Behavioural and Societal Sciences
Kampweg 5
3769 DE Soesterberg
Postbus 23
3769 ZG Soesterberg

www.tno.nl

T +31 88 866 15 00
F +31 34 635 39 77
infodesk@tno.nl

TNO-rapport

TNO 2013 R11811

Onderzoek naar de verschillen tussen menselijke en machinale scoring van de Toets Gesproken Nederlands

Datum	November 2013
Auteur(s)	Dr. Ir. J. M. Kessens Dr. O.A.J. Aarts
Aantal pagina's	43 (incl. bijlagen)
Aantal bijlagen	4
Opdrachtgever	Ministerie Sociale Zaken en Werkgelegenheid
Projectnummer	060.05128

Samenvatting

De Toets Gesproken Nederlands (TGN) is ontwikkeld om de mondelinge taalvaardigheid te toetsen van het Nederlands. De TGN toetst of een kandidaat voldoet aan een minimum mondeling taalvaardigheidsniveau door de antwoorden van een examenkandidaat automatisch te scoren met een spraakherkenner.

In 2007 is door TNO een onderzoek verricht naar de verschillen tussen machinale en menselijke scores van de TGN [5]. In 2007 werd een correlatiecoëfficiënt gevonden tussen de menselijke en machinale totaalscores van 0.90. Op grond van de aanbevelingen uit het onderzoeksrapport heeft het ministerie van Sociale Zaken en Werkgelegenheid (SZW) aan TNO gevraagd om het oorspronkelijke onderzoek te herhalen.

In dit onderzoek zijn twee representatieve steekproeven getrokken uit de TGN: een dataset met examens die afgenomen zijn in het binnenland en een dataset met examens uit het buitenland. Uit de resultaten van het onderzoek kunnen de volgende conclusies getrokken worden:

Correlatie tussen menselijke en machinale scores

- Op deelscore niveau zijn er aanzienlijke verschillen in de mate van correlatie tussen menselijke en machinale scores.
- Voor de totaalscores zijn de correlatiecoëfficiënten hoger, namelijk 0.80[1] voor de binnenlanddata en 0.83[2] voor de buitenlanddata. Deze waarden zijn lager dan de correlatiecoëfficiënt van 0.90 die werd gevonden in het onderzoek uit 2007.

Verschillen tussen menselijke en machinale scores

- De standaarddeviaties en het bereik van de menselijke scores zijn voor beide datasets aanzienlijk kleiner dan die van de machine scores.
- Er zijn verschillen in de gemiddelde deel- en totaalscores tussen mens en machine; de resultaten variëren echter per deelscores en per dataset. Voor de buitenlanddata geven de mensen gemiddeld significant lagere totaalscores dan de machine. Voor de binnenlanddata is het verschil in gemiddelde totaalscores niet significant.

Zak-/slaagpercentages

- Er zijn verschillen in de zak-/slaagpercentages volgens mens en machine; voor de binnenlanddata slagen minder kandidaten volgens de computer dan volgens de mens (51% volgens de computer versus 74% volgens de mens), terwijl voor de buitenlanddata juist meer kandidaten volgens de computer dan volgens de mens slagen (88% volgens de computer versus 79% volgens de mens).

[1] De boven- en ondergrens van het betrouwbaarheidsinterval zijn respectievelijk 0.73 en 0.85.
[2] De boven- en ondergrens van het betrouwbaarheidsinterval zijn respectievelijk 0.77 en 0.88.

Beoordelaarsovereenstemming

- Voor beide datasets geldt dat beoordelaars zeer consistent zijn in hun oordelen, d.w.z. er is een grote samenhang tussen de eerste en tweede score die ze geven over exact hetzelfde item en dezelfde kandidaat.
- De gemiddelde mate van overeenstemming tussen de oordelen van beoordelaars onderling zijn ‘redelijk’ (zinsbouw) en ‘bijna perfect’ (woordenschat) voor de inhoudelijke scores en ‘matig’ voor de kwalitatieve scores (uitspraak, vloeiendheid).

SZW heeft het consortium dat de toets heeft ontwikkeld (CINOP, Pearson en LTS) gevraagd om een reflectie te geven op de gevonden resultaten. De mogelijke verklaringen en aanbevelingen volgens CINOP, Pearson en LTS zijn:

Verklaring voor het verschil in de gevonden correlaties met het onderzoek uit 2007
De gevonden waarden van de correlaties in het huidige onderzoek zijn lager dan de correlatie die werd gevonden in het onderzoek uit 2007. Om praktische redenen waren er verschillen met de eerdere onderzoeksopzet:

- De omvang van de huidige steekproef is in dit onderzoek kleiner.
- De beoordelingssystematiek voor de inhoudelijke beoordeling is anders.
- Er is gebruik gemaakt van een bestaande groep beoordelaars, die ervaren zijn in het geven van dit type beoordelingen.

Om de eerste verklaring te toetsen kan een analyse uitgevoerd worden waarbij de binnen- en buitenlanddata als geheel worden geanalyseerd. Daarnaast kan het onderzoek herhaald worden op dezelfde wijze en onder dezelfde condities als in het oorspronkelijke onderzoek.

Verklaring voor de gevonden verschillen in totaalscores tussen mens en machine
Uit de resultaten blijkt dat de standaarddeviaties en het bereik van de menselijke scores voor beide datasets aanzienlijk kleiner zijn dan die van de machinescores. Dit kan als volgt verklaard worden:

- De neiging van beoordelaars tot “centrale tendentie”: Dit is het vermijden van de extremen van de schaal door beoordelaars.
- Een ander beoordelaarseffect is “normverschuiving”: Het aanpassen van de strengheid van de beoordelaar aan het gemiddelde niveau van de kandidaten.
- De betrouwbaarheid van de menselijke toetsscores kan laag zijn.

Centrale tendentie kan wellicht tegengegaan worden door beoordelaars vaker de extremen van de schaal te laten beoordelen. Een andere aanbeveling is om beoordelaars door middel van intervisie te wijzen op verschuiving van de norm. Tenslotte bevelen CINOP, Pearson en LTS aan om in een vervolgonderzoek de betrouwbaarheid van de oordelen van de computer en menselijke beoordelaars te betrekken.

Verklaring voor de verschillen in resultaat tussen de twee datasets
Voor de buitenlanddata geven de mensen significant lagere totaalscores dan de machine. Voor de binnenlanddata is het verschil in gemiddelde totaalscores niet significant. Deze verschillen in resultaat tussen de binnenlanddata en buitenlanddata kunnen gezocht worden in verschillen tussen de populaties:

- De achtergronden van de kandidaten: de populatie in het binnenland bevat waarschijnlijk gemiddeld lager opgeleide kandidaten dan de populatie in het buitenland.

Om deze verklaring te kunnen staven is het aan te bevelen om in vervolgonderzoek ook te beschikken over de achtergrondgegevens van kandidaten zoals de vooropleiding.

- Het taalniveau dat geleerd moet worden: kandidaten in het binnenland moeten de taal op een hoger niveau beheersen en hebben meer mogelijkheden om zich de taal eigen te maken ten opzichte van de kandidaten uit het buitenland. Hierdoor zullen de kandidaten in Nederland andere strategieën gebruiken om voor de toets te slagen dan de kandidaten uit het buitenland. De gegeven verklaring kan beter onderbouwd worden als een aanvullend onderzoek wordt verricht naar hoe kandidaten zich voorbereiden op de TGN.
- Het aantal keren dat een examen is herkanst: de kandidaten uit het buitenland herkansen de TGN minder vaak dan de kandidaten in Nederland. Dit betekent dat kandidaten in Nederland zich mogelijk anders gedragen dan kandidaten die het examen voor een eerste keer doen. Het verdient de aanbeveling voor de vergelijkbaarheid om een dataset te gebruiken die alleen uit kandidaten bestaat die voor de eerste keer het examen doen.

Summary

The “Toets Gesproken Nederlands” (TGN) is developed to test the oral language proficiency level in Dutch. The TGN tests whether a candidate possess a certain minimum proficiency level by scoring the answer to short questions with an automatic speech recognition system.

In 2007, TNO performed a study to tests whether the differences exist between the automatic and human scores of the TGN [5]. The correlation coefficient that was found between the human and machine total scores was 0.90. Based on the recommendations in this report, the Dutch Ministry of SZW asked TNO to repeat the earlier research.

In this research, two representative samples are drawn from the TGN: one dataset consisting of exams that have been conducted outside the Netherlands (‘buitenland’-data), and one from inside the Netherlands (‘binnenland’-data).
The following conclusions can be drawn based on the finding of this research:

Correlation between human and machine scores

- On sub score level, substantial differences exist between the degree of correlation between the human and machine scores.
- For the total scores, the correlation coefficients are in general higher, namely 0.80[3] for the ‘binnenland’-data en 0.83[4] for the ‘buitenland’-data. These values are lower than the correlation coefficient of 0.90 found in the 2007 research.

Differences between human and machine scores

- For both datasets, the standard deviations and the ranges of the human scores are considerably smaller than the standard deviations and ranges of the machine scores.
- Differences exist between the mean sub scores and total scores of the human judgers and the machine; however, the results vary per sub score and per dataset. For the ‘buitenland’-data the human judgers assign significant lower total scores than the machine. For the ‘binnenland’-data the difference in mean total scores is not significant.

Percentages passes and fails

- Differences exist between the percentages of passes and fails according to the machine and according to the human judgers; for the ‘binnenland’-data less candidates passes the test according to the computer than according to the human judgers (51% according to the computer versus 74% according to the humans), whereas for the ‘buitenland’-data more candidates passes the test according to the computer than according to the human judgers (88% according to the computer versus 79% according to the humans).

[3] The upper and lower bound of the 95% confidence interval are respectively 0.73 and 0.85.
[4] The upper and lower bound of the 95% confidence interval are respectively 0.77 and 0.88.

Rater agreement

- For both datasets, raters are very consistent in their judgments, i.e. the agreement is very high between the first and second score which they assign to exactly the same item of the same candidate.
- The mean agreement between two raters is qualified as ‘moderate’ (sentence build) and ‘almost perfect’ (vocabulary) for the content scores and ‘fair’ for the qualitative scores (pronunciation, fluency).

SZW has asked the consortium that developed the TGN (CINOP, Pearson en LTS) to reflect on the results that has been found. The possible explanations and recommendations for further research of CINOP, Pearson en LTS are:

Explanation for the difference in the correlations with the research of 2007
The values of the correlation coefficients in the current research are lower that the correlation coefficient that was found in 2007. For practical reasons, differences exist with the former research design:

- The size of the current sample is smaller in this research.
- The methodology of content scoring by the humans is different.
- The current pool of raters used in this research is different. These raters have a lot of experience with assigning this type of scores.

To test the first hypothesis, the complete dataset of ‘binnenland’- en ‘buitenland’-land data can be analyzed as one complete set. Besides, the research can be repeated using exactly the same methodology and under exactly the same conditions as the former research.

Explanation for the difference in human and machine scoring
The results show that, for both datasets, the standard deviations and ranges of the human scores are considerably smaller than of the machine scores. This can be explained as follows:

- “Central tendency”; the tendency of human judgers to avoid assigning extreme scores on a rating scale.
- “Standard shift”; the tendency of human judgers to adjust the strictness of their ratings on the mean proficiency level of the candidates.
- The reliability of the human scores can be low.

Central tendency may be avoided or diminished by submitting more exams that contains extremes to the judgers. Another recommendation is to point out the standard shift by means of peer review. Finally, another recommendation CINOP, Pearson en LTS recommend is to include the reliability of the human and machine scores in a follow-up study.

Explanation for the different results between the two datasets
For the ‘buitenland’-data the humans assign significantly lower total scores than the machine. For the ‘binnenland’-data the difference is not significant. The difference in results between the ‘binnenland’- and ‘buitenland’-data can be explained by differences in the populations:

- The background data of the candidates: de ‘binnenland’ population contains probably lower educated candidates compared to the ‘buitenland’ population. To verify this hypothesis, it is recommended to include the background data (like education) of the candidates in a follow-up research.

