten inleveren. Ze krijgen dan, wanneer CINOP heeft gecontroleerd of de toetsen helemaal zijn afgemaakt, in ruil voor hun inspanning de cadeaubon(nen).

Vragenlijstjes met aanvullende gegevens

- U ontvangt van CINOP voor elke kandidaat een vragenlijstje.
- LET OP: op de vragenlijsten staan steeds twee TIN-codes. Zorg dat de juiste naam is ingevuld. Dat wil zeggen dat CINOP straks de juiste achtergrondgegevens kan koppelen aan degenen die de toetsen met de bijbehorende TIN-codes heeft gemaakt.
- Beantwoord de vragen. LET OP: U hoeft geen extra NT2-toetsen af te nemen. U vult alleen gegevens in die u al in uw bezit hebt doordat de kandidaat onlangs, maar uiterlijk na 1 december 2003, een gestandaardiseerde toets heeft afgelegd.

- Bij de eerste vraag over het ‘CEF-niveau’ van de kandidaat vult u uw eigen oordeel in. Als u het CEF onvoldoende kent, vult u uw oordeel in aan de hand van de bekende ‘NT2-schaal’: 1, 2, 3… A1 staat dan voor 1, A2 voor 2, et cetera.

Als cursisten een nieuwe TIN-code hebben gekregen

- Het is mogelijk dat een pretest mislukt, doordat bijvoorbeeld de telefoonverbinding wordt verbroken voor het einde van de toets. De cursist kan het opnieuw proberen. Hij/zij moet dan echter een nieuwe TIN-code krijgen. TIN-codes kunnen maar één keer gebruikt worden.
- U geeft de cursist dan een nieuw setje. De cursist krijgt dan TWEE nieuwe TIN-codes. Hij/zij hoeft er daarvan natuurlijk maar één te gebruiken: die van de mislukte toets.
- Omdat CINOP alleen TIN-codes kent, is het nodig dat u in zo’n geval NOG eens de vragenlijst voor de cursist invult: nu de lijst met de **nieuwe** TIN-codes. CINOP krijgt in zo’n geval dus twee vragenlijsten met dezelfde gegevens over dezelfde cursist, maar met verschillende TIN-codes.
- Controleer ten slotte heel goed of uw registratieformulier nog klopt.

Afronding

- U verzamelt alle materialen: de vragenlijsten en de vier documenten die u aan de cursisten hebt uitgereikt.
- U stuurt CINOP alle materialen retour. Verwijder eerst de namen van de cursisten. CINOP is alleen geïnteresseerd in de TIN-codes en anonieme achtergrondgegevens.
- U ontvangt van CINOP een lijst met TIN-codes van de kandidaten die de toets hebben afgelegd en voor elke complete toets die is afgelegd een VVV-bon van 10 euro.
- U geeft de betrokken kandidaten hun vergoeding.

Bijlage 8

Ontwikkeling Inburgeringstoets NT2
Algemene informatie voor docenten

Voor nadere informatie en vragen:
Anne Kerkhoff (akerkhoff@cinop.nl), Anne Toorenaar (atoorenaar@cinop.nl)
Miranda de Jong (mjong@cinop.nl)

Telefoon: 073-6800724

Achtergronden

- Er worden twee instrumenten ontwikkeld: een toets 'Gesproken Nederlands' en een toets 'Geletterdheid'. De toets 'Gesproken Nederlands' toetst luister- en spreekvaardigheid en heeft een bereik van A1 tot en met B2 (eventueel C2). De toets 'Geletterdheid' toetst of iemand gealfabetiseerd is in het Latijnse schrift. De toets bestaat uit een aantal korte woorden, zinnen en een tekstje dat kandidaten moeten voorlezen. Het is nog zeer onzeker of de toets 'Geletterdheid' gebruikt gaat worden. Besluitvorming daarover vindt in de zomer van 2004 plaats in de Tweede Kamer.
- De toetsen worden afgenomen per telefoon. Bij de beoordeling wordt gebruik gemaakt van automatische spraakherkenning.
- De toetsen worden door CINOP ontwikkeld in opdracht van het ministerie van Justitie. CINOP werkt samen met LTS in Velp en het Amerikaanse bedrijf Ordinate.
- De toetsen zullen in het buitenland door het ministerie van Justitie gebruikt worden als 'toelatingsexamen' in het kader van aanvragen voor een machtiging tot voorlopig verblijf bij gezinshereniging en gezinsvorming. Over de precieze eisen die gesteld gaan worden, moet de Tweede Kamer in de zomer een besluit nemen. Het advies van de commissie Franssen luidt: geen eisen aan geletterdheid en voor de mondelinge vaardigheden niveau 'A1-min'.
- In het contract tussen CINOP en het ministerie van Justitie is vastgelegd dat de toetsen ontwikkeld worden in het kader van het streven naar een inburgeringexamen in Nederland dat gebaseerd is op de Europese portfoliomethodiek.
- Om een indruk te krijgen van de te ontwikkelen mondelinge toets, kunt u gratis een vergelijkbare toets Engels proberen via www.ordinate.com.

Samenvatting van de procedure

Ten behoeve van de constructie van de toetsen en de bouw van de spraakherkenner dienen alle opgaven te worden getest door allochtone en autochtone sprekers van het Nederlands. De kandidaten leggen de toetsen af per telefoon. Elke telefoon met een vaste lijn kan gebruikt worden. **Er kan helaas GEEN gebruik gemaakt worden van mobiele telefoons.** Ten behoeve van de validering van de toets zijn aanvullende gegevens over de achtergronden van de kandidaten nodig. De gegevens over de allochtone kandidaten worden door de docenten ingevuld. Autochtone kandidaten zullen de betreffende vragenlijst meestal zelf invullen. Natuurlijk is het docenten toegestaan om autochtone kandidaten daarbij te helpen.

