Research Center voor Examinering en Certificering (RCEC)

Beoordeling TNO-rapport ‘Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland’ (TNO-DV 2007 C053)

Dr. P.F. Sanders

Universiteit Twente

Research Center voor Examinering en Certificering (RCEC)

Beoordeling TNO-rapport 'Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland' (TNO-DV 2007 C053)

Dr. P.F. Sanders
Directeur Research Center voor Examinering en Certificering (RCEC)
Nieuwe Oeverstraat 50
6811 JB ARNHEM
E: piet.sanders@cito.nl
T: 026-3521414

Cito en Universiteit Twente
Arnhem, 2007

Beoordeling TNO-rapport 'Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland' (TNO-DV 2007 C053)

Samenvatting

Hoewel bij de methodologie van het TNO-onderzoek kritische kanttekeningen geplaatst kunnen worden met betrekking tot de keuze van het psychometrisch model, de koppeling van datasets en het (zelf) niet uitvoeren van de regressieanalyse, heeft TNO de door het Consortium gehanteerde methodologie zeker verbeterd. Met inachtneming van wat hiervoor is opgemerkt, heeft TNO terecht geconstateerd dat de cesuur voor de TGN te soepel ingesteld is.
Met de door TNO gehanteerde methodologie kan echter niet bepaald worden hoeveel strenger de cesuur zou moeten zijn. Voor dat laatste dient nader onderzoek plaats te vinden.

1. Inleiding

In het debat van 5 juli 2007 heeft de Tweede Kamer aangegeven dat zij een extra onderzoek wenst naar de cesuurbepaling van de Toets Gesproken Nederlands (TGN) in het kader van inburgering en naturalisatie. Het Ministerie van VROM, Directoraat-Generaal Wonen, Wijken en Integratie heeft het Research Center voor Examinering en Certificering (RCEC)[1] verzocht een antwoord te geven op de volgende onderzoeksvraag:

'Is het TNO-onderzoek naar de cesuurbepaling zorgvuldig uitgevoerd en is daarmee de conclusie gerechtvaardigd dat de cesuur voor de TGN te soepel is ingesteld?'

Het onderhavige rapport bestaat uit vier onderdelen. In het eerste onderdeel wordt de globale onderzoeksvraag van VROM nader afgebakend. In het tweede onderdeel wordt de uitvoering van het RCEC-onderzoek beschreven en wie bij de uitvoering van het RCEC-onderzoek betrokken waren. In het derde onderdeel worden de resultaten van het RCEC-onderzoek gepresenteerd. Het vierde onderdeel bevat de conclusies van het onderzoek en een aantal overwegingen voor de opdrachtgever.

2. Afbakening van het onderzoek

In de 'Verantwoording' van het TNO-rapport wordt aangegeven dat TNO verantwoordelijk was voor de opzet van het onderzoek maar dat aan de uitvoering andere partijen bijgedragen hebben. Die andere partijen waren de leden van het Consortium dat de TGN ontwikkeld heeft, namelijk CINOP, LTS en Ordinate. Daarnaast hebben de onderzoekers van TNO zich bij de opzet en voortgang van het onderzoek laten adviseren door een externe deskundige op het gebied van psychometrie.

Het onderzoeksrapport van TNO betreft twee onderzoeksvragen. Deze twee vragen zijn:
- Onderzoeksvraag 1: Zijn er substantiële verschillen in de beoordelingen tussen het systeem dat automatisch uitslagen genereert en menselijke beoordelingen?
- Onderzoeksvraag 2. Is de zak-/slaaggrens van de TGN op het goede niveau ingesteld?

Gezien de opdracht aan het RCEC heeft het onderzoek van het RCEC zich uitsluitend gericht op onderzoeksvraag 2. Op pagina 22 van het TNO-rapport wordt vermeld dat TNO voor het vaststellen van de zak-/slaaggrens dezelfde methode gebruikt als die door het Consortium gebruikt is zodat de nieuwe resultaten te vergelijken zijn met de eerder verkregen resultaten. In een vooronderzoek heeft TNO de door het Consortium gebruikte methode bestudeerd en toelichting gevraagd en verkregen van het Consortium. Dat laatste was nodig omdat, zoals ook het RCEC heeft kunnen constateren, de beschrijving van de methodologie in de publicatie 'Verantwoording Toets Gesproken Nederlands' van het Consortium ontoereikend was.

Door TNO werd tijdens het vooronderzoek geconstateerd dat er twee verbeteringen mogelijk waren in de methode die door het Consortium gebruikt was om verschillende datatsets gezamenlijk te analyseren. Vandaar dat TNO onderzoeksvraag 2 in twee deelvragen heeft opgesplitst.