- The language proficiency level that has to be learned varies: candidates in the Netherlands need to have a higher proficiency level compared to the foreign candidates. Also, they have had more opportunities to learn the language compared to the foreign candidates. For this reason, they may use other strategies to pass the test, compared to the foreign candidates. Additional research on how candidates prepare for the TGN examination is required to explore this further.
- The number of times that the TGN is re-done: The candidates from abroad fail for the TGN exam less often than the candidates in the Netherlands.
  This means that the candidates in the Netherlands will behave differently compared to the foreign candidates. It is recommended to use a dataset that only includes data of candidates who make the TGN exam for the very first time.

Inhoudsopgave

1 Inleiding

1.1 Achtergrond onderzoek

De Toets Gesproken Nederlands (TGN) is ontwikkeld om de mondelinge taalvaardigheid te toetsen in het Nederlands [7]. De TGN toetst of een examenkandidaat voldoet aan een minimum mondeling taalvaardigheidsniveau.

De TGN wordt zowel in het buitenland als in Nederland afgenomen.
De examenkandidaten in het buitenland bestaan uit buitenlanders die zich duurzaam in Nederland willen vestigen. Men legt het examen af bij de Nederlandse ambassade of consulaat-generaal in het eigen land. De TGN vormt ook onderdeel van het inburgeringsexamen in Nederland en wordt afgenomen bij de afnamelocaties van de Dienst Uitvoering Onderwijs(DUO). Het vereiste mondelinge taalvaardigheidsniveau voor de TGN is A1 voor het examen afgenomen in het buitenland, en A2 voor het examen afgenomen in Nederland. Deze normering is gerelateerd aan de niveaus van het Gemeenschappelijk Europees Referentiekader (Common European Framework (CEF), 2001) [8].

De TGN is ontwikkeld door een consortium bestaande uit CINOP [1], LTS [2] en Pearson [3] en is gebaseerd op de Phonepass-technologie van Pearson.
Deze technologie toetst taalvaardigheid door de antwoorden van examenkandidaten automatisch te scoren met een spraakherkenner. Als een examenkandidaat de TGN vier keer zonder succes heeft afgelegd is het mogelijk om een herbeoordeling aan te vragen. Het examen wordt dan beoordeeld door menselijke beoordelaars. Op basis van de menselijke oordelen stelt het systeem vervolgens de toetsuitslag opnieuw vast. Dit proces van menselijke herbeoordeling wordt uitgevoerd door CINOP[5].

In 2007 is door TNO een onderzoek verricht naar de verschillen tussen machinale scores gegenereerd met het scoringssysteem van de TGN en scores die zijn gegeven door menselijke beoordelaars [5]. De steekproef in 2007 bestond uit examens afgenomen in het buitenland. Er werd toen een correlatiecoëfficiënt gevonden tussen de menselijke en machinale totaalscores van 0.90. Aangezien er in deze steekproef echter relatief weinig kandidaten waren met een score onder de zak-/slaaggrens, heeft TNO geadviseerd om de verschillen tussen menselijke en machinale scoring nogmaals te onderzoeken indien er een substantiëler aantal kandidaten zou zijn met een score onder de zak-/slaaggrens. Op grond van deze aanbevelingen heeft het ministerie van SZW aan TNO gevraagd om het oorspronkelijke onderzoek te herhalen. In dit onderzoek is een representatieve steekproef genomen uit TGN examens die in het binnenland en in het buitenland zijn afgenomen.

[5] Voor de Wet Inburgering vindt dit plaats vanaf 2 april 2013, voor de Wet Inburgering Buitenland vanaf 1 juli 2013.

1.2 Werking en toetsopbouw TGN

De TGN meet het gemak waarmee kandidaten gesproken Nederlands kunnen verstaan en begrijpen en hierop adequaat en verstaanbaar in het Nederlands kunnen reageren. De TGN bestaat uit 48 opgaven of items.
Ieder examen is uniek in de zin dat per examen de opgaven (per item type) willekeurig worden getrokken uit de item bank met mogelijke items. Er zijn drie itemtypes:

1 *Korte vragen* (14 items); een korte vraag met een kort antwoord, bijvoorbeeld "Kun je rijst eten of drinken?" (antwoord: eten).
2 *Tegenstellingen* (10 items); de kandidaat moet het tegengestelde van een woord zeggen, bijv. "Niet" (antwoord: wel).
3 *Herhaalopdrachten* (24 items). de kandidaat krijgt de opdracht een zin letterlijk te herhalen, bijv. "Daar heb ik nog nooit van gehoord".

Op basis van automatische analyse van de antwoorden van kandidaten op de toets opgaven worden vier deelscores gegenereerd:

1 Woordenschat.
2 Zinsbouw.
3 Uitspraak.
4 Vloeiendheid.

De deelscore "woordenschat" is gebaseerd op analyse van de antwoorden van de kandidaten op de korte vragen en tegenstellingen. De overige drie deelscores zijn gebaseerd op automatische analyse van de antwoorden van de kandidaten op de herhaalopdrachten.

De spraakherkenner genereert de deelscores op twee manieren: Voor de inhoudelijke deelscores *zinsbouw* en *woordenschat* bepaalt de spraakherkenner de inhoudelijke correctheid van de spraak ("Wat is er gezegd?"). Voor de kwalitatieve deelscores *uitspraak* en *vloeiendheid* scoort de spraakherkenner de spraakkwaliteit ("Hoe wordt het gezegd?"). De totaalscore op de toets wordt verkregen door de vier deelscores te schalen, in te perken en te middelen tot een totaalscore. Hierbij weegt iedere deelscore even zwaar mee. Voor een gedetailleerde omschrijving van de opbouw, werking en ontwikkeling van de TGN, zie "Verantwoording Toets Gesproken Nederlands, CINOP, 2005" [7].

2 Aanpak onderzoek

2.1 Verantwoording onderzoek

TNO was verantwoordelijk voor de opzet, uitvoer en rapportage van het onderzoek. Daarbij is er samengewerkt met de partijen die de TGN hebben ontwikkeld:

- CINOP was verantwoordelijk voor het verzamelen en vastleggen van de menselijke scores.
- Pearson was verantwoordelijk voor de selectie van de steekproeven, voor het opleveren van de machinale scores, en het omzetten van de menselijke deelscores naar totaalscores.

In totaal zijn er zes voortgangsgesprekken geweest tussen TNO, CINOP en SZW. Doel van deze gesprekken was; het vastleggen van de steekproef, het vastleggen van het onderzoeksprotocol, het monitoren van de voortgang van de data-verzameling en het bespreken van de resultaten van het onderzoek.
Aangezien Pearson een Amerikaans bedrijf is, is het onderzoeksprotocol in het Engels opgesteld en is bij dit rapport een Engelse samenvatting toegevoegd.

2.2 Relatie met het onderzoek uit 2007

Uitgangspunt voor het onderzoeksprotocol was het herhalen van het onderzoek dat TNO in 2007 heeft uitgevoerd. Vanwege praktische randvoorwaarden wijkt dit onderzoek in de uitvoer af van de eerdere opzet. Belangrijkste verschillen zijn:

- In het huidige onderzoek wordt de omzetting van de menselijke deelscores naar deelscores door Pearson uitgevoerd. Hiervoor wordt de zogenaamde “MRT scoring logic” gebruikt. De reden hiervoor is dat deze MRT logic ook in de praktijk wordt toegepast bij het omzetten van de menselijke itemscores naar deelscores tijdens het herbeoordelingsproces. Hierdoor zullen we met dit onderzoek de vraag beantwoorden: “Zijn er substantiële verschillen tussen de machinale scores en menselijke scores zoals deze verkregen worden tijdens het herbeoordelingsproces?”
- En ander verschil is dat de inhoudelijke scores (woordenschat en zinsbouw) niet worden verkregen door woord-voor-woord transcripties te maken, maar door direct het aantal correcte woorden te laten scoren door de menselijke beoordelaars. De reden hiervoor is een praktische; binnen de beperkte doorlooptijd van het onderzoek was het niet mogelijk om woord-voor-woord transcripties te laten maken. Door dezelfde beoordelaars in te zetten die worden gebruikt voor het herbeoordelen van de TGN examens werd het praktisch haalbaar om de data te verzamelen binnen de door het ministerie van SZW gestelde korte doorlooptijd van het onderzoek. Daarnaast is het voordeel van deze opzet dat de menselijke scores in het onderzoek nu direct vergelijkbaar zijn met de menselijke scores die gebruikt worden bij de herbeoordeling van de examens.

2.3 Onderzoeksopzet

Voordat begonnen is met de uitvoer van het onderzoek heeft TNO een onderzoeksopzet voorgesteld. Deze is vervolgens besproken met de betrokken partijen (CINOP & Pearson) en het ministerie SZW, om te toetsen op praktische haalbaarheid en inhoudelijke correctheid. Op 16 september is het definitieve onderzoeksprotocol vastgesteld. Details van het onderzoeksprotocol zijn te vinden in bijlage A. Hieronder worden de hoofdlijnen beschreven.

2.4 Onderzoeksvragen

De hoofdonderzoeksvraag luidt als volgt:

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

Het onderzoek naar de verschillen tussen menselijke en machinale scores van de TGN bevat de volgende subvragen:

1 Zijn er verschillen in de gemiddelde machinale scores en de gemiddelde menselijke scores?
2 Zijn er deelscores waar het verschil tussen menselijke en machinale scores groter is dan bij andere deelscores?
3 Wat is de correlatie tussen de machinale en menselijke scores, voor de deelscores en de totale toetsscore?
4 Wat zijn de percentages (niet)geslaagd volgens de machine en de percentages (niet) geslaagd volgens mens?

2.5 Dataverzameling

Uitgangspunt bij de dataverzameling is dat deze een representatieve steekproef vormt van de praktijk. In totaal zijn twee datasets verzameld:

1 *Binnenlanddata*; Dit zijn de data die zijn verzameld in Nederland. Deze examens zijn afgenomen bij de afnamelocaties van de DUO.
2 *Buitenlanddata*; Dit zijn de data die zijn verzameld in het land van herkomst van de examenkandidaten. Deze examens zijn afgelegd bij de Nederlandse ambassade of consulaat-generaal in het eigen land.

De TGN wordt afgenomen via de telefoon. Technisch verschil in de afname in Nederland en in het buitenland is dat er een ander telefoonnet gebruik wordt. In Nederland is dit het zogenaamde "Public Switched Telephone Network (PSTN), terwijl in het Buitenland een speciaal en beveiligt telefoonnet van het Ministerie van Buitenlandse Zaken (MFA-net) wordt gebruikt. Bij het MFA-net treedt er pakketverlies op en is het signaal enigszins vertraagd. In het onderzoek "verantwoording TGN" [7] is beschreven hoe de toets is aangepast om te corrigeren voor de nadelige invloed van het MFA-net.

Het vereiste mondelinge taalvaardigheidsniveau voor de TGN is A1 (TGN totaalscore=26) voor het examen afgenomen in het buitenland, en A2 (TGN totaalscore=37) voor het examen afgenomen in Nederland. De twee datasets zijn afkomstig uit de periode 1 januari 2012 t/m 1 april 2013. Examens zijn zo gekozen dat deze representatief zijn voor de complete dataset.

Alleen geldige examens zijn geselecteerd, volgens dezelfde criteria als gebruikt bij het onderzoek in 2007 [5]; examens van ambassadepersoneel, examen waarvoor de machine geen oordeel kon geven en onvolledige examens zijn uitgesloten van het onderzoek.

2.6 Methode

In het onderzoek zijn voor de twee datasets zowel menselijke als machinale itemscores verzameld. In tabel 1 staat een overzicht van de menselijke scores. Per dataset werden minimaal 125 examens beoordeeld. Per examenkandidaat werden 91 oordelen gegeven over 45 items. Hiertoe werd iedere respons van een respons van een kandidaat telkens voorgelegd aan vier beoordelaars die willekeurig uit de pool van beoordelaars werden getrokken. Voor de inhoudelijke scores (woordenschat en zinsbouw) werd de mate van correctheid gescoord, voor de kwalitatieve scores (vloeiendheid en uitspraak) werd een kwaliteitsoordeel gegeven volgens de CEF-schaal [8]. De exacte omschrijving van deze beoordelingen staan beschreven in het TNO-rapport uit 2007 [5].