Privacy
Om de privacy van de kandidaten te beschermen, wordt er gewerkt met TIN-codes (TIN staat voor 'Toets Identificatie Nummer'). De docent en de kandidaten zijn de enige personen die TIN-codes kunnen verbinden aan namen van kandidaten. CINOP, LTS en Ordinate beschikken NIET over namen van kandidaten. Het is dan ook uitgesloten dat gegevens over individuele kandidaten in handen van derden komen.

Wat houdt de medewerking voor uw deelnemers in?

- Alle deelnemers krijgen alle noodzakelijke informatie over de twee toetsen en de wijze van afname op papier. Allochtone cursisten krijgen daar van hun docent klassikaal een korte uitleg bij. U geeft die uitleg aan de hand van de onze instructies. Waar nodig kunt u uiteraard ook autochtone cursisten helpen met de instructies. Voor de meeste autochtone kandidaten zal dat echter niet nodig zijn.
- De deelnemers krijgen twee persoonlijke TIN-codes en één telefoonnummer dat gratis gebeld kan worden.
- Via het gratis telefoonnummer bellen de deelnemers twee keer met de computer van Ordinate.
- De eerste keer toetsen ze de TIN-code voor de toets 'Gesproken Nederlands' in. Ze krijgen daarop via de computer telefonisch hun opgaven te horen. Ze spreken hun antwoorden via de telefoon in. In totaal kost dat naar schatting maximaal 15 minuten.
- De opgaven voor de schriftelijke toets hebben de deelnemers via u ontvangen. Ze houden het blad met de opgaven bij de hand als ze voor de tweede keer bellen. Nadat ze hun TIN-code voor de schriftelijke toets hebben ingetoetst, krijgen ze via de telefoon instructies om de opgaven op het toetsblad voor te lezen.
- De deelnemers kunnen de toets op elk gewenst moment voor 10 mei en op elke gewenste plaats afnemen. De enige beperking is dat zij gebruik moeten maken van een vaste telefoonlijn, *dus geen mobiele telefoon.*

Deelnemers ontvangen voor hun deelname per afgelegde toets een VVV-bon ter waarde van 10 euro. Deelnemers die beide toetsen doen, ontvangen dus bonnen ter waarde van 20 euro. Deelnemers krijgen alleen een vergoeding als ze alle opgaven hebben gedaan. *Een toets die niet is afgemaakt, is niet geldig.* Iedere reactie waaruit blijkt dat de deelnemer de opgave beluisterd heeft – ook reacties als 'ik weet het niet' - geldt als een geldig antwoord. Deelnemers die de verbinding met de computer voor het einde van de toets verbreken, dan wel helemaal niets zeggen, krijgen geen vergoeding. CINOP kan via de computer van Ordinate controleren of de toetsen zijn afgemaakt.
LET OP: CINOP zal t.z.t de uitbetaling in de vorm van cadeaubonnen per TIN-code aan u zenden. U moet dus zelf het verband tussen TIN-code en de namen van cursisten registreren t.b.v de uitbetaling.

Wat houdt uw medewerking in?
Docenten leveren op twee manieren een bijdrage aan de ontwikkeling van de toetsen:

- Docenten bemiddelen tussen kandidaten en CINOP: ze benaderen de kandidaten, informeren hen en reiken de vergoeding uit.
- Docenten vullen voor CINOP een korte vragenlijst in met aanvullende gegevens over de allochtone deelnemers. Docenten hoeven daarvoor geen extra toetsen af te nemen: het gaat uitsluitend om toetsgegevens die reeds beschikbaar zijn doordat de cursisten deze toetsen al hebben afgelegd.
- Voor uw eigen informatie is er ruimte om aan de bovenzijde van het formulier de naam van de betrokken cursist in te vullen. Wij verzoeken u echter dringend deze naam voor inzending aan CINOP te verwijderen. De gegevens worden gebruikt om de toetsen te valideren. Het gaat om achtergrondgegevens en gegevens over de taalvaardigheid in het Nederlands van de deelnemers. Autochtone deelnemers zullen het vragenlijstje meestal zelf invullen.

Bijlage 9

Inburgeringstoets NT2

Algemene informatie voor de kandidaten

Uitgangspunten

- U werkt mee aan de ontwikkeling van een toets Nederlands als tweede taal ('Nederlands voor buitenlanders').
- De toets heeft twee delen: een toets 'Gesproken Nederlands' en een toets 'Geschreven Nederlands'. De mondelinge toets test luisteren en spreken. De schriftelijke toets test of u Nederlandse woorden en zinnen hardop kunt voorlezen.
- De toetsen worden afgenomen via de telefoon.
- De toetsen worden ontwikkeld door CINOP in samenwerking met LTS en Ordinate.
- De toetsen worden ontwikkeld in opdracht van het ministerie van Justitie.

Wat houdt uw medewerking in?

- U maakt via een telefoon gratis twee korte toetsen. U kunt de toetsen op school afleggen, of thuis. U kunt elke **vaste** telefoon gebruiken, maar geen **mobiele** telefoon. U kunt altijd bellen, 7 dagen per week, 24 uur per dag.

Wat hebt u nodig om de toetsen te doen?

- U krijgt van uw docent drie formulieren:
  - een blad met het telefoonnummer dat u moet bellen en twee TIN-codes;
  - een blad met instructies bij de toets Gesproken Nederlands;
  - een blad met instructies en opgaven bij de toets Geschreven Nederlands.

Hoeveel tijd kost u dit?

- Uw deelname aan de mondelinge toets kost ongeveer 20 minuten.
- Uw deelname aan de schriftelijke toets kost ongeveer 10 minuten.

Wat levert het op?

- U werkt mee aan de constructie van een belangrijke toets.
- Als u de toets helemaal afmaakt, dat wil zeggen op alle opgaven reageert, ontvangt u via uw docent een cadeaubon van 10 euro per toets. Als u allebei de toetsen afmaakt, krijgt u dus voor 20 euro cadeaubonnen.

Wat gebeurt er met uw gegevens? Blijft u anoniem?