[1] Het RCEC, een samenwerkingsverband van de Universiteit Twente en het Cito, is een onafhankelijk onderzoeksinstituut op het gebied van examinering en certificering.

De twee onderzoeksvragen zijn:
- Onderzoeksvraag 2a: Vinden we dezelfde zak-/slaaggrens als we de schalingsmethode herhalen met de twee verbeteringen die uit het vooronderzoek volgen?
- Onderzoeksvraag 2b: Wordt dezelfde schaal gevonden voor de data verzameld in het binnen- versus buitenland?

Wat betreft het onderzoek naar onderzoeksvraag 2b wordt op pagina 37 van het TNO-rapport opgemerkt dat 'Vanwege de bias (volgens TNO waarschijnlijk veroorzaakt door de schattingsmethode die de FACETS-software gebruikt), het verschil in telefoonnetwerk en het feit dat conclusies over verhaaltjes navertellen niet één op één naar de machine te vertalen zijn, kunnen geen harde conclusies getrokken worden over de schaling in het buitenland.'

Gegeven de conclusies van TNO over onderzoeksvraag 2b en gegeven ook de opdracht van VROM aan het RCEC, heeft het onderzoek van het RCEC zich dan ook uitsluitend gericht op het onderzoek van TNO naar onderzoeksvraag 2a:

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

3. Uitvoering van het onderzoek

De onderzoeksaanpak voor het RCEC-onderzoek bestond uit:
1. Documentonderzoek.
2. Bevraging van de auteurs van het TNO-rapport.
3. Rapportage.

Ad 1. Documentonderzoek

Voor het documentonderzoek waren door het Ministerie van VROM een aantal publicaties en onderzoeksrapporten ter beschikking gesteld. Voor de beantwoording van de onderzoeksvraag bleken met name twee publicaties relevant te zijn. De eerste publicatie betrof de 'Verantwoording Toets Gesproken Nederlands', een publicatie van het CINOP van 2005 met als auteurs: A. Kerkhof, P. Poelmans, J. de Jong en M. Lening. De tweede publicatie betrof het onderzoeksrapport 'Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland', een publicatie van TNO (TNO-DV 2007 C053) van 2007 met als auteurs: J. Kessens en G. Jacobusse. Het documentonderzoek werd uitgevoerd door:
- Dr. P. Sanders, psychometricus/directeur van het Research Center voor Examinering en Certificering.
- Dr. ir. T. Eggen, senior medewerker/psychometricus bij het Psychometrisch Onderzoek-centrum van het Cito.

Ad 2. Bevraging van de auteurs van het TNO-rapport

Op 23 augustus 2007 heeft bijna drie uur overleg plaatsgevonden over het onderzoeksrapport van TNO. Deelnemers aan het overleg waren:
- Drs. H-J. Vink, businessunitmanager bij TNO.
- Dr. ir. J. Kessens, onderzoeker spraaktechnologie bij TNO Defensie en Veiligheid en medeauteur van het voornoemde onderzoeksrapport.
- Drs. G. Jacobusse, statisticus bij TNO Kwaliteit van Leven en medeauteur van het voornoemde onderzoeksrapport.
- Dr. P. Sanders, psychometricus/directeur van het Research Center voor Examinering en Certificering.

- Prof. dr. G. Maris, senior medewerker/psychometricus bij het Psychometrisch Onderzoek-centrum van het Cito en hoogleraar Psychometrie aan de Universiteit van Amsterdam.
- Dr. T. Bechger, senior medewerker/psychometricus bij het Psychometrisch Onderzoek-centrum van het Cito.
- Dr. A. Béguin, senior medewerker/psychometricus bij het Psychometrisch Onderzoek-centrum van het Cito.

Ad 3. Rapportage

Dr. P. Sanders, de directeur van het RCEC, heeft de rapportage verzorgd. Hij is daarbij inhoudelijk ondersteund door de vier voornoemde psychometrici die werkzaam zijn bij het Cito. Het RCEC, in de persoon van dr. P. Sanders, is eindverantwoordelijk voor de rapportage.

4. Resultaten van het RCEC-onderzoek

Het RCEC-onderzoek heeft geresulteerd in vier kritiekpunten op het TNO-onderzoek. Het eerste kritiekpunt betreft het psychometrisch model waarmee de data geanalyseerd zijn, het tweede kritiekpunt betreft het bepalen van de regressiefuncties, het derde kritiekpunt betreft de bepaling van de afstanden van de grenswaarden en het vierde kritiekpunt betreft de plaatsing van de grenswaarden, met name die van de zak-/slaaggrens of cesuur.