Tabel 1 Overzicht van de menselijke scores.

Deelscore	Type score per item	Menselijke score per item	Aantal items per examenkandidaat
Woordenschat	Dichotoom: [0,1]	Correctheid antwoord	22
Zinsbouw	Polytoom: [0 - max. aantal woorden]	Aantal correcte woorden	23
Vloeiendheid	Polytoom: [0-7]	CEF-oordeel	23
Uitspraak	Polytoom: [0-7]	CEF-oordeel	23

Vervolgens zijn de 91 menselijke scores met de MRT logic omgezet naar 4 deelscores en 1 totaalscore. Daarnaast zijn ook per examenkandidaat 4 machinale deelscores en 1 totaalscore verkregen. Om de subvragen te beantwoorden zijn de onderstaande analyses uitgevoerd met de menselijke en machinale scores (voor de deelscores en de totaalscores).

2.6.1 Beschrijvende statistiek van de data

Om een indruk te krijgen van de scoreverdeling is een aantal kengetallen van de scoreverdelingen onderzocht (minimum-maximum, gemiddelde, standaarddeviatie en de scheefheid van de scoreverdeling).

2.6.2 Verschillen in gemiddelde scores

De verschillen tussen de gemiddelde machinale en menselijke scores zijn berekend en statistisch getoetst (gepaarde t-test).

2.6.3 Correlatie tussen menselijke en machinale scores

De Pearson's correlatie coëfficiënt is gebruikt om de samenhang tussen de menselijke en machinale scores te berekenen. Deze coëfficiënt wordt als volgt berekend:

$$r_{x\,y} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{(n-1)s_x\ s_y}$$

Waarbij:
$\overline{x}$ = steekproefgemiddelde van menselijke scores
$\overline{y}$ = steekproefgemiddelde van machinale scores
S_x = steekproef standaarddeviatie van de menselijke scores
S_y = steekproef standaarddeviatie van de machinale scores

2.6.4 Grafische weergave menselijke en machinale scores

De correlatie tussen de machine en de menselijke score zal ook visueel worden weergegeven in een puntenwolk. Voor het gemak is ook de regressielijn in de figuur weergegeven.

2.6.5 De zak-/slaagpercentages volgens mens en machine

Een ander aspect van belang zijn zak-/slaagpercentages volgens de machine en volgens de menselijke beoordelaars. Voor de binnenlanddata zijn deze berekend voor het afkappunt 37 (A2), en voor de buitenlanddata voor het afkappunt 26 (A1). Daarnaast kan de "odds ratio's[6]" worden berekend. Dit kwantificeert hoeverre eigenschap A ("pass" door een machine) is geassocieerd met eigenschap B ("pass" door een menselijke beoordelaar) volgens onderstaande formule:

$$odds\ ratio = \frac{N_{gezakt\ machine,\ \ gezakt\ mens}\ N_{geslaagd\ machine,\ \ geslaagd\ mens}}{N_{gezakt\ mens,\ \ geslaagd\ machine}\ N_{geslaagd\ mens,\ \ gezakt\ machine}}$$

Waarbij:
N = aantal kandidaten in een bepaalde cel.

2.6.6 De intra-beoordelaarsovereenstemming

De intra-beoordelaarsovereenstemming is berekend op basis van dubbele oordelen over exact hetzelfde item van dezelfde kandidaat (Cronbach's Alpha). Deze maat geeft aan in hoeverre beoordelaars geneigd zijn om dezelfde score te geven aan exact hetzelfde antwoord op een vraag/item.

2.6.7 De inter-beoordelaarovereenstemming

De betrouwbaarheid tussen beoordelaars onderling is berekend op basis van alle menselijke deelscores die zijn gegeven. Hiertoe is de Cohen's Kappa referentie [9] berekend voor alle mogelijke combinaties van beoordelaars. De Cohen's Kappa is een maat voor de overeenstemming van de beoordeling van twee verschillende beoordelaars. De formule voor Cohen's Kappa (κ) ziet er als volgt uit:

$$\kappa = \frac{\Pr\ (a) - \Pr\ (e)}{1 - \Pr\ (e)}$$

Waarbij:
Pr (a) = de relatieve geobserveerde overeenstemming tussen beoordelaars.
Pr (e) = de overeenstemming op basis van kans (waarbij de geobserveerde data worden gebruikt om deze overeenstemming op basis van kans te berekenen).

[6] De odds ratio bij deze steekproef omvang zijn erg gevoelig zijn voor de celvullingen. Echter doordat de sample random getrokken is mag verondersteld worden dat de odds ratio's redelijk stabiel zijn.

Als de beoordelaars perfect overeenstemmen dan is $\kappa = 1$. Als de overeenstemming niet hoger is dan de geschatte overeenkomst op basis van kans dan is, $\kappa = 0$.
Tabel 2 geeft de kwalificaties voor de κ-waarden groter dan nul volgens Landis & Koch, 1977 [10]:

Tabel 2 Kwalificaties van κ-waarde volgens Landis & Koch (1977).

κ-waarde	kwalificatie
0.00-0.20	Zwak
0.21-0.40	Matig
0.41-0.60	Redelijk
0.61-0.80	Sterk
0.81-1.00	Bijna perfect

3 Resultaten binnenlanddata

3.1 Beschrijvende statistiek

De beschrijvende statistieken van binnenlanddata staan in tabel 3. In totaal zijn er 137 examens beoordeeld. Om een eerste indruk te krijgen van de data is het bereik (de minimale score en de maximale score), het gemiddelde, de standaarddeviatie (s.d.) en de scheefheid gepresenteerd. De standaarddeviatie geeft informatie over de variantie van de desbetreffende variabele. Als de data van die variabele bijvoorbeeld normaal verdeeld is, valt 95% van alle waarden binnen $\mu - 2\sigma$ en $\mu + 2\sigma$ (zie bijlage B). De scheefheid is een statistiek met informatie over de mate waarin de betreffende gegevens van de variabele een Gaussische verdeling volgt. Als deze statistiek kleiner is dan -1 of groter is dan +1 zou er sprake kunnen zijn van scheefheid (een asymmetrie of een niet-normale verdeling).

Er zijn verschillen tussen menselijke en machinale scores. Gemiddeld zijn de scores van de mensen hoger dan die van de machine (de gemiddelde totaalscores is 41.10 bij de mens en 40.28 bij de machine). Verder zijn de standaarddeviaties (s.d.) van de machinale scores aanzienlijk groter dan die van de menselijke beoordelaars. Dit houdt in dat de scores die de machine genereert aanzienlijk meer variatie vertonen dan de scores die de mensen geven. Dit geldt met name voor de kwalitatieve scores (*vloeiendheid* en *uitspraak*). De scheefheidstatistiek geeft geen aanleiding om te veronderstellen dat de gegevens niet normaal verdeeld zijn.

Tabel 3 Beschrijvende statistiek van de binnenlanddata (N=137).

Score	min - max	gemiddelde	s.d.	scheefheid
Woordenschat mens	11.6 – 80.7	44.62	14.58	-0.02
Woordenschat machine	2 – 90	45.00	17.75	0.46
Zinsbouw mens	1.6 – 83.7	51.99	13.64	-0.72
Zinsbouw machine	4 – 90	48.27	16.49	0.45
Vloeiendheid mens	12.1 – 42.9	30.32	4.17	-0.92
Vloeiendheid machine	0 – 90	37.56	24.05	0.19
Uitspraak mens	10.7 – 57.6	37.49	6.04	-0.68
Uitspraak machine	0 – 90	30.30	22.23	0.55
Totaal mens	**12.4 – 66.1**	**41.10**	**8.69**	**-0.30**
Totaal machine	**8 – 90**	**40.28**	**16.45**	**0.55**

3.2 Verschillen in gemiddelde scores

In tabel 4 staan de verschillen in gemiddelde scores weergegeven: De 1ste kolom geeft aan om welke score het gaat, de 2de kolom geeft het verschil in gemiddelde scores tussen machine en mens, de 3de en 4de kolom geven respectievelijk de onder- en bovengrens van het 95% betrouwbaarheidsinterval. De significantie van het verschil is gebaseerd op een gepaarde t-test. De t-waarde (toetsgrootheid) staat in de laatste kolom.

Tabel 4 laat zien dat er verschillen zijn in de gemiddelde scores van mens en machine. Een positief getal betekent dat de machine gemiddeld hoger scoort dan de mens.

Met een gepaarde t-test onderzochten we of er mogelijke significante verschillen bestaan tussen de mens en machine scores. De resultaten in tabel 4 laten zien dat er significante verschillen zijn tussen de gemiddelde deelscores *zinsbouw*, *uitspraak* en *vloeiendheid*. Voor *zinsbouw* en *uitspraak* geeft de machine gemiddeld lagere scores dan de mens (respectievelijk 3.73 en 7.19 lager), voor *vloeiendheid* scoort de machine gemiddeld juist hoger (7.24 hoger). Voor de totaalscore is het verschil kleiner en niet significant.

Tabel 4 Verschillen in gemiddelde scores gebaseerd op een gepaarde t-test (dF=136).

Binnenland	verschil (machine – mens)	ondergrens 95%-interval	bovengrens 95%-interval	t-waarde
Woordenschat	0.40	-0.95	1.73	0.57
Zinsbouw	-3.73***	-5.23	-2.23	-4.94
Vloeiendheid	7.24***	3.51	10.97	3.84
Uitspraak	-7.19***	-10.34	-4.04	-4.52
Totaal	**-0.82**	**- 2.65**	**1.01**	**-0,89**

*** = sign. p<.001

3.3 Correlatie tussen menselijke en machinale scores

Ondanks deze verschillen kan er nog steeds een hoge samenhang zijn tussen de machinale en de menselijke beoordelingen. Dit berekenen we aan de hand van de Pearson's correlatie coëfficiënt. Daarnaast zijn ook de 95% betrouwbaarheids-intervallen berekend (ondergrens en bovengrens 95%-interval). Een betrouwbaarheidsinterval is een maat voor de betrouwbaarheid van een schatter (hier de correlatiecoëfficiënten). Dat wil zeggen dat bij een groot aantal herhalingen van de studie de werkelijke correlatie in 95% van de gevallen binnen dit interval zou vallen.

Tabel 5 laat de correlatiecoëfficiënten zien tussen menselijke en machinale scores. De correlaties voor de kwalitatieve deelscores *vloeiendheid* en *uitspraak* zijn lager dan de correlaties voor de inhoudelijke deelscore *woordenschat* en *zinsbouw*. Voor de totaalscore is de correlatiecoëfficiënt 0.80.

Tabel 5 Correlatie coëfficiënt tussen de menselijke en machinale scores.