- Deelname aan de toets is helemaal **anoniem**. *Niemand krijgt van uw docent of school gegevens over uw naam of adres.* CINOP, LTS en Ordinate kennen alleen de TIN-codes. De opdrachtgever, het ministerie van Justitie, krijgt geen enkel gegeven over TIN-codes, kandidaten of scholen.

Bijlage 10

Vragenlijst Pretest

In te vullen door docent

Naam Cursist: -- 6359 4096

Afknippen voor inzending aan CINOP 27728605

TINcode Mondeling	6359 4096
TINcode Schriftelijk	27728605

CursistNummer		
Opleidingsniveau	| 0-6 jaar | 7-12 jaar | Meer dan 12 jaar	← Aantal jaren onderwijs in land van herkomst
Alfagroep	Ja | Nee	
Geboorteland		
Geboortejaar		
Geslacht	m | v	
Jaren in Nederland		
UitstrooomPerspectief	Professioneel | Sociaal | Educatief	

CEF - niveau	Gesprekken voeren *(schatting docent)*	Onder A1 / A1/ A2 / B1 / B2 / C1 / C2		
	Lezen *(schatting docent)*	Onder A1 / A1/ A2 / B1 / B2 / C1 / C2		
Staatsexamen I	Lezen		Score:	
	Luisteren		Score:	
	Spreken		Score:	
	Schrijven		Score:	
Staatsexamen II	Lezen		Score:	
	Luisteren		Score:	
	Spreken		Score:	
	Schrijven		Score:	
Profieltoets	Lezen	Hoog / Laag	Uitslag:	
	Luisteren	Hoog / Laag	Uitslag:	
	Spreken	Hoog / Laag	Uitslag:	
	Schrijven	Hoog / Laag	Uitslag:	
	Aantal uren NT2 op deze ROC:			
NIVOR	Lezen	Niveau:	Score:	
	Luisteren	Niveau:	Score:	
	Spreken	Niveau:	Score:	
	Schrijven	Niveau:	Score:	
ICE-Traject	Lezen	Serie L / S	Score:	
	Luisteren	Serie L / S	Score:	
	Spreken	Serie L / S	Score:	
	Schrijven	Serie L / S	Score:	
ItemDito	Lezen	Niveau:	Score:	
	Spreken	Niveau:	Score:	
	Schrijven	Niveau:	Score:	
NT2CAT	Lezen	Niveau:	Score:	
	Luisteren	Niveau:	Score:	

8389 NN-Literacy-1 8044 NN-DUTCH-75 001 00 001

Bijlage 11

INTERVIEWPROTOCOL

Mondelinge Interactie

Dataverzameling Amsterdam

Protocol voor interview ten behoeve van beoordeling

Interviewers maken tijdens gesprek aantekeningen op het Beoordelingsformulier Interviewer.
Beoordelaars maken tijdens gesprek aantekeningen op het Beoordelingsformulier Beoordelaar.
Interviewers moeten het idee geven dat zij daadwerkelijk geïnteresseerd zijn in de antwoorden van de kandidaat: zij willen echt graag de antwoorden horen en maken hiervan ook een notitie.

Beoordelaars mengen zich niet in het gesprek. Zij nemen een onopvallende positie in, observeren en noteren hun observaties.

Het gesprek is "adaptief" opgebouwd. Het is niet nodig met alle kandidaten alle rubrieken af te werken. Als het bij 4 al heel moeilijk gaat, dan 5 en 6 overslaan. Gaat het bij 4 redelijk goed, dan 5 proberen. Lukt dat ook wel, dan door naar 6.

Na afloop van het gesprek, wanneer de kandidaat vertrokken is, blijft het stil. Interviewers en beoordelaars noteren ieder afzonderlijk hun globale oordeel. Dat oordeel is definitief. Pas na het vastleggen hiervan bespreken interviewer en beoordelaar onderling eventueel hun bevindingen. Eenmaal ingevulde oordelen worden echter NIET meer gewijzigd.

Wederzijdse begroeting. De interviewer begroet de kandidaat
1.1 De kandidaat groet terug.

Kennismaking: Introductie van de aanwezigen door interviewer (indien nog niet bekend).
2.1 Interviewer vraagt naar naam en persoonsgegevens kandidaat.

3 Introductie
(*Toelichting op het doel van het gesprek: de telefoontoets evalueren. Wij willen graag weten wat de kandidaten vinden van de toets en of de toets goed meet.)*

4 Vragen stellen over achtergrond. Interviewer vraagt naar
4.1 land van herkomst,
4.2 taalachtergrond,
4.3 lengte verblijf in Nederland,
4.4 familie of andere contacten in Nederland,
4.5 hoeveelheid Nederlandse les

5 Vragen naar ervaringen met de telefoontoets Interviewer vraagt
5.1 of alles goed verlopen is.
5.2 of de kandidaat de thuistoets gemaakt heeft, en wanneer.
5.3 of de toetsen op school gemakkelijker waren dan de thuis gemaakt toets.
5.4 welk deel/welke delen van de toets de kandidaat het moeilijkst vond en waarom.

6 Vragen naar mening over een aantal onderwerpen. Interviewer vraagt
6.1 of de kandidaat denkt dat de toets echt spreekvaardigheid toetst.
6.2 wat de kandidaat vindt van de stad Amsterdam.
6.3 wat de kandidaat het meeste opviel bij aankomst in Nederland
6.4 wat de kandidaat verwacht van de Nederlandse les hier op het ROC.

7 Afsluiting: Interviewer neemt afscheid van de kandidaat, dankt en reikt cadeaubon uit
7.1 kandidaat bedankt en of neemt afscheid.