4.1 Multifacet rating scale model en FACETS

De data die door TNO geanalyseerd zijn, betreffen responses op items (taaluitingen) van kandidaten die door verschillende beoordelaars met behulp van de CEF-schaal beoordeeld zijn. TNO is in navolging van het Consortium bij de analyse van de data (van de vier deelvaardigheden) uitgegaan van het multifacet rating scale model (MRSM) en heeft de analyses uitgevoerd met het programma FACETS. Het MRSM is naar mening van het RCEC echter geen geschikt model indien meerdere beoordelaars dezelfde uitingen beoordelen.
Volgens het MRSM is een nieuwe beoordeling van dezelfde uiting hetzelfde als een beoordeling van een nieuwe uiting. Ter illustratie een voorbeeld. Stel dat bij een kunstschaatswedstrijd de onfortuinlijke deelneemster valt en dat de 10 juryleden die dit zien haar prestatie beoordelen. Een andere kunstschaatsster valt echter bij 10 opeenvolgende kunstschaatswedstrijden wat door 1 jurylid gezien wordt die 10 keer haar prestatie beoordeelt. Deze twee beoordelingssituaties zijn voor het MRSM gelijk wat in het geval van de twee kunstschaatsers betekent dat ze als even vaardig beoordeeld worden. Het MRSM houdt echter ten onrechte geen rekening met de conditionele afhankelijkheid in de beoordelingen. Beoordelingen van dezelfde uiting zijn namelijk alleen onafhankelijk conditioneel op de kwaliteit van de uiting maar niet conditioneel op de vaardigheid zoals door het MRSM verondersteld wordt. In plaats van het MRSM hadden naar mening van het RCEC zowel het Consortium als TNO een analysemodel moeten gebruiken waarin of expliciet rekening gehouden wordt met deze conditionele afhankelijkheid of waarin wel op een juiste wijze met de afhankelijkheden omgegaan kan worden.

De data zijn door het Consortium en TNO geanalyseerd met het programma FACETS. Dit programma gebruikt de joint maximum likelihood schattingsmethode waarvan bekend is dat die tot inconsistente parameterschattingen aanleiding geeft. Omdat in het onderhavige geval sprake is van een zeer onvolledige dataverzameling omdat niet alle beoordelaars alle taaluitingen van alle kandidaten beoordelen, zou die inconsistentie wel eens problematisch kunnen zijn. Het TNO-rapport lijkt hier op pagina 35 bij de bespreking van onderzoeksvraag 2b ook naar te verwijzen: 'Onzekerheid over de maximum-likelihood parameterschattingen resulteert in schattingen die bias (vertekening) naar de extremen hebben.' Behalve bij de schattingsmethode kunnen ook bij de gehanteerde passingsmaten kanttekeningen geplaatst worden.

4.2 Bepaling van de regressiefuncties

Op pagina 23 van het TNO-rapport wordt aangegeven dat met behulp van regressiefuncties de relatie bepaald is tussen de theta's van FACETS en de theta's van TGN. Twee opmerkingen hierover. De eerste opmerking is dat het RCEC het een vreemde gang van zaken vindt dat de regressieanalyse niet door TNO uitgevoerd is maar door een lid van het Consortium. Dat deze werkwijze noodzakelijk was om volgens het Consortium bedrijfsgevoelige informatie te beschermen, vindt het RCEC geen valide argument. De tweede opmerking is dat de resultaten van de regressieanalyse, met name ook een visuele presentatie van de thetaverdelingen van FACETS en TGN, ontbreken.