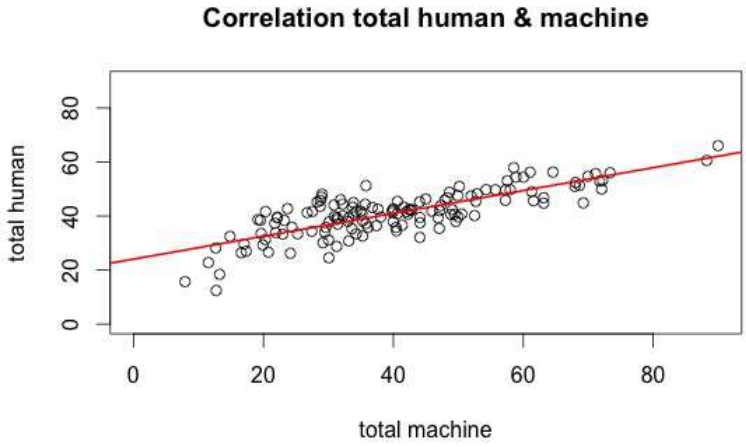
Binnenland	Pearson's correlatie coëfficiënt	ondergrens 95%-interval	bovengrens 95%-interval
Woordenschat	.90***	.86	.93
Zinsbouw	.85***	.79	.89
Vloeiendheid	.54***	.41	.65
Uitspraak	.68***	.58	.76
Totaal	**.80*** **	**.73**	**.85**

*** = sign. p<.001

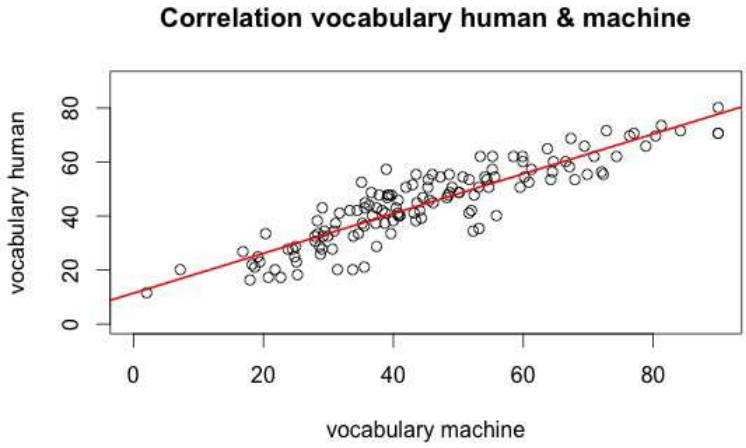
3.4 Grafische weergaven van de correlatie tussen mens en machine

De visuele weergave van de correlatie tussen de machine en de menselijke scores wordt getoond in de onderstaande figuren (1a t/m 1e). De puntenwolk voor de machinale en menselijke totaalscores wordt weergegeven, evenals de puntenwolken voor de deelscores. Voor het gemak is de regressielijn ook in de figuur weergegeven.

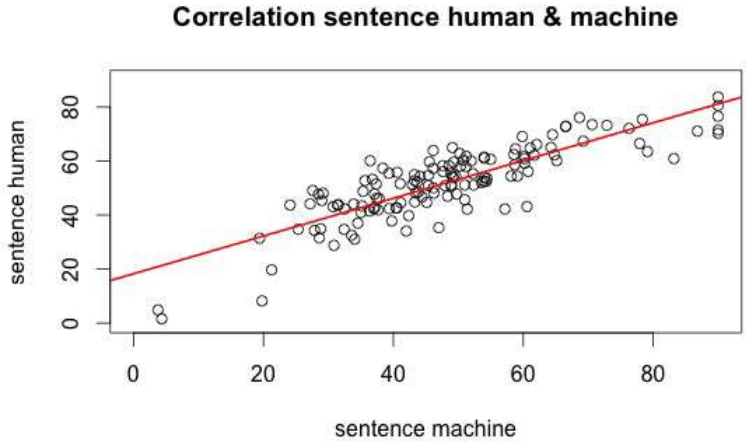
Wat opvalt in de figuren is dat de machine een breder score-bereik benut dan de menselijke beoordelaars. Dit geldt vooral voor de kwalitatieve scores (*uitspraak* en *vloeiendheid*). Verder zijn de extreme scores (0 en 90) vaker aanwezig bij de machine dan bij de mens.



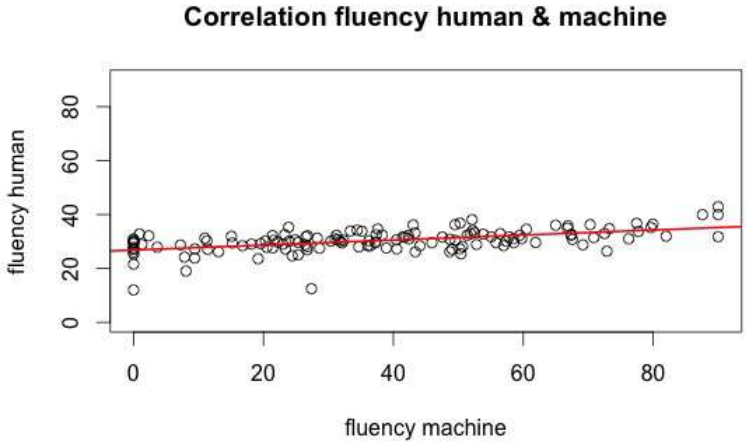
Figuur 1a Correlatie mens en machine (totaalscores); r=0.80.



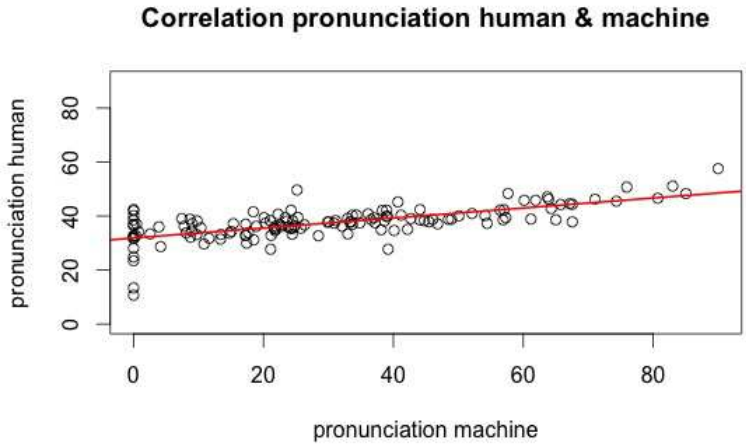
Figuur 1b Correlatie mens en machine (woordenschat); r=0.90.



Figuur 1c Correlatie mens en machine (zinsbouw); r=0.85.



Figuur 1d Correlatie mens en machine (vloeiendheid); r=0.54.



Figuur 1e Correlatie mens en machine (uitspraak); r=0.68.

3.5 Zak-/slaag-percentages volgens de mens en machine

Een ander aspect van belang zijn zak-/slaagpercentages volgens mens en machine. Voor de binnenlanddata worden deze percentages berekend voor een cut-off score van 37 (A2). De resultaten zijn weergegeven in tabel 6. Volgens de menselijke beoordelaars is 74% van de kandidaten geslaagd, tegen 51% volgens de machine.

Tabel 6 Zak-/slaagpercentages volgens mens en machine voor de binnenlanddata.

Binnenland	%geslaagd	%gezakt
Menselijke beoordelaars	74	26
Machine	51	49

Bij de verschillen is het informatief om na te gaan waar dit precies optreden. Daarom is er een kruistabel gemaakt om dit inzichtelijk te maken, zie tabel 7. De resultaten laten zien dat de verschillen vooral optreden in de cel ‘gezakt machine, geslaagd mens’, veel minder in de cel ‘geslaagd machine, gezakt mens’.

Tabel 7 Kruistabel met percentages gezakte en geslaagde kandidaten volgens mens en machine, de aantallen staan tussen haakjes.

Binnenland	Gezakt mens	Geslaagd mens	Totaal
Gezakt machine	22% (30)	27% (37)	49%
Geslaagd machine	4% (6)	47% (64)	51%
Totaal	26%	74%	100%
Odds ratio 9%			

3.6 Intra- en interbeoordelaarsovereenkomst

In totaal hebben 15 beoordelaars de menselijke scores voor de binnenlanddata gegeven. De intra-beoordelaarsovereenkomst werd onderzocht door het berekenen van de Cronbach's Alpha voor menselijke beoordelaars (N=15). Hiertoe is een deel van de data gebruikt waarbij een beoordelaar exact hetzelfde item van dezelfde kandidaat twee keer hebben beoordeeld.

Tabel 8 toont dat beoordelaars zeer consistent zijn in hun oordelen: Alle Cronbach's Alpha’s voor de deelscores zijn boven 0.90. Op basis daarvan kunnen we vaststellen dat de tweede score zeer vergelijkbaar is met de eerste score die gegeven werd.

Dit geldt voor zowel de totaalscores als de deelscores. Echter, kijken we naar de tijden tussen de eerste en tweede beoordeling dan vindt deze relatief snel plaats, waardoor mogelijk geheugeneffecten kunnen optreden[7].

Tabel 8 Intrarater-overeenkomst voor de verschillende deelscores.

Binnenland	Cronbach's alpha	N
Woordenschat	.988	656
Zinsbouw	.976	288
Vloeiendheid	.941	187
Uitspraak	.941	183
Totaal	**.986**	**1314**

Voor de verschillende deelscores is de Cohen's kappa berekend voor de matrix van alle mogelijke N x N beoordelaars (N=15). De Cohen's kappa wordt berekend op elke kandidaat/item combinaties die beide beoordelaars gescoord hebben.
Dit resulteert in 4 kruistabellen die te vinden zijn in bijlage C. De kappa's waarvoor het aantal items kleiner is dan 40 zijn niet weergegeven (n.a. = not applicable).
De gemiddelde Cohen's kappa en gemiddeld aantal items is weergegeven in tabel 9. Hierbij zijn alleen de kappa's meegenomen waarvoor het aantal items groter is dan 40.

Tabel 9 laat zien dat de mate van overeenkomst tussen beoordelaars sterk varieert per deelscore: Voor de deelscore *woordenschat* is de gemiddelde kappa "bijna perfect", voor *zinsbouw* "matig" en voor *vloeiendheid* en *uitspraak* "redelijk".

Tabel 9 Gemiddelde Cohen's κ, de kwalificatie volgens [10] en het gemiddeld aantal items.

Binnenland	Gem. Cohen's κ	kwalificatie	Gem. N
Woordenschat	0,889	bijna perfect	185
Zinsbouw	0,570	redelijk	217
Vloeiendheid	0,312	matig	188
Uitspraak	0,240	matig	188

[7] Met CINOP/Pearson was afgesproken dat er één dag tussen de eerste en tweede beoordeling zou zijn. Dit bleek niet het geval, de gemiddelde tijd (median) tussen de eerste en tweede beoordeling bedroeg 4 minuten. Uitgaande van gemiddeld 200 beoordelingen per uur (bron CINOP), zitten er dus ruim 13 beoordelingen tussen de eerste en tweede beoordeling.

4 Resultaten buitenlanddata

4.1 Beschrijvende statistiek

De beschrijvende statistieken van buitenlanddata staan in tabel 10. In totaal zijn er 137 examens beoordeeld. De gemiddelde scores verschillen van de binnenlanddata. Hier geven de menselijke beoordelaars gemiddeld juist lagere scores dan de machine (33.1 versus 41.5). Weer geldt dat de standaarddeviaties voor de machine aanzienlijk groter zijn dan die van de menselijke scores.
De scheefheidstatistiek geeft wederom geen aanleiding om te veronderstellen dat de gegevens niet normaal verdeeld zijn.

Tabel 10 Beschrijvende statistiek van de buitenlanddata (N=137).

Buitenland	min - max	gemiddelde	s.d.	scheefheid
Woordenschat mens	10.0 – 42.1	27.1	4.3	-0.89
Woordenschat machine	2 – 87	37.9	14.2	0.32
Zinsbouw mens	2.05 – 77.3	33.5	15.2	0.21
Zinsbouw machine	0 – 90	43.8	17.2	0.20
Vloeiendheid mens	10.0 – 59.1	32.5	6.9	-0.15
Vloeiendheid machine	0 – 85	44.5	20.1	-0.18
Uitspraak mens	10.0 – 65.6	33.2	9.6	0.23
Uitspraak machine	2 – 84	39.8	17.7	0.22
Totaal mens	**10.0 – 65.6**	**33.1**	**9.5**	**0.25**
Totaal machine	**4 – 83**	**41.5**	**14.1**	**0.20**

4.2 Verschillen in gemiddelde scores

Tabel 11 laat zien dat alle verschillen in de gemiddelde scores van mens en machine positief en significant zijn. Dit betekent dat voor alle deelscores de machine hogere scores genereert dan de menselijke beoordelaars. De verschillen zijn het grootst voor *woordenschat*, *zinsbouw* en *vloeiendheid*. Ook voor de totaalscore scoort de machine significant hoger dan de mens (8.37 punten hoger).

Tabel 11 Verschillen in gemiddelde scores gebaseerd op een gepaarde t-test (dF=136).