Bijlage 12 BEOORDELINGSSCHAAL GESPREKSVAARDHEID

Aanwijzingen voor de Beoordeling van gespreksvaardigheid met de CEF-schaal

Vooraf: De afstand van helemaal geen kennis van een taal tot volledige beheersing is heel erg groot. Er zijn niet veel mensen die een tweede of vreemde taal ooit volledig leren beheersen. Wanneer we die grote afstand tussen helemaal geen beheersing en volledige beheersing verdelen in vier, zes of zelfs acht niveaus, blijft ieder van die niveaus nog een behoorlijke grote afstand overbruggen. Het kan voor sommige mensen dan ook jaren duren om van een bepaald niveau naar een hoger niveau te komen. Ook binnen een niveau is de afstand nog groot. Mensen bovenin dat niveau, die bijna aan het volgende niveau toe zijn, kunnen heel wat meer dan mensen die nog maar net de ondergrens van dat niveau zijn gepasseerd. Binnen een niveau komen dus nog grote onderlinge verschillen voor.

Niveaubepaling Bij het bepalen van het niveau kunnen we het beste in twee stappen te werk gaan.

Stap 1 Bepaal eerst het niveau op grond van de hoofdindeling:

Globale observatie	Kenmerken	Karakteristiek	CEF-Niveau
Gesprekken met personen op dit niveau verlopen moeiteloos. Ze spreken vlot en hebben geen of een nauwelijks merkbaar buitenlands accent.	Communicatie verloopt wat de taal betreft probleemloos en zonder misverstanden - over alle onderwerpen; - zeer vlot en gemakkelijk; - nauwelijks of geen grammaticale fouten; - zeer grote woordenschat; - nauwelijks of geen accent.	Vaardig	C
In gesprekken met personen op dit niveau hoeft men niet of nauwelijks rekening te houden met hun taalvaardigheid. Zij kunnen zich zelfstandig redden in het Nederlands. Een eventueel buitenlands accent is niet storend voor de communicatie.	Communicatie verloopt wat de taal betreft met weinig of geen moeite soms is wat nadere precisering, herhaling of omschrijving noodzakelijk - over vertrouwde onderwerpen; - redelijk normaal tempo; - zonder storende fouten; - een ruime woordenschat binnen eigen ervaring; - zonder storend accent.	Onafhankelijk	B
Met personen op dit niveau kan men een gesprek voeren als men maar rekening houdt met hun taalniveau: langzaam praten, herhalen. Vanwege hun buitenlands accent kunnen zij soms minder goed verstaanbaar zijn.	Communicatie is afhankelijk vaan de bereidheid tot medewerking van de gesprekspartners - over alledaagse, bekende onderwerpen; - met inspanning en een aangepast spreektempo; - grammaticale fouten komen veel voor; - het vocabulaire is beperkt; - de uitspraak kan soms hinderlijk zijn voor het begrip.	Essentieel	A
Met personen op dit niveau is het niet mogelijk een gesprek te voeren. Zij begrijpen weinig of niets en zijn zeer moeilijk of niet te verstaan.	Communicatie is indien al mogelijk zeer beperkt en geheel afhankelijk van de inspanningen van de gesprekspartners - over een zeer beperkt aantal feitelijk zaken; - in een zeer laag tempo met duidelijke articulatie; - alleen in losse woorden zonder grammaticale structuur; - vocabulaire is beperkt tot enkele woorden; - door de uitspraak zijn personen vaak onverstaanbaar.	Rudimentair of Nul	Onder A

Stap 2 Op de volgende pagina staan meer gedetailleerde beschrijvingen. De hoofdniveaus staan aangegeven in de meest linkse kolom. Ieder hoofdniveau is ingedeeld in twee subniveaus. Zoek bij het in Stap 1 gekozen hoofdniveau het subniveau waarvan de beschrijving het beste past bij uw indruk van de gespreksvaardigheid van de cursist.

Beoordelingsschaal Gespreksvaardigheid op de CEF-Niveaus

C	**C2**	***Brengt betekenisnuances nauwkeurig en op natuurlijke wijze over*** Kan spontaan en met een natuurlijke vloeiendheid ook langere interventies verrichten. Vertoont daarbij een consistente grammaticale en fonologische beheersing van gevarieerd en complex taalgebruik met inbegrip van een juist gebruik van verbindingswoorden en voegwoorden. Kan moedertaalsprekers moeiteloos verstaan.
	C1	***Drukt zich vloeiend en spontaan uit in duidelijke, goedgestructureerde spraak.*** Kan zich spontaan en vloeiend uitdrukken, bijna moeiteloos in een gelijkmatig lopend taalgebruik. Heeft een duidelijke en natuurlijke uitspraak. Kan intonatie variëren en gebruikt klemtoon om delen te benadrukken. Maakt zelden fouten. Vertoont beheersing van verbindingswoorden en voegwoorden. Verstaat praktisch iedere moedertaalspreker, maar moet wellicht soms om bevestiging vragen.
B	**B2**	***Brengt informatie en standpunten helder en zonder merkbare moeite over.*** Kan eenheden taal met een redelijk evenwichtig tempo produceren met weinig merkbare pauzes. Heldere uitspraak en intonatie. Fouten leiden niet tot misverstanden. Helder, samenhangend betoog, echter soms enigszins "springerig". Kan in detail standaard moedertaalsprekers ook in een lawaaierige omgeving verstaan.
	B1	***Communiceert begrijpelijk de belangrijkste punten m.b.t. vertrouwde zaken.*** Kan op begrijpelijke wijze doorspreken, hoewel evident pauzerend voor het plannen en herstellen van grammaticale en lexicale elementen Uitspraak is begrijpelijk hoewel bij tijden gekleurd door een buitenlands accent en ook uitspraakfouten optreden. Redelijk correct gebruik van een algemeen repertoire in voorspelbare situaties. Kan eenvoudige losse elementen verbinden tot een samenhangend geheel. Kan duidelijk sprekende moedertaalsprekers volgen, maar moet soms om herhaling vragen.
A	**A2**	***Communiceert basisinformatie over werk, achtergrond, familie, vrije tijd, etc.*** Kan zichzelf in korte zinnen verstaanbaar maken, hoewel pauzes, valse starts, en herformuleringen evident aanwezig zijn Uitspraak is over het algemeen helder genoeg om te worden verstaan ondanks een duidelijk buitenlands accent. Gebruikt een beperkt aantal eenvoudige structuren correct, maar maakt systematisch elementaire fouten. Kan woordgroepen verbinden met eenvoudige voegwoordeen zoals "en", "maar", en "omdat". Kan zich tot hem/haar richtende, duidelijk sprekende moedertaalsprekers verstaan, wanneer zonodig om herhaling gevraagd kan worden.
	A1	***Doet eenvoudige uitspraken over persoonlijke gegevens en bekende onderwerpen.*** Kan omgaan met zeer korte, geïsoleerde, voornamelijk standaard uitingen. Veel pauzes om te zoeken naar uitdrukkingen en om minder bekende woorden uit te spreken. Spreekt met sterk buitenlands accent. Begrijpt de strekking van direct tot hem/haar gerichte en duidelijk gesproken vragen.
Onder A	**A1-min**	***Kan met behulp van losse woorden zaken van direct persoonlijk belang communiceren.*** Gebruikt losse woorden, enkele standaarduitdrukkingen en elementaire beleefdheidsfrases maar is vanwege uitspraak moeilijk te verstaan. Begrijpt eenvoudige direct tot hem/haar gerichte en met zorg gesproken vragen naar of mededelingen over personalia, en een beperkt aantal concrete alledaagse begrippen. Kan vragen over dergelijke zaken soms ook met een of meer losse woorden beantwoorden. Conversatie is echter niet mogelijk
	***Onder*A1-min**	***Spreekt op een niveau dat lager is dan het bij A1-min beschreven niveau. Zou als toerist niet zonder hulp kunnen 'overleven'.*** Beheerst zo weinig woorden en/of uitdrukkingen dat verbale communicatie niet mogelijk is. Kan wellicht met veel hulp en begrip van de gesprekspartner enkele vragen naar eigen naam en adres of andere persoonlijke gegevens begrijpen. Kan dergelijke vragen echter meestal niet beantwoorden. Ook basishandelingen, zoals het maken van een afspraak of het geven of begrijpen van eenvoudige routebeschrijvingen, zijn niet mogelijk, hoewel soms met veel gebaren enige communicatie kan worden bereikt.