4.3 Bepaling afstanden grenswaarden

Op pagina 22 - 27 van het TNO-rapport worden de drie stappen van de schalingsmethode beschreven. In stap 1 worden de grenswaarden geschat van de CEF-schaal die opgedeeld is in 8 discrete CEF-niveaus. In stap 2 worden de regressiefuncties bepaald waarmee de vaardigheidsschattingen verkregen met de FACETS-analyse op de TGN vaardigheidsschaal geprojecteerd worden (zie ook punt 4.2). In stap 3 worden de TGN vaardigheidsschattingen uit stap 2 met behulp van een transformatiefunctie afgebeeld op de TGN rapportageschaal van TGN die van 10 tot en met 80 loopt.
Op pagina 26 - 30 van het TNO-rapport wordt stap 1 van de schalingsmethode besproken en met de door TNO voorgestelde verbeteringen gepresenteerd. Voor het bepalen van de CEF-grenswaarden heeft het Consortium gebruik gemaakt van twee datasets die apart van elkaar verzameld zijn. Om de twee datasets gezamenlijk met FACETS te kunnen analyseren, moeten de data voor twee van de drie facetten (kandidaten, items en beoordelaars) overlap vertonen. De datasets vertonen echter geen overlap via facet 1 (kandidaten) omdat er geen ge-meenschappelijke kandidaten zijn en ook niet via facet 2 (items) omdat in dataset 1 andere itemtypes gebruikt worden dan in dataset 2. De onderzoekers van TNO constateerden dat de koppeling die het Consortium via imputatie van 20 CEF-beoordelingen van een kunstmatige beoordelaar tot stand had gebracht, niet correct was uitgevoerd. Op basis van twee aannames heeft TNO een verbetering van de koppeling tussen de twee datasets voorgesteld en uitgevoerd. Gegeven dat een aantal beoordelaars in beide datasets voorkomt, is de eerste aanname dat aangenomen mag worden dat een beoordelaar in beide datasets even streng is. De tweede aanname is om tijdelijk één item uit dataset 1 gelijk te stellen aan één item uit dataset 2. Hierdoor ontstaat een koppeling die het mogelijk maakt om de datasets samen te analyseren. Deze analyse werd door TNO 36 keer met steeds andere items herhaald met als doel na te gaan of de schattingen van de zak-/slaaggrenzen gevoelig zijn voor deze manier van koppelen. Naar aanleiding van de resultaten van de analyses merkt TNO op pagina 30 op 'dat de andere manier van datakoppeling geen gevolgen heeft voor de tussen de CEF-grens-waardenFacets die worden gevonden.', en concludeert op pagina 33 dat 'De gevonden afstanden tussen CEF-grenswaarden komen goed overeen met de gevonden afstanden.'

Op pagina 26 van het TNO-rapport wordt terecht opgemerkt dat het probleem dat beide datasets niet samen te analyseren zijn, theoretisch onoplosbaar is. Hoewel de resultaten van de analyses onderling vergelijkbaar zijn en goed overeenkomen met die van het Consortium, is dat volgens het RCEC onvoldoende bewijs voor de conclusie van TNO dat de afstanden tussen de CEF-grenswaarden juist zouden zijn. Deze conclusie mag volgens het RCEC alleen maar getrokken worden op basis van onderzoek waarbij de 2 datasets wel gezamenlijk geanalyseerd (kunnen) worden door een koppeling via 'common persons' en/of 'common items'. De datasets hadden trouwens wel gekoppeld kunnen worden via data van TGN maar van deze mogelijkheid heeft TNO geen gebruik gemaakt of kunnen maken omdat het bedrijfsveilige informatie betrof.

4.4 Plaatsing grenswaarden

Op pagina 24 en 25 wordt in het TNO-rapport het volgende opgemerkt. 'Het doel van de FACETS-analyse is om CEF-grenswaarden te vinden die een plaats hebben op de taalvaardigheidsschaal van de kandidaten, onafhankelijk van beoordelaarsstrengheid. De CEF-grenswaarden die uit de FACETS-analyse volgen zijn juist wel relatief aan beoordelaars-strengheid en itemmoeilijkheid. Er moet een keuze gemaakt worden om de relatieve CEF-grenswaarden uit de FACETS-analyse uit te drukken op een vaste plaats op de kandidaat-vaardigheidsschaal. Een logisch verdedigbare keuze is om de CEF-grenswaarden uit te drukken ten opzichte van de gemiddelde beoordelaarsstrengheid en itemmoeilijkheid. De schaal is al verankerd ten opzichte van een gemiddelde beoordelaarsstrengheid van 0. De gecentraliseerde CEF-grenswaarden moeten echter nog gecorrigeerd worden voor de gemiddelde itemmoeilijkheid.'
Op pagina 31 en 32 presenteert TNO de resultaten van de correctie voor de gemiddelde beoordelaarsstrengheid en itemmoeilijkheid. Vergeleken met de oorspronkelijke schalings-methode van het Consortium waarbij alleen gecorrigeerd werd voor de gemiddelde beoordelaarsstrengheid, zouden volgens TNO de CEF-grenswaarden, dus ook de zak-/slaaggrens of cesuur, met ruim één CEF-niveau aangepast moeten worden.