Buitenland	verschil (machine – mens)	ondergrens 95%-interval	bovengrens 95%-interval	t-waarde
Woordenschat	10.80***	8.71	12.89	10.22
Zinsbouw	10.24***	7.59	12.90	7.64
Vloeiendheid	12.00***	8.98	15.03	7.86
Uitspraak	6.60***	4.42	8.78	5.99
Totaal	**8.37*** **	**6.99**	**9.75**	**12.00**

*** = sign. p<.001

4.3 Correlaties tussen menselijke en machinale scores

Tabel 12 laat de correlatiecoëfficiënten zien tussen menselijke en machinale scores voor de buitenlanddata. De correlaties zijn relatief zwak voor de deelscores (vooral voor *woordenschat* (0.55), *zinsbouw* (0.54) en *vloeiendheid* (0.47)), maar is een stuk hoger voor de totaalscore (0.83).

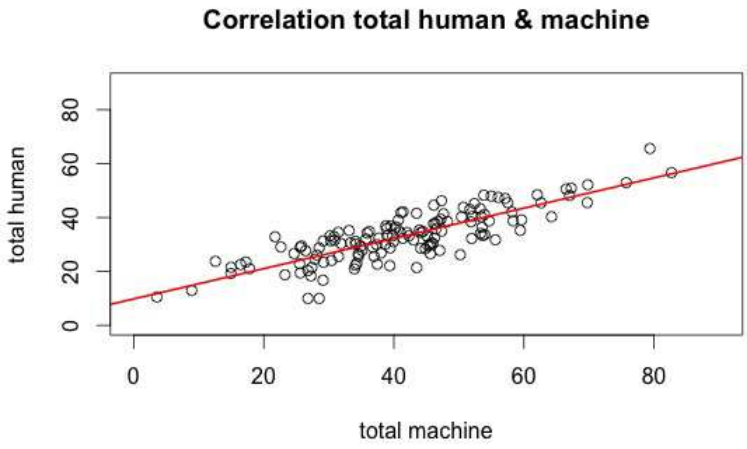
Tabel 12 Correlatie coëfficiënt tussen de menselijke en machinale scores

Buitenland	Pearson's correlatie coëfficiënt	ondergrens 95%-interval	bovengrens 95%-interval
Woordenschat	0.55***	0.43	0.66
Zinsbouw	0.54***	0.41	0.65
Vloeiendheid	0.47***	0.32	0.59
Uitspraak	0.70***	0.61	0.78
Totaal	**0.83*** **	**0.77**	**0.88**

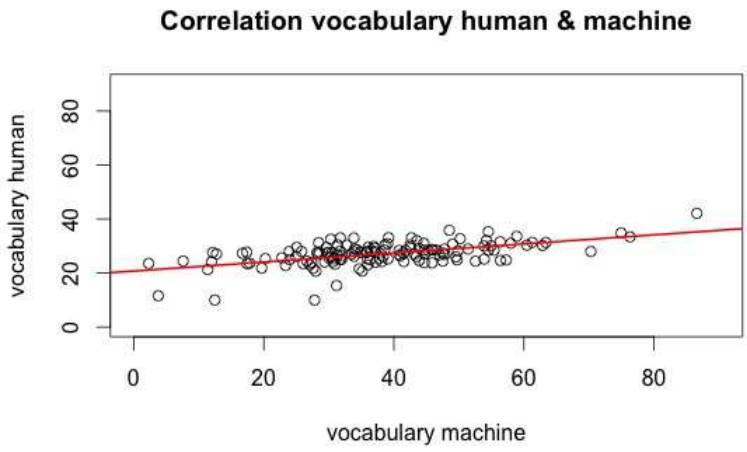
*** = sign. p<.001

4.4 Grafische weergaven van de correlatie tussen mens en machine

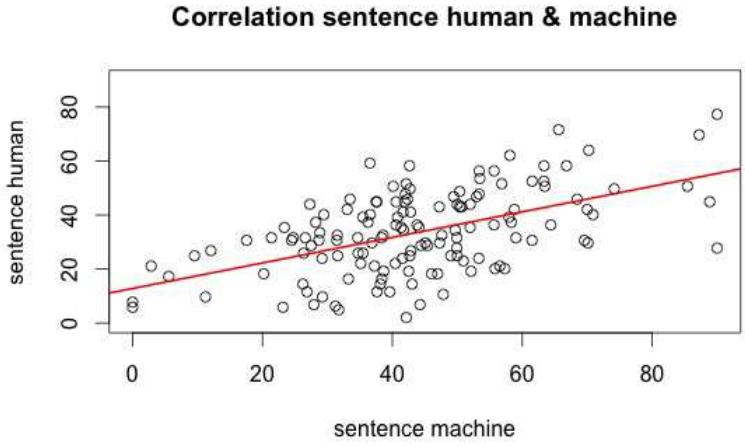
De visuele weergave van de correlatie tussen de machinale en de menselijke scores wordt getoond in onderstaande figuren (2a t/m 2e). Wat weer opvalt in de figuren is dat de machine een breder score-bereik benut dan de menselijke beoordelaars. Verder zijn de extreme scores (0 en 90) ook weer vaker aanwezig bij de machine dan bij de mens.



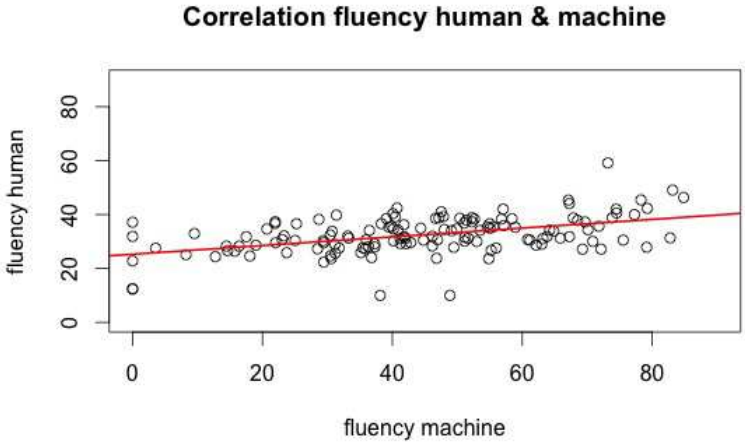
Figuur 2a Correlatie mens en machine (totaalscores); r=0.83.



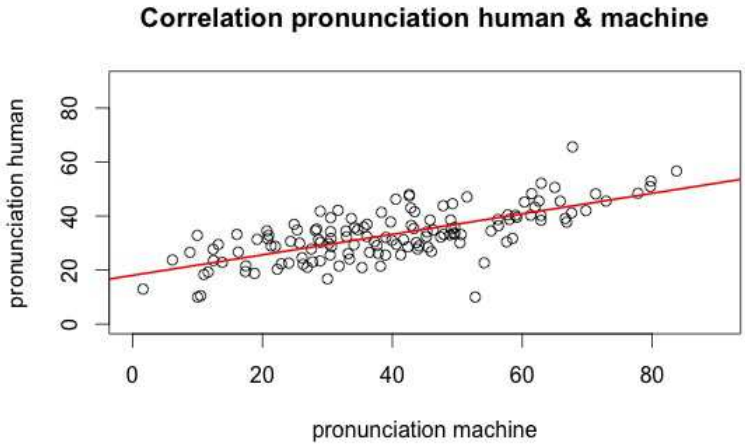
Figuur 2b Correlatie mens en machine (woordenschat); r=0.55.



Figuur 2c Correlatie mens en machine (zinsbouw); r=0.54.



Figuur 2d Correlatie mens en machine (vloeiendheid); r=0.47.



Figuur 2e Correlatie mens en machine (uitspraak); r=0.70.

4.5 Zak-/slaag-percentages volgens de mens en machine

Wederom zijn de zak-/slaagpercentages volgens mens en machine berekend. Voor de buitenlanddata worden deze percentages berekend voor een cut-off score van 26 (A1). De resultaten zijn weergegeven in tabel 13. Volgens de menselijke beoordelaars is 79% van de kandidaten geslaagd, tegen 88% volgens de machine.

Tabel 13 Zak-/slaagpercentages volgens mens en machine voor de binnenlanddata.

Buitenland	%geslaagd	%gezakt
Menselijke beoordelaars	79	21
Machine	88	12

Bij de verschillen is het informatief om na te gaan waar deze precies plaatsvinden. Daarom is er een kruistabel gemaakt, zie tabel 14. De resultaten laten zien dat de verschillen vooral optreden in de cel ‘geslaagd machine, gezakt mens’, en minder in de cel ‘geslaagd mens, gezakt machine’. Dit is tegengesteld aan de resultaten die zijn gevonden voor de binnenlanddata.

Tabel 14 Kruistabel met percentages gezakte en geslaagde kandidaten volgens mens en machine, de aantallen staan tussen haakjes.

Buitenland	Gezakt mens	Geslaagd mens	Totaal
Gezakt machine	8% (11)	4% (5)	12%
Geslaagd machine	13% (29)	75% (103)	88%
Totaal	21%	79%	100%
Odds ratio 8%			

4.6 Intra- en inter-beoordelaarsovereenstemming

De menselijke scores voor de buitenlanddata zijn door 15 beoordelaars gemaakt (N=15). Voor de buitenlanddata overlappen 14 beoordelaars met de binnenlanddata, beoordelaar 15 is uniek voor de binnenlanddata en beoordelaar 16 is uniek voor de buitenlanddata.

Tabel 15 toont dat beoordelaars wederom zeer consistent zijn in hun oordelen; de tweede score is zeer vergelijkbaar met de eerste score die gegeven werd. Dit komt tot uiting in de Cronbach's Alpha's, deze zijn boven de 0.90. Dit geldt voor zowel de totaalscores als de deelscores[8].

Tabel 15 Cronbach’s alpha voor de verschillende deelscores.

Buitenland	Cronbach’s alpha	N
Woordenschat	.974	463
Zinsbouw	.972	463
Vloeiendheid	.939	580
Uitspraak	.954	468
Totaal	.974	**2039**

[8] De gemiddelde tijd (median) tussen de eerste en tweede beoordeling bedroeg 5 minuten in de buitenland-data. Uitgaande van gemiddeld 200 beoordelingen per uur (bron CINOP), zitten er dus bijna 17 beoordelingen tussen de eerste en tweede beoordeling.

Voor de verschillende deelscores is de Cohen's kappa berekend voor de matrix van alle mogelijke N x N beoordelaars (N=15), zie bijlage D. De gemiddelde Cohen's kappa's staan in Tabel 16. Hierbij zijn weer alleen de kappa's meegenomen waarvoor N>40.

Tabel 16 laat zien dat de mate van overeenkomst tussen beoordelaars weer sterk varieert per deelscore. De kwalificaties per deelscore zijn gelijk aan de binnenlanddata: Voor de deelscore *woordenschat* is de gemiddelde kappa "bijna perfect", voor *zinsbouw* "redelijk" en voor *vloeiendheid & uitspraak* "matig".

Tabel 16 Gemiddelde Cohen's κ per deelscore, de kwalificatie volgens [10] en het gemiddeld aantal items.

Binnenland	Gem. Cohen's κ	kwalificatie	Gem. N
Woordenschat	0,887	bijna perfect	165
Zinsbouw	0,459	redelijk	190
Vloeiendheid	0,286	matig	188
Uitspraak	0,247	matig	188

5 Conclusies

Uit de resultaten kunnen de volgende conclusies getrokken worden:

Correlatie tussen menselijke en machinale scores

- Op deelscore niveau zijn er aanzienlijke verschillen in de mate van correlatie tussen menselijke en machinale scores.
- Voor de totaalscores zijn de correlatiecoëfficiënten hoger, namelijk 0.80[9] voor de binnenlanddata en 0.83[10] voor de buitenlanddata. Deze waarden zijn lager dan de correlatiecoëfficiënt van 0.90 die werd gevonden in het onderzoek uit 2007.

Verschillen tussen menselijke en machinale scores

- De standaarddeviaties en het bereik van de menselijke scores zijn voor beide datasets aanzienlijk kleiner dan die van de machine scores.
- Er zijn verschillen in de gemiddelde deel- en totaalscores tussen mens en machine; de resultaten variëren echter per deelscores en per dataset. Voor de buitenlanddata geven de mensen gemiddeld significant lagere totaalscores dan de machine. Voor de binnenlanddata is het verschil in gemiddelde totaalscores niet significant.