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

Bijlage 13

EFFECTEN VAN DE KEUZE VAN DE CESUUR

Plaats van de cesuur en theoretische percentages terechte beslissingen

De gebruiker van de toets dient zich te realiseren dat bij iedere gerapporteerde toetsscore een meetfout wordt gemaakt. Dit geldt voor alle toetsen, dus ook voor de TGN. Hoewel zeker is dàt er een meetfout wordt gemaakt, kennen wij voor individuele kandidaten niet de omvang van de meetfout en ook niet de richting waarin de fout optreedt. De ware score van de kandidaat kan hoger of lager liggen dan de gerapporteerde score. Wij kunnen op grond van de verzamelde data wel een schatting maken van de grootte van de meetfout voor de gemiddelde kandidaat op een bepaald niveau. Deze schattingen staan vermeld in Tabel 6.3.

De gebruiker van de toets kan een afweging maken welke gevolgen van de meetfout als meest kostbaar moeten worden beschouwd: is het ernstiger dat een kandidaat ten onrechte wordt afgewezen, of is juist onterecht slagen een minder aanvaardbaar risico? Wanneer men de in Tabel 5.5. gepresenteerde grensscores hanteert als beslispunten, is het risico gemiddeld – voor de gehele populatie van kandidaten – op onterecht zakken even groot als het risico op onterecht slagen.

De opdrachtgever heeft dus ten aanzien van de keuze voor een cesuur drie mogelijkheden met drie verschillende consequenties.

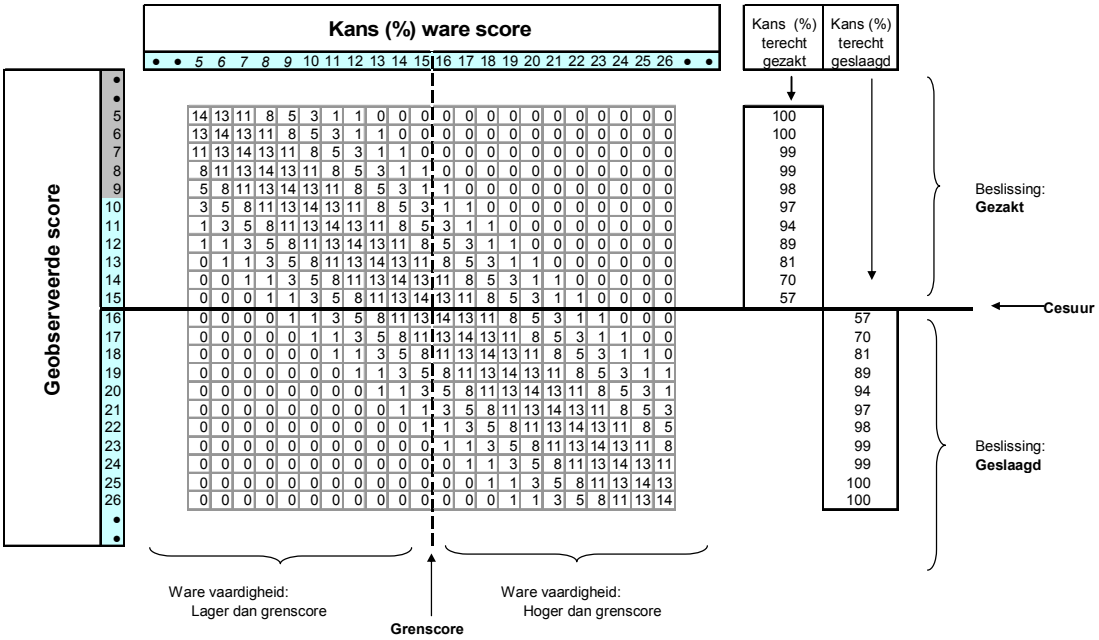
Plaats van de cesuur	Consequenties
Precies bij de grensscore	Het aantal foute beslissingen ten voordele van kandidaten is even groot als het aantal foute beslissingen ten nadele van kandidaten.
	Het risico dat kandidaten met een voldoende vaardigheid worden afgewezen is even groot als het risico dat kandidaten met een onvoldoende vaardigheid worden toegelaten.
Bij een score onder de grensscore	Het aantal foute beslissingen ten voordele van kandidaten is groter dan het aantal foute beslissingen ten nadele van kandidaten.
	Het risico dat kandidaten met een voldoende vaardigheid worden afgewezen is kleiner dan het risico dat kandidaten met een onvoldoende vaardigheid worden toegelaten.
Bij een score boven de grensscore	Het aantal foute beslissingen ten voordele van kandidaten is kleiner dan het aantal foute beslissingen ten nadele van kandidaten.
	Het risico dat kandidaten met een voldoende vaardigheid worden afgewezen is groter dan het risico dat kandidaten met een onvoldoende vaardigheid worden toegelaten.