Op pagina 31 van het TNO-rapport stelt TNO: 'Een logisch verdedigbare keuze is om de CEF-grenswaarden uit te drukken ten opzichte van de gemiddelde beoordelaarsstrengheid en itemmoeilijkheid.' Volgens het RCEC resulteert de correctie voor gemiddelde beoordelaars-strengheid door het Consortium in een populatieafhankelijke cesuur, d.w.z. dat de cesuur hoger komt te liggen in geval de populatie van kandidaten vaardiger is, en dat het gemiddelde van de CEF-grenswaarden gelijk is aan het populatiegemiddelde. De correctie voor gemiddelde beoordelaarsstrengheid en itemmoeilijkheid door TNO resulteert volgens het RCEC in een itemafhankelijke cesuur, d.w.z. dat de cesuur hoger komt te liggen als de items moeilijker zijn, en dat het gemiddelde van de CEF-grenswaarden gelijk is aan de gemiddelde itemmoeilijkheid.

Bij gebrek aan de feitelijke toetsresultaten heeft het RCEC op basis van de data die vermeld staan in het TNO-rapport enige simulaties uitgevoerd. Die simulaties bevestigen dat de methodologie van het Consortium fout is en dat de door het Consortium voorgestelde cesuur inderdaad te soepel is. TNO heeft de door het Consortium gemaakte methodologische fout hersteld en stelt terecht een strengere cesuur voor. Volgens het RCEC is de tekortkoming van de door TNO gehanteerde methodologie echter dat daarmee niet bepaald kan worden hoeveel strenger de cesuur zou moeten zijn wat TNO echter wel doet.
Het voorgaande kan toegelicht worden aan de hand van de extra onderzoeksresultaten in bijlage G van het TNO-rapport. De eerste en laatste kolom van Tabel G.1 in Bijlage G laten zien dat de gemiddelde oordelen van deskundige beoordelaars het meest overeenkomen met de door TNO gecorrigeerde TGN scores. Volgens de (beperkte) simulaties die het RCEC op basis van de parameters zoals gerapporteerd in de Figuren G.2a, G.2b en G2c, uitgevoerd heeft, blijkt echter

dat de variantie in de oordelen[2] vaak dusdanig groot is dat het gemiddelde geen recht doet aan de beoordeling die aan de kandidaat toegekend wordt. Het door TNO voorgestelde cesuurvoorstel zou terecht zijn indien uit nader onderzoek van de feitelijke data zou blijken dat de variantie tussen beoordelaars en items gering is. Dat laatste is echter in tegenspraak met de in Bijlage G gerapporteerde parameterschattingen.

5. Conclusie en overwegingen

Het Ministerie van VROM, Directoraat-Generaal Wonen, Wijken en Integratie heeft het Research Center voor Examinering en Certificering (RCEC) verzocht een antwoord te geven op de volgende onderzoeksvraag:

'Is het TNO-onderzoek naar de cesuurbepaling zorgvuldig uitgevoerd en is daarmee de conclusie gerechtvaardigd dat de cesuur voor de TGN te soepel is ingesteld?'

Het RCEC-onderzoek heeft tot de volgende conclusies geleid:

1. Hoewel bij de methodologie van het TNO-onderzoek kritische kanttekeningen geplaatst kunnen worden met betrekking tot de keuze van het psychometrisch model, de koppeling van datasets en het (zelf) niet uitvoeren van de regressieanalyse, heeft TNO de door het Consortium gehanteerde methodologie zeker verbeterd.
2. Met inachtneming van wat hiervoor is opgemerkt, heeft TNO terecht geconstateerd dat de cesuur voor de TGN te soepel ingesteld is.
3. Met de door TNO gehanteerde methodologie kan echter niet bepaald worden hoeveel strenger de cesuur zou moeten zijn. Voor dat laatste dient nader onderzoek plaats te vinden.

Naar aanleiding van het onderzoek en de conclusies, wil het RCEC de opdrachtgever nog de volgende overwegingen meegeven:

- De eerste overweging heeft betrekking op de scoringsregel die bij de TGN gehanteerd wordt. Een scoringsregel maakt de kandidaat in ieder geval duidelijk wat de minimum- en maximumscore is die op een toets of een examen behaald kan worden en wat individuele items of opdrachten aan het zakken of slagen voor het examen kunnen bijdragen. Transparantie van de scoringsregel is dan ook een eerste vereiste voor toetsen of examens, ongeacht of de resultaten in ruwe scores of afgeleide scores gerapporteerd worden. In de publicaties over de TGN wordt niet of onvoldoende duidelijk gemaakt hoe de scores die behaald worden op de vier onderscheiden vaardigheden een bepaald CEF-niveau opleveren. De gebruikte technologie bij de afname van de TGN zou geen reden mogen zijn die noodzakelijke transparantie niet te verschaffen.

[2] Variantie tussen de oordelen is opgebouwd uit alle beoordelingen die de beoordelaars aan alle items/taaluitingen van een kandidaat gegeven hebben.