Zak-/slaagpercentages

- Er zijn verschillen in de zak-/slaagpercentages volgens mens en machine; voor de binnenlanddata slagen minder kandidaten volgens de computer dan volgens de mens (51% volgens de computer versus 74% volgens de mens), terwijl voor de buitenlanddata juist meer kandidaten volgens de computer dan volgens de mens slagen (88% volgens de computer versus 79% volgens de mens).

Beoordelaarsovereenstemming

- Voor beide datasets geldt dat beoordelaars zeer consistent zijn in hun oordelen, d.w.z. er is een grote samenhang tussen de eerste en tweede score die ze geven over exact hetzelfde item en dezelfde kandidaat.
- De gemiddelde mate van overeenstemming tussen de oordelen van beoordelaars onderling zijn ‘redelijk’ (zinsbouw) en ‘bijna perfect’ (woordenschat) voor de inhoudelijke scores en ‘matig’ voor de kwalitatieve scores (uitspraak, vloeiendheid).

[9] De boven- en ondergrens van het betrouwbaarheidsinterval zijn respectievelijk 0.73 en 0.85.
[10] De boven- en ondergrens van het betrouwbaarheidsinterval zijn respectievelijk 0.77 en 0.88.

6 Aanbevelingen

De resultaten in dit onderzoek hebben een nadere verklaring nodig. Op basis van de data die TNO ter beschikking heeft gekregen verdiend het in ieder geval de aanbeveling om nader te kijken naar de menselijke beoordeling. De standaarddeviaties en het bereik van de menselijke scores zijn voor beide datasets aanzienlijk kleiner dan die van de machine scores. Dit impliceert mogelijke leer- en kadereffecten bij de menselijke beoordelaars.

TNO heeft echter onvoldoende beschikking over achtergrondgegevens van de data om aanbevelingen te kunnen doen ten aanzien van de overige resultaten. Om deze reden is met SZW afgesproken dat het consortium van CINOP, Pearson en LTS als reactie op dit rapport mogelijke verklaringen en aanbevelingen opschrijft. Deze reactie is hieronder letterlijk weergegeven.

Mogelijke verklaringen door CINOP, Pearson en LTS

TNO heeft in opdracht van het Ministerie van Sociale Zaken en Werkgelegenheid (ministerie van SZW) een herhalingsonderzoek van het onderzoek uit 2007 verricht met als hoofdvraag:

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

Het ministerie van SZW heeft CINOP gevraagd een reflectie met mogelijke verklaringen te schrijven op de gevonden resultaten uit dit rapport. Hieronder staat een beschrijving van mogelijke verklaringen voor de door TNO gerapporteerde resultaten. Om deze verklaringen te kunnen staven wordt aanbevolen nader onderzoek te verrichten. Daar waar mogelijk wordt de verklaring afgesloten met een aanbeveling voor verder onderzoek.

De hoofdconclusie van het onderzoek vermeldt dat de correlatie tussen menselijke en machinale totaalscores 0.80 (binnenlanddata) is en 0.83 (buitenlanddata). Dit is een lagere correlatie dan gevonden is in het TNO-onderzoek van 2007 (toen was de correlatie 0.90). Mogelijk kan een deel van dit verschil verklaard worden door het verschil in het onderzoeksdesign dat gebruikt is in 2007 ten opzichte van het design van het huidige onderzoek. Dat onderzoeksdesign verschilt in een aantal aspecten, namelijk:

- De grootte van de dataset; in 2007 zijn 500 kandidaten onderzocht, terwijl in dit onderzoek 274 examenkandidaten zijn onderzocht, waarbij geen uitspraken over het totaal aantal kandidaten wordt gedaan, maar gedifferentieerd in twee representatieve groepen van elk 137 kandidaten.
- De beoordelaarssystematiek; in 2007 zijn woord-voor-woord transcripties gebruikt, terwijl in het huidige onderzoek beoordelaars het aantal correcte woorden hebben gescoord.
- De ervaring van de beoordelaars; in 2007 zijn beoordelaars specifiek voor het onderzoek getraind, terwijl er nu gebruik is gemaakt van een bestaande groep beoordelaars.

In een nader onderzoek en/of analyse zou de totale steekproef als één steekproef behandeld kunnen worden, zodat over een grotere steekproef uitspraken gedaan kunnen worden. Beoordelaars zijn immers ook niet op de hoogte geweest tot welke groep een te beoordelen item behoort. Om te onderzoeken of de gevonden resultaten verklaard kunnen worden door de verschillen in opzet van het onderzoek, is het aan te bevelen om de dataverzameling op dezelfde wijze en condities te laten plaatsvinden als in het oorspronkelijke onderzoek. Dit betekent zowel de items te laten transcriberen als ook een pool beoordelaars te trainen met dit doel.

Op basis van de resultaten wordt verder door TNO aanbevolen om als mogelijke oorzaak voor de verschillen tussen beoordelingen van de computer en menselijke beoordelaars nader te kijken naar de menselijke beoordelaars. De verschillen tussen mens en machine zijn op de totaalscore kleiner dan op de deelscores, waarbij het verschil bij de binnenlanddata niet significant is. Hierbij moet opgemerkt worden dat de toets geconstrueerd is om een betrouwbare uitspraak over de totaalscore te geven en dat de deelscores niet bedoeld zijn om daar uitspraken over te doen.

De verschillen tussen mens en computer zijn met name groot op de kwalitatieve aspecten *uitspraak* en *vloeiendheid*. Uit de resultaten blijkt dat de computer steeds de gehele scoreschaal gebruikt (minima en maxima respectievelijk dicht bij nul en bij 90). De beoordelaars gebruiken soms de helft of minder van de scoreschaal; zij maken dus veel minder onderscheid tussen de kandidaten. De reden dat beoordelaars geen gebruik maken van de hele schaal wordt mogelijk veroorzaakt door een bekende systematische afwijking die beoordelaars kunnen vertonen, namelijk ‘centrale tendentie’. Met andere woorden, beoordelaars vermijden extreem geformuleerde categorieën en zitten met hun beoordelingen steeds rond het midden van de beoordelingsschaal. Een ander beoordelaarseffect dat hier het verschil tussen oordelen van de computer en beoordelaars zou kunnen verklaren, is normverschuiving (de neiging van een beoordelaar om de strengheid van de beoordelingen aan te passen aan het gemiddelde niveau van de te beoordelen kandidaten).

De neiging tot centrale tendentie op te heffen blijft een uitdaging.
Door beoordelaars vaker examens te laten beoordelen die ook aan de bovenkant van de schaal zitten, kan geprobeerd worden om beoordelaars meer gebruik te laten maken van de hele schaal. Een aanbeveling om de zogenaamde normverschuiving tegen te gaan is om de groep beoordelaars regelmatig door middel van intervisie te wijzen op het verschuiven van de norm. Een andere aanbeveling is om in een vervolgonderzoek de betrouwbaarheid van de oordelen van de computer en menselijke beoordelaars erin te betrekken. De scores voor de betrouwbaarheid van de oordelen van de computer en menselijke beoordelaars ontbreken in dit rapport. Deze zijn van belang om de correlaties tussen mens en machine te kunnen duiden. Immers de hoogte van de correlatie tussen mens en machine wordt beperkt door betrouwbaarheid van beide scores. Dat beoordelaars consistent zijn op itemniveau betekent nog niet direct dat hun toetsscores betrouwbaar zijn. Gegeven de soms zeer geringe variantie in hun scores is het waarschijnlijk dat de betrouwbaarheid van hun toetsscores laag is.

Uit de resultaten van het onderzoek wordt daarnaast geconcludeerd dat er verschillen zijn in de gemiddelde totaalscores tussen mens en machine. Het rapport concludeert: "*Er zijn verschillen in de gemiddelde deel- en totaalscores tussen mens en machine; de resultaten variëren echter per deelscores en per dataset. Voor de buitenlanddata geven de mensen significant lagere totaalscores dan de machine. Voor de binnenlanddata is het verschil niet significant."* Mogelijke verklaring voor de verschillen in resultaten is het verschil tussen de twee populaties van de datasets.

Er kunnen grofweg drie verschillen benoemd worden: de achtergronden van de kandidaten, het taalniveau dat geleerd moet worden en de manier waarop dat niveau bereikt wordt en het aantal keren dat het examen herkanst wordt.

Een eerste verschil tussen de populaties wordt veroorzaakt door de achtergronden van de kandidaten: de buitenlanddata bestaan uit data die zijn verzameld in het buitenland. Het betreft hier kandidaten die de TGN in het land van herkomst hebben afgelegd op een ambassade of consulaat-generaal. Alle vreemdelingen[11] die zich vrijwillig voor langere tijd in Nederland willen vestigen in het kader van gezinsvorming, gezinshereniging of als geestelijk bedienaar, dienen dit examen af te leggen. Hierdoor bestaat de populatie uit kandidaten met alle mogelijke verschillen in achtergronden. De binnenlanddata bestaan daarentegen uit data die verzameld zijn in Nederland; het betreft hier kandidaten die de TGN op een afnamelocatie van DUO hebben gemaakt. Kandidaten die in Nederland inburgeren hebben een keuze hoe zij dit willen doen. Dit kan door het inburgeringsexamen te doen, maar bijvoorbeeld ook door het staatsexamen Nederlands als Tweede Taal of een beroepsopleiding te doen. Het staatsexamen NT2 toetst een hoger niveau van het Nederlands (B1 of B2). Hoger taalvaardige en/of hoger opgeleide kandidaten kunnen om die reden vaker voor het staatsexamen kiezen. Hierdoor bestaat de populatie, die de TGN in het binnenland aflegt uit gemiddeld lager opgeleide kandidaten dan de populatie die de TGN in het land van herkomst aflegt. Om de verklaring echt te kunnen staven is het aan te bevelen om in vervolgonderzoek ook te beschikken over achtergrondgegevens van kandidaten zoals de vooropleiding.

Een tweede verschil tussen de twee populaties is de context waarin kandidaten de taal kunnen leren en het taalniveau dat zij moeten aantonen om te slagen voor het examen. De kandidaten die deelnemen aan het examen in het buitenland, bereiden zich voornamelijk in het land van herkomst voor op het examen (en leren dus het Nederlands als vreemde taal). Zij moeten taalniveau A1 halen. Kandidaten die examen in Nederland doen, kunnen zich daarentegen hierop in Nederland voorbereiden (zij leren het Nederlands als tweede taal), waardoor zij veel mogelijkheden hebben om zich de taal eigen te maken; de taal wordt immers overal om hen heen gebruikt.

---

[11] Het basisexamen inburgering moet worden afgelegd door vreemdelingen (bron: www.ind.nl):

- in de leeftijd van 18 jaar tot de AOW-gerechtigde leeftijd;
- die vóór hun komst naar Nederland in het bezit moeten zijn van een mvv;
- die voor een niet-tijdelijk verblijfsdoel naar Nederland komen in de zin van de Wet inburgering[2] of geestelijk bedienaar zijn; en
- die niet zijn vrijgesteld of ontheven van het basisexamen inburgering (art. 3 en 5 Wet inburgering).

Zij moeten het taalniveau A2 hebben om te slagen voor het examen (wat één ERK-niveau hoger is dan het taalniveau A1). Dit leidt ertoe dat kandidaten die examen doen in het buitenland mogelijk andere strategieën gebruiken dan kandidaten die examen in Nederland doen (namelijk strategieën die passen bij het de hoogte van het taalniveau). Bovendien leiden bepaalde antwoordpatronen in het buitenland mogelijk al tot een voldoende op A1, terwijl die antwoordpatronen in het binnenland niet voldoende zijn om het niveau A2 te halen. Hierdoor zullen er andere strategieën gebruikt worden. Door het gebruik van 'verkeerd aangeleerde' strategieën voor TGN binnenland zakken sommige kandidaten ook bij herkansingen, wat blijkt uit de antwoordpatronen van deze kandidaten.
Deze kandidaten antwoorden op een dusdanige van normale spraak afwijkende wijze, dat zij lage scores ontvangen voor de deelvaardigheden *vloeiendheid* en *uitspraak*. De gegeven verklaring kan beter onderbouwd worden als een aanvullend onderzoek wordt verricht naar hoe kandidaten zich voorbereiden op de TGN.