Naarmate de gerapporteerde score van een kandidaat verder afligt van de grensscore voor een bepaald niveau wordt de kans kleiner dat op grond van deze score een onterechte beslissing over een kandidaat wordt genomen.

Figuur 1 brengt - gegeven de meetfout - de relatie tussen geobserveerde score en ware score in beeld alsmede het daaruit volgende risico van onterecht zakken of slagen bij een bepaalde score indien voor het A1-min niveau de cesuur precies bij de geschatte grensscore van 16 wordt gelegd. Verticaal zijn geobserveerde scores uitgezet en horizontaal de ware scores. Scores lager dan 10 worden gerapporteerd als 10. Deze zijn daarom donkerder gearceerd. In de cellen is de kans (afgerond op hele procenten) aangegeven dat een bepaalde ware score hoort bij een bepaalde geobserveerde score.

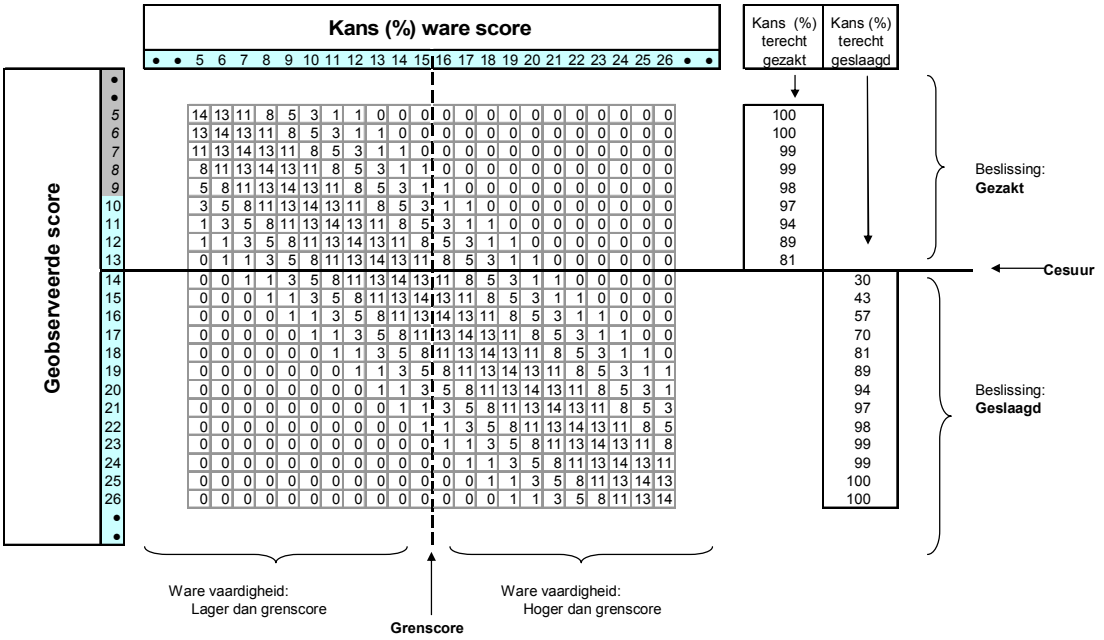
Nemen we bijvoorbeeld een geobserveerde score van 18, dan kan worden afgelezen dat er (theoretisch) geen enkele kans bestaat dat de ware score van deze kandidaat in feite 10 of lager is. Er bestaat 1 procent kans dat de ware score voor deze kandidaat 11 is, en ook 1 procent dat de score 12 is, 3 procent dat deze 13 is, enzovoorts. Op de ware scoreschaal is aangenomen dat een ware score van 16 en hoger duidt op een vaardigheid die hoger is dan de minimaal vereiste vaardigheid voor niveau A1-min. Alle ware scores lager dan 16 duiden op een vaardigheid die geringer is dan de minimaal vereiste vaardigheid. De grens tussen onvoldoende en voldoende vaardigheid op de ware scoreschaal is aangegeven met een verticale stippellijn. De grens tussen het toekennen van een 'voldoende' of een 'onvoldoende' is op de gerapporteerde scoreschaal aangegeven met een doorgetrokken horizontale lijn.

Het effect op de beslissing 'geslaagd' of 'gezakt' kan aan de rechterzijde van de figuur worden afgelezen. De kans dat een score van 16 overeenkomt met enige ware score boven de ondergrens van A1min is de som van alle bij een geobserveerde score van 16 behorende ware scores boven de ondergrens. De kans op een onterechte beslissing is (100% - de kans op een terechte beslissing). De kans dat de beslissing juist is om aan de geobserveerde score van 16 de beslissing 'geslaagd' te hechten is dus bij de gegeven meetfout 57%. De kans op een onterechte beslissing (de kandidaat is geslaagd, maar diens ware vaardigheid is in feite lager dan de ondergrens van A1min) is daarmee (100-57=) 43%. Bij een geobserveerde score van 15, net onder de cesuur, treffen we een omgekeerde verhouding aan tussen de kansen op terechte en onterechte beslissingen.