- De tweede overweging betreft de afname van de TGN in de praktijk. Bij de TGN wordt gebruik gemaakt van een itembank die volgens de publicaties die het RCEC geraadpleegd heeft uit ongeveer 1000 items bestaat. Aan een kandidaat worden 45 items uit de itembank voorgelegd en aan een andere kandidaat 45 (gedeeltelijk) andere items. Hoewel de items aselect uit de op moeilijkheidsgraad gestratificeerde itembank getrokken worden, geeft dat geen garantie dat de verschillende toetsen voor wat betreft moeilijkheidsgraad vergelijkbaar zijn. Indien de cesuur op de TGN schaal voor alle random getrokken toetsen gelijk is, zou dat betekenen dat afhankelijk van de moeilijkheidsgraad van de toets, kandidaten bevoor- of benadeeld worden.
- De derde overweging is om onderzoek te (laten) doen naar andere methoden die in de literatuur en praktijk gebruikt worden om de cesuur vast te stellen. Het nadeel van de methode die nu voor de TGN gehanteerd wordt, is moeilijk door de gebruikers van de TGN te bevatten. Hierdoor ontbreekt de mogelijkheid tot een transparante legitimering van de cesuur naar de doelgroep.

Arnhem, 5 september 2007

Dr. P.F. Sanders
Directeur Research Center voor Examinering en Certificering (RCEC)
Nieuwe Oeverstraat 50
6811 JB ARNHEM
E: piet.sanders@cito.nl
T: 026-3521414

Research Center voor Examinering en Certificering (RCEC)

Addendum bij Beoordeling TNO-rapport 'Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland' (TNO-DV 2007 C053)

Dr. P.F. Sanders

Universiteit Twente

Research Center voor Examinering en Certificering (RCEC)

Addendum bij Beoordeling TNO-rapport 'Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland' (TNO-DV 2007 C053)

Dr. P.F. Sanders
Directeur Research Center voor Examinering en Certificering (RCEC)
Nieuwe Oeverstraat 50
6811 JB ARNHEM
E: piet.sanders@cito.nl
T: 026-3521414

Cito en Universiteit Twente
Arnhem, 2007

Addendum bij Beoordeling TNO-rapport 'Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland' (TNO-DV 2007 C053)

Samenvatting

Naar aanleiding van de conclusies van het RCEC rapport van 5 september 2007 heeft de Directie Inburgering & Integratie, afdeling BU van het ministerie van VROM, het RCEC verzocht verder onderzoek te doen. Het RCEC heeft de data die in het oorspronkelijke TNO onderzoek gebruikt zijn onderzocht en is van mening dat als gevolg van de verschillen die er tussen beoordelingen van beoordelaars bestaan bij het beoordelen/classificeren van de taaluitingen van de kandidaten en de tekortkomingen van de data geen methodologisch verantwoorde cesuur kan worden vastgesteld. Het RCEC heeft op verzoek van de opdrachtgever de RCEC beoordeling van het TNO rapport (TNO-DV 2007 C053) en een verslag met de resultaten van het onderzoek van de data ter beschikking gesteld aan TNO en met TNO besproken. Naar aanleiding van de bespreking van de RCEC beoordeling van het TNO rapport en het aanvullende verslag, heeft ook TNO de data nogmaals onderzocht. De onderzoekers van TNO blijven van mening dat de verschillen tussen de beoordelingen van beoordelaars niet dusdanig groot zijn dat daardoor geen methodologisch verantwoorde cesuur vastgesteld zou kunnen worden.

Gegeven het voorgaande geeft het RCEC de opdrachtgever de volgende twee overwegingen:

1. Het RCEC heeft de conclusie van het TNO onderzoek dat de cesuur van de TGN te soepel ingesteld is, bevestigd. Het RCEC en TNO blijven van mening verschillen of met de data die nu voorhanden zijn een methodologisch verantwoordbare cesuur vastgesteld kan worden. Aanbevolen wordt nieuw onderzoek naar de verantwoordbaarheid van de TGN cesuur te laten uitvoeren. Hierbij dient er wel rekening mee te worden gehouden dat op basis van dat onderzoek wellicht ook geconcludeerd zou kunnen worden dat het met de TGN in zijn huidige opzet niet mogelijk is om de doelgroep van kandidaten voldoende betrouwbaar, dat wil zeggen met relatief geringe percentages ten onrechte gezakte en geslaagde kandidaten, te kunnen beoordelen als zijnde van een lager dan A1min niveau of van een A1min niveau.