Een derde verschil tussen de populaties is, dat kandidaten die de TGN in het buitenland maken minder vaak de TGN herkansen dan kandidaten die de TGN in Nederland maken. Hierdoor bevat de steekproef van de binnenlanddata mogelijk meer kandidaten die al meerdere keren zijn op geweest voor de TGN dan de steekproef van de buitenlanddata. Een examenkandidaat die voor een tweede (of meerdere) keer opgaat, zal zich mogelijk anders gedragen dan een kandidaat die voor de eerste keer examen doet. De populatie van het onderzoek van 2007 bestond overigens uit alleen kandidaten die voor het eerst examen TGN deden. Het verdient aanbeveling voor de vergelijkbaarheid om een dataset te gebruiken die alleen uit kandidaten bestaat die voor de eerste keer het examen doen.

7 Referenties

[1]Onderdeel van CINOP Advies B.V., 's-Hertogenbosch, Nederland, www.cinop.nl.

[2]Language Testing Services, Velp, Nederland.

[3]NCS Pearson Inc., Menlo Park, California, USA, www.VersantTest.com.

[4]Tegenwoordig vormt de afdeling Inburgering onderdeel van het ministerie Sociale Zaken en Werkgelegenheid (SZW).

[5]Kessens, J.M. & Jacobusse, G. (2007). Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland (TNO –DV 2007 C053). Soesterberg: TNO.

[6]Ongecorrigeerde correlatiecoëfficiënt (Pearson's Product Moment Correlation).

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

[8]Common European Framework of Reference for Languages, Council of Europe, 2001.

[9]Landis, J. R., & Koch,G.G. (1977), The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

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

A Onderzoeksprotocol

Research protocol – “Differences between automatic and human scoring of the TGN (2013)”

Datum: 16-09-2013
Authors: Judith Kessens, Olav Aarts, Gert Jacobusse

Research question

In 2007, a study was conducted by TNO to answer two research questions regarding the validity of the “Toets Gesproken Nederlands“ (TGN) (Kessens & Jacobusse, 20071).The ministry of SZW (“Sociale Zaken en Werkgelegenheid”) asked TNO to repeat part of the research (research question 1). The research question is formulated as follows:

“Are there substantial differences between the automatic and human scoring of the TGN?”

Automatic scoring of the TGN

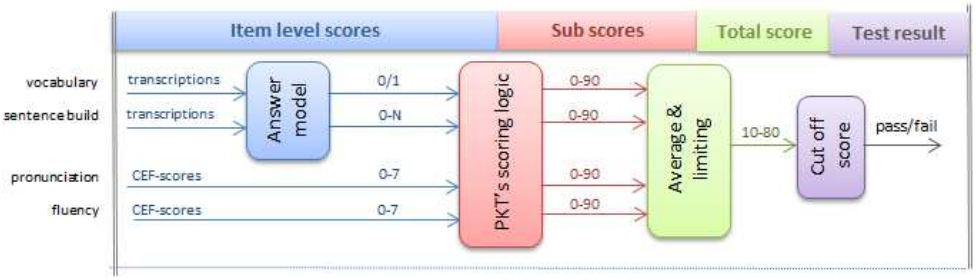
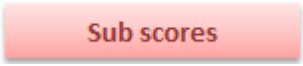


Figure 1 Automatic scoring of the TGN.

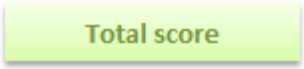
In Figure 1, the automatic scoring of the TGN is schematically presented. For each of the four aspects (vocabulary, sentence build, pronunciation and fluency), four types of scores are generated (each type of scores is indicated with a different color in Figure 1). The four types of scores are explained in detail below:

Item level scores

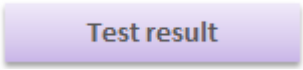
For each aspect, the responses of the candidates are automatically processed by the speech processor. For the aspects vocabulary and sentence build, the speech processor generates word-to-word transcriptions. Subsequently, the transcriptions are transformed into correctness scores, by applying the answer model.
The answer model is used to determine how many words are correct, irrespective of the way words are pronounced. For the aspects, pronunciation and fluency, the quality of the speech is estimated. This done by calculating quality scores, which are based on qualitative measures of the speech (like rate of speech, duration of pauses, sound probabilities).



Next, the item level scores are transformed in subscores by applying PKT's scoring logic. This logic includes; Item Response Theory (IRT) transformations for the content scores, and Non-Linear Transformations (NLT) for the qualitative scores. Also, the subscores are limited within the scoring range of 0-90.



Next, the four subscores are aggregated in the total test score, by taking the average and limiting the range to 10-80.



Finally, the total score is transformed into a pass or a fail by applying the cut-off score.

Approach in 2007

The setup of the research in 2007 differed from the current research. In 2007, the human scores for the rejudge procedure were transformed to subscores by applying the PKT's scoring logic. Therefore, in this research, the methodology of the PKT's scoring logic was used for the human scoring.
Both the *scoring logic* and the *preciseness of the automatic scores* are evaluated at once:

- *Preciseness of the automatic scores*: Both humans and machine deliver exactly the same item level content scores, i.e. word-to-word transcriptions of the answer of a candidate. In this way the preciseness of the transcriptions and CEF-scores are implicitly evaluated. However, possible errors in machine itemscores can be only be explained by impreciseness of the automatic transcriptions or CEF-scores.
- *Facets & PKT's scaling and limiting*: The transformation of the human itemscores in subscores is obtained with the same methodology, namely; 1) Multi-facets Rasch Analysis, 2) scaling, and 3) limiting. The human subscores were partly generated with the automatic system: The Multi-facets Rasch analysis was performed by TNO, whereas the last two steps (scaling and limiting) were done by PKT. The machine subscores are directly generated with the PKT's scoring logic. In this way, the correctness of part of the PKT's scoring logic (i.e. the Multi-facets Rasch Analysis) was implicitly evaluated.

In Figure 2, the human scoring methodology of 2007 is schematically presented. The correlation coefficients were calculated at the level of subscores and total scores.

To summarize, a low correlation can be explained by:
1 Impreciseness of the automatic transcriptions/CEF-scores.
2 Impreciseness of the human transcriptions/CEF-scores.
3 Differences in the Multi-facets Rasch Analysis applied by TNO and PKT.

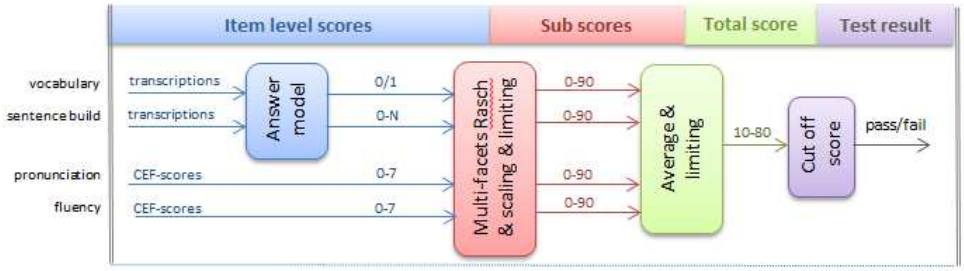


Figure 2 Human scoring of the TGN (research 2007).

Approach in 2013

In the current research, we propose to use the MRT scoring logic for the conversion of the human itemscores to subscores. The rationale behind this choice is that the MRT logic is currently used to transform the human rejudge scores into subscores. With this research design, we will be able to answer the question "are there substantial differences in the human scores obtained during reassessment and the automatic scores?".Another difference with the set up in 2007 is that the correctness of the content aspects (vocabulary and sentence build) is directly scored by the raters. The rationale behind this is a practical one: Within the limited time scope of this project it is not feasible to recruit, train and instruct raters for the transcription task. However, by using the raters that are currently performing the reassessment procedure, no extra recruitment and training is needed.
Both the *scoring logic* and the *preciseness of the automatic scores* are evaluated at once:

- *Preciseness of the automatic scores*: Both humans and machine delivered exactly the same item level content scores, i.e. correctness scores and CEF scores. The implication of using correctness scores is that possible errors in machine itemscores can be explained by impreciseness of the automatic transcriptions and/or incorrectness of the answer model, or by impreciseness of the automatic CEF-scoring.
- *PKT's scoring logic vs. MRT scoring logic*: The transformation of the item levels scores in subscores are obtained with the PKT's scoring logic for the machine scores and with the MRT scoring logic for the human scores. The implication is that possible errors in human subscores can also be explained by incorrectness of the MRT logic.

In Figure 2, the human scoring methodology of 2013 is schematically presented. The correlation coefficients are calculated at the level of subscores and total scores. To summarize, a low correlation can be explained by:

1 Impreciseness of the automatic transcriptions & answer model /CEF-scores.
2 Impreciseness of the human correctness scores/CEF-scores.
3 Differences in the MRT logic and the PKT's scoring logic.

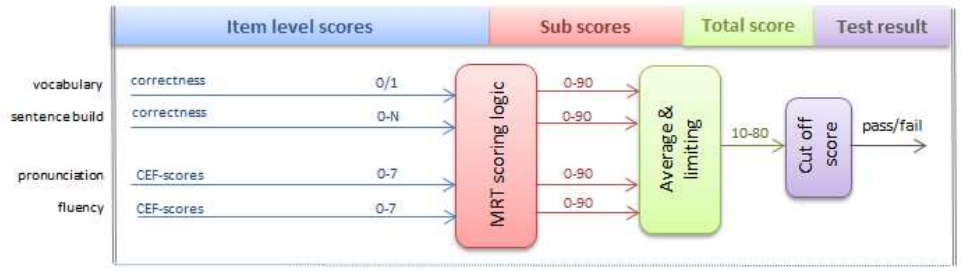


Figure 3 Human scoring of the TGN (research 2013).

Data collection

The data obtained for this research consists of two independent data sets. The data sets will be randomly selected from exams that have been conducted outside the Netherlands ("Buitenland data"), and inside the Netherlands ("Binnenland data") in the period: 1st of January 2012 t/m 1st of April 2013. The selection criteria are exactly the same as the research of 2007. In total, two data sets of 125 valid exams each will be collected.

To draw a representative sample of the population the following procedure is followed: First the mean and standard deviation of all PSTN test-takers was calculated (excluding status 1 and status 2 tests; between January 1, 2012 and March 31, 2013) for the overall and four subscores. Then, a script is written that kept drawing samples of 137 test-takers until their mean and standard deviation for the overall and four subscores matched. It was checked whether the sample distribution reflected the population distribution by comparing the score distribution histogram of the selected sample to that of the population, and indeed this appeared to be the case.

For practical reasons and to avoid halo-effects, each candidate's response will be presented to 4 randomly chosen raters from a pool of N raters. In Table 1, the various types of human scores that will be obtained have been summarized.

Table 1 Type of scores for the human raters.

Aspect	Type of score	Human scores	Number of items per candidate
Vocabulary	dichotomeous: 0/1	Correctness	22
Sentence build	polytomeous: 0-max.	Correctness	23
Fluency	polytomeous: 0-7	CEF-rating	23
Pronunciation	polytomeous: 0-7	CEF-rating	23

Methodology and results

To correctly analyse the two data sets, we first take some preliminary steps to examine the data. The data will be checked for expected range, mean, extreme scores, extreme cases (i.e. exam-candidate combinations), and skewness. Incorrect, impossible or faulty scores will either be omitted or corrected. Possible non-normal distributions will be transformed into normal, Gaussian distributions.