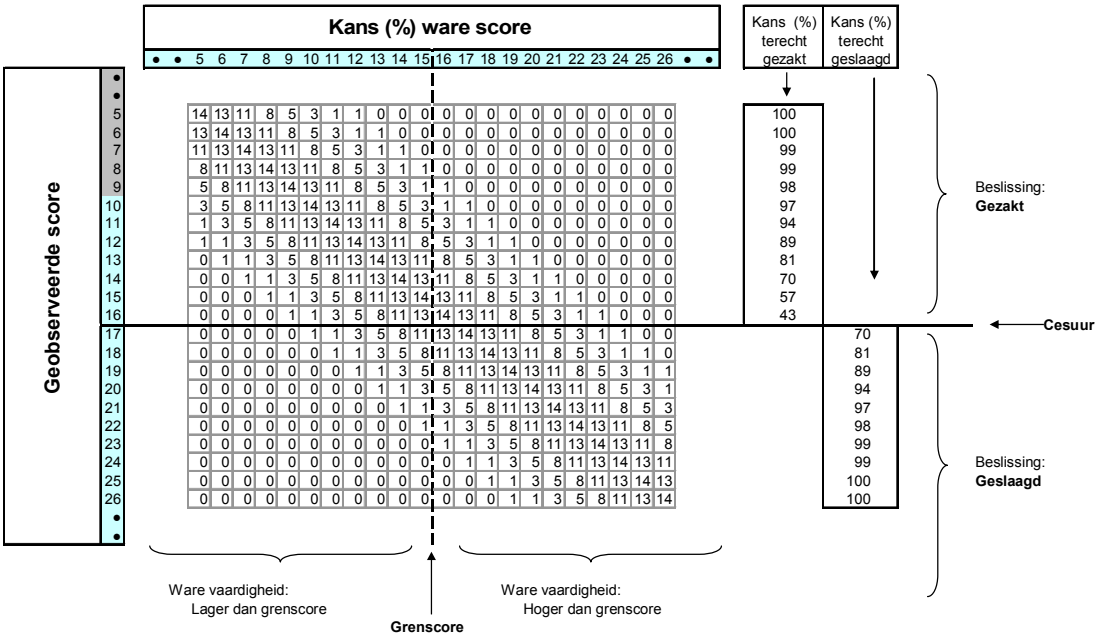
Figuur 1: Plaats van de cesuur precies bij de grensscore van A1-min (=16)

Figuur 2 toont de effecten van een keus voor de plaatsing van de cesuur twee scorepunten lager op de geobserveerde scoreschaal. Uit deze figuur blijkt dat de zekerheid van de beslissing 'gezakt' duidelijk toeneemt. Bij geen enkele geobserveerde score bestaat er minder dan 81% kans dat deze beslissing terecht is. Daar staat tegenover dat de kans dat een kandidaat onterecht slaagt fors toeneemt. Bij de laagste geobserveerde score waaraan het predikaat 'geslaagd' wordt toegekend (14) bestaat er een kans van (100-30=) 70% dat deze beslissing onjuist is.



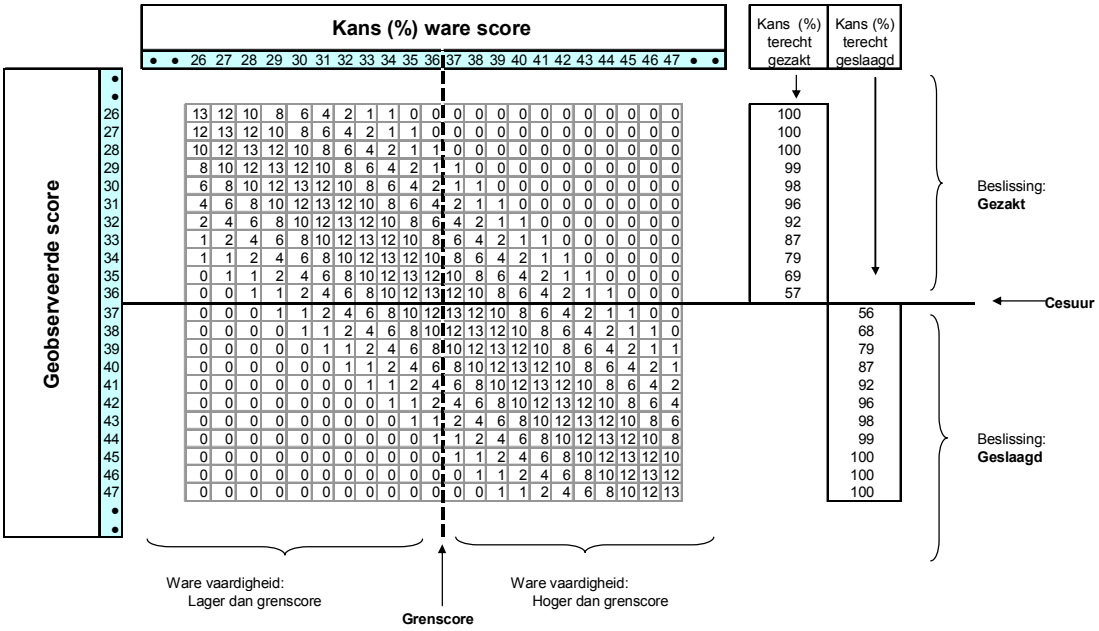
Figuur 2: Plaats van de cesuur bij een score onder de grensscore van A1-min (<16)

Figuur 3 toont de effecten van een keus voor de plaatsing van de cesuur een scorepunt hoger op de geobserveerde scoreschaal. Uit deze figuur blijkt dat de zekerheid van de beslissing 'geslaagd' duidelijk toeneemt. Daar staat echter tegenover dat nu juist de kans dat een kandidaat ten onrechte zakt veel groter is dan bij beide vorige keuzen.