2. In afwachting van de uitkomsten van voornoemd onderzoek wordt voorgesteld de cesuur op de TGN aan te passen overeenkomstig het voorstel dat geformuleerd is in de brief van de minister van Wonen, Wijken en Integratie van 29 mei aan de Tweede Kamer. Bij de nieuwe, hogere cesuur zullen uiteraard meer kandidaten zakken dan bij de oorspronkelijke cesuur maar een verwacht zakpercentage van vijfentwintig procent kan vergeleken met de vele examentoetsen die bijvoorbeeld jaarlijks in het voortgezet onderwijs worden afgenomen zeker niet uitzonderlijk hoog genoemd worden. Met de hogere cesuur correspondeert een beheersingspercentage van iets meer dan dertig procent op onderscheiden onderdelen van de TGN. Voor het onderdeel 'woordenschat' betekent dit bijvoorbeeld dat nu 7 van de 22 gesproken antwoorden het goede antwoord moeten bevatten terwijl dat voorheen 4 goede antwoorden waren. Het nieuwe beheersings-percentage mag dan ook als alleszins redelijk beschouwd worden.

1. Inleiding

In het RCEC rapport van 5 september 2007 wordt naar aanleiding van de beoordeling die het Research Center voor Examinering en Certificering (RCEC)[3] op verzoek van de Directie Inburgering & Integratie, afdeling BU van het ministerie van VROM verricht heeft van het TNO rapport 'Onderzoek naar de kwaliteit van het inburgeringsexamen buitenland' (TNO-DV 2007 C053) het volgende geconcludeerd:

1. Hoewel bij de methodologie van het TNO onderzoek kritische kanttekeningen geplaatst kunnen worden met betrekking tot de keuze van het psychometrisch model, de koppeling van datasets en het (zelf) niet uitvoeren van de regressieanalyse, heeft TNO de door het Consortium gehanteerde methodologie zeker verbeterd.
2. Met inachtneming van wat hiervoor is opgemerkt, heeft TNO terecht geconstateerd dat de cesuur voor de TGN te soepel ingesteld is.
3. Met de door TNO gehanteerde methodologie kan echter niet bepaald worden hoeveel strenger de cesuur zou moeten zijn. Voor dat laatste dient nader onderzoek plaats te vinden.

In de derde conclusie wordt door het RCEC voorgesteld om nader onderzoek te doen. De data die het mogelijk maakten dat onderzoek uit te voeren zijn op verzoek van VROM door het Consortium aan het RCEC ter beschikking gesteld. Het RCEC heeft de data geanalyseerd en de onderzoeksresultaten op verzoek van VROM besproken met TNO. Het RCEC heeft de data ter beschikking gesteld aan TNO die dezelfde data ook geanalyseerd heeft.

Het onderhavige rapport bestaat uit drie onderdelen. In het eerste onderdeel wordt de uitvoering van het RCEC onderzoek beschreven en wie bij de uitvoering van het RCEC onderzoek betrokken waren. In het tweede onderdeel worden de resultaten van het RCEC onderzoek en de resultaten van het overleg met TNO besproken. Het derde onderdeel bevat de conclusies van het onderzoek en overwegingen voor de opdrachtgever.

2. Uitvoering van het onderzoek

De werkzaamheden van het RCEC ten behoeve van het onderhavige onderzoek bestonden uit:

1. Analyse van de data die door het Consortium ter beschikking waren gesteld.
2. Overleg met TNO over het RCEC rapport en de data analyse.

Ad 1. Data analyse

Voor het onderzoek is gebruik gemaakt van data die ter beschikking gesteld zijn door het Consortium in de persoon van Dr. J.H.A.L. de Jong van LTS. De data betreffen de beoordelingen van 534 examenkandidaten. Van elke kandidaat zijn de volgende data beschikbaar:

- de beoordelingen op de gehanteerde CEF-schaal (<A1min, A1min, A1, A2, B1, B2, C1, C2) van zijn/haar taaluitingen van een aantal beoordelaars;
- zijn/haar vaardigheidsschatting op de theta-schaal;
- zijn/haar TGN score volgens het Consortium;
- zijn/haar TGN score volgens TNO;
- indicatie of toetsdata gebaseerd waren op dataset 1 of dataset 2.

[3] Het RCEC, een samenwerkingsverband van de Universiteit Twente en het Cito, is een onafhankelijk onderzoeksinstituut op het gebied van examinering en certificering.

Om te kunnen bepalen of met deze data een verantwoordbare cesuur vastgesteld kon worden, werden twee analyses uitgevoerd:

1. Bepaling van de consistentie van beoordelingen van beoordelaars

   Van elke kandidaat is nagegaan of diens taaluitingen door de beoordelaars consistent beoordeeld werden. De conclusie is dat bij zeer veel kandidaten de verschillen in beoordelingen door de beoordelaars zodanig groot waren dat een gemiddelde beoordeling van een kandidaat een onterechte beoordeling geweest zou zijn.