To answer the research question we propose the following steps:

- Though we assume the raters are preselected and screened it is advisable to estimate the reliability of the raters by an intra-class correlation (i.e. Cronbach's Alpha or ICC).
- Average machine and human sub- and total scores and standard deviations.
- Estimate the reliability between the raters. If this might be relatively low, assess if this is due to one (or more) raters and possibly omit them from further analysis.
- Correlations (Pearson) between machine and human score for total and sub scores.
- Graphical representation with the axes of the machine and human score for all subscores and the total score (i.e. scatter plots and correlation table).
- Examine whether there are differences in correlation between human and machine scores for various subscores and the total score. This examination will also be done for the two data sets (Binnenland and Buitenland data).
- Crosstabs: percentage (not) passed by the machine / (not) passed by humans. For both the 26 and 37 total cut-off scores.
- If it turns out that the correlation between the data sets differ substantially from the research of 2007, it is possible to check whether the correlation with attenuation correction does comply with the research of 2007. To perform a split-half method it is likely that a larger dataset is required. Possible methods are Guttman split half – correcting for underestimation of reliability – and (un)equal-length Spearman Brown (in case of unequal number of items. For a robust test-retest at best only 5% of the items of an exam-candidate combination need to be retested.

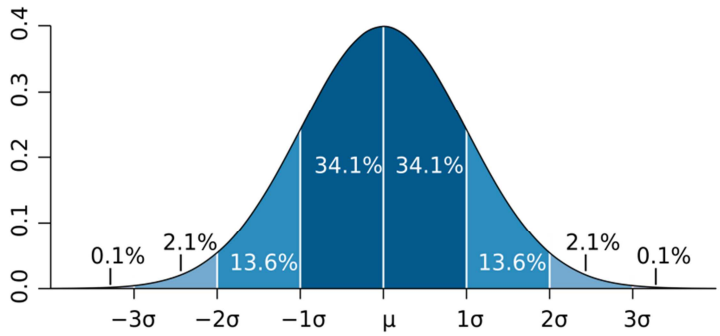
PKT/CINOP provides TNO:

- Machine subscores and total scores.
- Item level human scores together with the formulas implemented in the MRT logic to convert human itemscores to subscores and the formula to convert the four subscores to a human total score.
- Test-retest scores for both datasets (2-4% of the items for the Binnenland data, 5% of the items for the Buitenland data), preferably with time stamps.

References

1 Kessens, J.M. & Jacobusse, G. (2007). Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland (TNO –DV 2007 C053). Soesterberg: TNO.
2 Linacre, J.M. (1988), A Computer Program for the Analysis of Multi-Faceted Data, Chicago, IL: Mesa Press.

B Normale verdeling



Figuur 4 Standaard normale verdeling.

C Cohen's κ -s deelscores binnenland

Tabel 2 Cohen's κ-s voor de deelscore woordenschat, binnenland.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		n.a..	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
2			0,9	1,0	0,8	1,0	n.a.	n.a.	0,9	n.a.	n.a.	0,9	0,8	n.a.	n.a.
3				1,0	0,9	0,9	1,0	0,9	1,0	n.a.	n.a.	0,9	1,0	0,9	n.a.
4					0,9	0,9	0,8	0,9	1,0	n.a.	n.a.	0,9	1,0	0,8	n.a.
5						0,9	n.a.	0,9	0,9	n.a.	n.a.	0,9	0,8	0,8	n.a.
6							0,8	0,9	0,9	n.a.	n.a.	0,9	1,0	0,9	n.a.
7								0,3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
8									0,9	n.a.	n.a.	0,9	1,0	0,9	n.a.
9										n.a.	n.a.	0,9	0,9	0,9	n.a.
10											n.a.	n.a.	n.a.	n.a.	n.a.
11												n.a.	n.a.	n.a.	n.a.
12													1,0	0,9	n.a.
13														n.a.	n.a.
14															n.a.
15															

Tabel 3 Cohen's κ-s voor de deelscore zinsbouw, binnenland.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
2			0,6	0,7	0,6	0,5	0,7	0,5	0,7	n.a.	0,5	0,5	0,6	0,7	n.a.
3				0,7	0,6	0,5	0,6	0,5	0,6	n.a.	0,5	0,5	0,6	0,6	n.a.
4					0,7	0,5	1,0	0,5	0,5	n.a.	0,5	0,5	n.a.	0,6	n.a.
5						0,5	0,6	0,5	0,7	n.a.	0,5	0,5	0,5	0,6	n.a.
6							0,6	0,4	0,5	n.a.	0,3	0,5	0,6	0,5	n.a.
7								0,5	0,7	n.a.	0,5	0,6	0,6	0,6	n.a.
8									0,4	n.a.	0,5	0,6	0,7	0,5	n.a.
9										n.a.	0,6	0,6	0,5	0,6	n.a.
10											n.a.	n.a.	n.a	n.a	n.a.
11												0,5	n.a	n.a	n.a.
12													0,7	0,6	n.a.
13														0,4	n.a.
14															n.a.
15															

Tabel 4 Cohen's κ-s voor de deelscore vloeiendheid, binnenland.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		n.a	n.a	0,4	0,4	0,3	0,3	0,2	n.a	0,5	n.a	0,3	0,4	n.a	n.a
2			n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a
3				-0,1	n.a	n.a	-0,1	0,0	n.a	n.a	n.a	n.a	n.a	n.a	n.a
4					0,3	0,3	0,3	0,0	n.a	n.a	n.a	0,4	0,6	0,5	n,a
5						0,4	0,3	0,0	0,4	0,4	n.a	0,4	0,3	0,4	n,a
6							0,5	0,1	0,9	0,2	n.a	0,4	0,5	0,2	n.a
7								0,0	0,6	0,3	n.a	0,4	0,6	0,3	n.a
8									0,1	0,1	n.a	0,1	0,0	0,1	n.a
9										0,4	n.a	0,5	n.a	n.a	n.a
10											n.a	0,3	n.a	0,4	n.a
11												n.a	n.a	n.a	n.a
12													0,6	0,4	n.a
13														0,3	n.a
14															n.a
15															

Tabel 5 Cohen's κ-s voor de deelscore uitspraak, binnenland.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		n.a.	0,4	n.a.	0,0	0,3	0,1	0,1	0,1	0,2	n.a.	0,3	n.a.	n.a.	n.a.
2			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3				n.a.	0,2	0,3	0,3	0,1	n.a.	0,3	n.a.	0,1	n.a.	n.a.	n.a.
4					0,3	0,1	n.a.	0,4	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
5						0,2	0,2	0,2	0,1	0,3	n.a.	0,1	0,3	0,1	n.a.
6							0,3	0,1	0,1	0,3	n.a.	0,3	0,4	0,0	n.a.
7								0,2	0,2	0,6	n.a.	0,2	0,3	0,1	n.a.
8									0,2	0,3	n.a.	0,1	0,2	0,0	n.a.
9										0,2	n.a.	0,0	0,4	n.a.	n.a.
10											n.a.	0,1	0,3	n.a.	n.a.
11												n.a.	n.a.	n.a.	n.a.
12													0,3	0,0	n.a.
13														n.a.	n.a.
14															n.a.
15															

D Cohen's κ -s deelscores buitenland

Tabel 6 Cohen's κ-s voor de deelscore woordenschat, buitenland.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16
1		n.a.	0,8	0,9	0,9	1,0	0,8	0,9	0,6	n.a.	0,9	0,6	n.a.	n.a.	n.a.
2			1,0	n.a.	0,8	0,8	0,8	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3				0,9	0,9	1,0	0,9	0,9	1,0	n.a.	n.a.	0,9	n.a.	0,9	n.a.
4					0,9	0,9	0,9	0,9	n.a.	n.a.	0,9	0,9	n.a.	n.a.	n.a.
5						0,9	0,9	0,9	0,8	n.a.	0,9	0,9	n.a.	0,9	n.a.
6							0,9	0,9	1,0	n.a.	0,9	0,9	n.a.	0,9	n.a.
7								0,9	1,0	n.a.	0,7	0,9	n.a.	n.a.	n.a.
8									1,0	n.a.	n.a.	1,0	n.a.	n.a.	n.a.
9										n.a.	1,0	1,0	n.a.	n.a.	n.a.
10											n.a.	n.a.	n.a.	n.a.	n.a.
11												0,9	n.a.	n.a.	n.a.
12													n.a.	0,8	n.a.
13														n.a.	n.a.
14															n.a.
15															

Tabel 7 Cohen's κ-s voor de deelscore zinsbouw, buitenland.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16
1		n.a.	n.a.	0,1	0,5	0,5	0,5	n.a.	0,5	n.a.	0,5	0,5	n.a.	n.a.	n.a.
2			n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3				n.a.	0,7	0,6	0,6	n.a.	n.a.	n.a.	0,1	0,5	0,3	n.a.	n.a.
4					0,6	0,5	0,5	n.a.	0,4	n.a.	0,5	0,4	0,3	n.a.	n.a.
5						0,5	0,6	n.a.	0,5	n.a.	0,6	0,5	0,7	n.a.	n.a.
6							0,5	n.a.	0,5	n.a.	0,4	0,4	0,5	n.a.	n.a.
7								n.a.	n.a.	n.a.	0,5	0,5	0,3	n.a.	n.a.
8									n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
9										n.a.	0,4	n.a.	n.a.	n.a.	n.a.
10											n.a.	n.a.	n.a.	n.a.	n.a.
11												0,5	0,4	n.a.	n.a.
12													0,2	n.a.	n.a.
13														n.a.	n.a.
14															n.a.
15															

Tabel 8 Cohen's κ-s voor de deelscore vloeiendheid, buitenland.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16
1		n.a.	0,1	n.a.	0,4	0,4	0,4	0,1	n.a.	n.a.	n.a.	0,3	n.a.	n.a.	n.a.
2			n.a.	n.a.	0,2	0,3	0,3	0,1	0,7	0,2	n.a.	0,4	0,6	0,2	n.a.
3				n.a.	n.a.	0,1	0,2	0,1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4					n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
5						0,4	0,4	0,1	0,3	0,4	0,0	0,3	0,3	0,4	n.a.
6							0,5	0,1	0,5	0,4	0,1	0,4	0,6	0,4	n.a.
7								0,1	n.a.	0,3	n.a.	0,5	n.a.	0,4	n.a.
8									0,1	0,0	n.a.	0,1	0,2	0,2	n.a.
9										0,3	n.a.	0,2	0,7	0,0	n.a.
10											n.a.	0,3	0,0	n.a.	n.a.
11												0,1	n.a.	n.a.	n.a.
12													0,4	0,3	n.a.
13														n.a.	n.a.
14															n.a.
15															

Tabel 9 Cohen's κ-s voor de deelscore uitspraak, buitenland.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16
1		0,2	0,1	n.a.	0,2	0,3	0,3	n.a.	n.a.	n.a.	0,2	0,3	0,1	n.a.	0,4
2			0,2	n.a.	0,1	0,2	0,2	n.a.	n.a.	n.a.	0,1	0,2	0,1	n.a.	0,1
3				n.a.	0,2	0,3	0,1	n.a.	n.a.	n.a.	n.a.	0,3	0,0	n.a.	n.a.
4					n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
5						0,3	0,3	n.a.	n.a.	n.a.	0,1	0,2	0,3	n.a.	0,3
6							0,4	n.a.	n.a.	n.a.	0,2	0,3	0,4	n.a.	0,4
7								n.a.	n.a.	n.a.	0,2	0,3	n.a.	n.a.	0,4
8									n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
9										n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
10											n.a.	n.a.	n.a.	n.a.	n.a.
11												0,4	0,3	n.a.	0,2
12													0,3	n.a.	0,3
13														n.a.	n.a.
14															n.a.
15															

Distributielijst

Onderstaande instanties/personen ontvangen een volledig exemplaar van het rapport.

1/5	Pépin Cabo, Directie Integratie en Samenleving, Ministerie van SZW
6	Jo Fond Lam, CINOP
7	Prof. Dr. John H.A.L. de Jong, LTS
8	Jennifer Manning & Suzuki Masanori, NCS Pearson, Inc.
9	TNO, vestiging Groningen (Eemsgolaan), dr. O.A.J. Aarts
10/12	TNO, vestiging Soesterberg, dr.ir. J.M. Kessens
13/14	TNO, (archief) vestiging Soesterberg