Figuur 3: Plaats van de cesuur bij een score boven de grensscore van A1-min (>16)

Dezelfde berekeningen kunnen worden gemaakt voor de cesuur voor het A2 niveau. Figuur 4 toont de consequenties van de plaatsing van de cesuur precies bij de grensscore voor niveau A2 (=37). We zien ook hier dat de percentages terechte beslissingen bij de scores vlak onder en vlak boven de cesuur voor gezakt en voor geslaagd ongeveer van gelijke grootte zijn. Verplaatsing van de cesuur naar een lagere score of naar een hogere score heeft ook vergelijkbare gevolgen als bij de cesuur voor A1-min.



Figuur 4: Plaats van de cesuur precies bij de grensscore van A2 (=37)

Plaats van de cesuur en percentages gezakten

De keuze van de plaats van de cesuur heeft uiteraard ook gevolgen voor de percentages gezakten. Het is niet mogelijk op dit moment daar empirische gegevens over te leveren aangezien de populatie van deelnemers aan de examens nog niet bekend is. Gelet op de prijs die kandidaten voor het afleggen van een examen moeten betalen en rekening houdend met de beschikbaarheid van oefentoetsen waarmee kandidaten een reële schatting van hun slaagkansen op het examen kunnen maken, zou men kunnen veronderstellen dat vóór deelname aan het examen zelfselectie plaats zal vinden en dat daarom de percentages gezakten laag zullen zijn. Bij gebrek aan informatie over de mate waarin deze zelfselectie plaats zal vinden nemen we aan dat de verdeling van kandidaten in Amsterdam en in de groepen 1 t/m 4 van het MFA-Fit experiment representatief is voor de populatie deelnemers aan het examen in het buitenland en de verdeling van de pretest steekproef voor de populatie die aan de examens in Nederland zal deelnemen. De pretestdata zijn echter om twee redenen minder representatief: pretestkandidaten hebben geen oefentoetsen gemaakt en zij werden vanzelfsprekend geconfronteerd met de ongeselecteerde itemverzameling. Beide redenen hebben een negatieve invloed op hun scores. In werkelijkheid verwachten we daarom geringere percentages gezakten. Tabel 1 geeft een overzicht van de schattingen. De met de geschatte grensscore overeenkomende cesuur is vet gedrukt.

Tabel 1: Percentages gezakten bij verschillende cesuurkeuzen voor A1-min en voor A2

A1-min (Amsterdam & MFA-Fit data)			A2 (Pretest data)	
Minimum voldoende score	% gezakt		Minimum voldoende score	% gezakt
14	17.2%		35	23.0%
15	19.8%		36	24.8%
16	**20.2%**		**37**	**26.3%**
17	21.4%		38	28.1%
18	23.7%		39	29.4%

Plaats van de cesuur en schatting effectieve percentages terechte beslissingen

Wanneer we de schattingen van de vorige twee paragrafen combineren kunnen we een schatting maken van de effectieve percentages terechte (en onterechte) beslissingen bij de keuze voor verschillende cesuren. We weten immers uit Figuur 1 dat bij de keuze van een cesuur bij 16 er 57% van de kandidaten met die score van 16 onterecht slagen. Maar we schatten ook op basis van het MFA-Fit experiment, dat er 1.2% (21.4% - 20.2%) kandidaten zijn met die score van 16.

Dit betekent dus dat er dat er minder dan 1 kandidaat per honderd kandidaten (43% van 1.2%) is met een score van 16 die onterecht het predikaat 'geslaagd' krijgt toegekend. Op dezelfde wijze kunnen we schatten hoeveel kandidaten met een score van 17 onterecht slagen, enzovoorts met een score van 18, 19. Tellen we deze schattingen bij elkaar op dan krijgen we een schatting van het totale percentage kandidaten dat bij een de keuze van de cesuur bij 16 ten onrechte zou slagen: 1.8%. Deze berekeningen kunnen we voor een cesuur bij 16 ook maken voor 'terecht' of 'onterecht' zakken. En vervolgens kunnen we dezelfde berekeningen maken voor een aantal alternatieve cesuren. Tabel 2 geeft een overzicht van de schatting van de percentages terechte en onterechte beslissingen voor verschillende cesuren bij A1-min.

Tabel 2: Schatting terechtheid beslissingen bij verschillende cesuren voor A1-min (data Amsterdam en MFA-Fit)

Minimum	Gezakt		Geslaagd		Totaal	
voldoende score	Onterecht	Terecht	Onterecht	Terecht	Onterecht	Terecht
14	0.9%	16.3%	3.9%	78.9%	4.8%	95.2%
15	1.7%	18.2%	2.0%	78.1%	3.7%	96.3%
16	**1.8%**	**18.4%**	**1.8%**	**78.0%**	**3.6%**	**96.4%**
17	2.5%	18.9%	1.3%	77.3%	3.8%	96.2%
18	4.1%	19.6%	0.6%	75.7%	4.7%	95.3%

Dezelfde schattingen kunnen ook gemaakt worden voor de verschillende cesuren voor niveau A2 op grond van de geobserveerde verdeling van de NMS deelnemers aan de pretest. Tabel 3 geeft hiervan een overzicht.

Tabel 3: Schatting terechtheid beslissingen bij verschillende cesuren voor A2 (data pretest)

Minimum	Gezakt		Geslaagd		Totaal	
voldoende score	Onterecht	Terecht	Onterecht	Terecht	Onterecht	Terecht
35	1.1%	21.9%	3.0%	74.0%	4.1%	95.9%
36	1.9%	22.9%	2.0%	73.2%	3.8%	96.2%
37	**2.7%**	**23.6%**	**1.3%**	**72.4%**	**4.0%**	**96.0%**
38	4.0%	24.2%	0.7%	71.1%	4.7%	95.3%
39	5.0%	24.4%	0.5%	70.1%	5.4%	94.6%