2. Cesuurbepaling door middel van contrasterende groepen
   Uitgaande van voornoemde gegevens van 534 kandidaten is met behulp van een cesuur-bepalingsprocedure, geheten contrasterende groepen, geprobeerd een verantwoordbare cesuur te vinden. De cesuurbepalingsprocedure komt er bij deze data op neer dat we voor elke kandidaat nagaan of 80% of meer van de beoordelingen de kandidaat als lager dan A1min of als A1min of hoger classificeren. Uitgaande van 80% worden 82 kandidaten als lager dan A1min geclassificeerd, 325 kandidaten als A1min of hoger, terwijl over 127 kandidaten minder dan 80% overeenstemming was. Van deze drie groepen kandidaten wordt achtereenvolgens een cumulatieve verdeling van hun TGN scores bepaald. Voor elke mogelijke cesuur kan nu bepaald worden wat binnen deze populatie van kandidaten de kans is dat een kandidaat ten onrechte zakt, dan wel ten onrechte slaagt. Indien beide kansen voor een beoogde cesuur voldoende klein zijn, heeft men een voor deze populatie verantwoordbare cesuur. Indien de populatie echter niet representatief is voor de doelpopulatie, of indien de doelpopulatie naar verwachting over de tijd verandert van samenstelling dan kan men geen verantwoorde cesuur vaststellen. De data die door het Consortium ter beschikking gesteld waren, bestonden uit twee datasets, dataset 1 en dataset 2. Beide datasets bleken zodanige tekortkomingen te hebben dat op basis van deze datasets geen verantwoorde cesuur vastgesteld kon worden. Zo bevatte dataset 1 zeer weinig beoordelingen van taaluitingen van de kandidaten, terwijl dataset 2 niet representatief was voor de doelpopulatie omdat deze dataset nagenoeg geen <A1min kandidaten bevatte. Beide datasets zouden dan ook tot nogal verschillende cesuren geleid hebben.

Conclusie

Gegeven het gebrek aan consistentie van de beoordelingen van beoordelaars en de tekortkomingen van de data, is de conclusie dat het niet mogelijk is om met deze data een verantwoordbare cesuur vast te stellen.

Ad 2. Overleg met TNO over het RCEC rapport en de aanvullende data analyse.

Op 24 oktober 2007 heeft drie uur overleg plaatsgevonden over het RCEC rapport en de aanvullende data analyse. Deelnemers aan het overleg waren:
- Drs. H-J. Vink, businessunitmanager bij TNO.
- Dr. ir. J. Kessens, onderzoeker spraaktechnologie bij TNO Defensie en Veiligheid en medeauteur van het voornoemde onderzoeksrapport.
- Drs. G. Jacobusse, statisticus bij TNO Kwaliteit van Leven en medeauteur van het voornoemde onderzoeksrapport.
- Dr. P. Sanders, psychometricus/directeur van het Research Center voor Examinering en Certificering.
- Prof. dr. G. Maris, senior medewerker/psychometricus bij het Psychometrisch Onderzoek-centrum van het Cito en hoogleraar Psychometrie aan de Universiteit van Amsterdam.
- Dr. T. Bechger, senior medewerker/psychometricus bij het Psychometrisch Onderzoek-centrum van het Cito.

- Dr. A. Béguin, hoofd van het Psychometrisch Onderzoek- en Kenniscentrum van het Cito.

Samenvatting van het overleg tussen TNO en het RCEC

- Op verschillende plaatsen in het RCEC rapport wordt gesproken over ‘de door TNO gehanteerde methodologie’. Deze terminologie zou volgens TNO de indruk kunnen wekken dat TNO een methode voorgesteld zou hebben waarmee een cesuur bepaald kan worden. Dat laatste is volgens TNO uitdrukkelijk niet het geval.
- Op basis van de aanvullende informatie van TNO, lijkt de tweede overweging op pagina 10 van het RCEC rapport over mogelijke verschillen in moeilijkheidsgraad van verschillende toetsen inderdaad niet terecht.
- Aan de conclusie van het RCEC dat op basis van de data geen verantwoordbare cesuur kan worden vastgesteld, is nagenoeg het gehele overleg besteed. De onderzoekers van het RCEC hebben nogmaals hun methodologische bezwaren tegen de werkwijze van TNO omstandig toegelicht. De onderzoekers van TNO wilden echter eerst zelf de data analyseren voordat zij voornoemde conclusie wilden bevestigen of ontkennen. Op basis van hun analyse van de data menen de onderzoekers van TNO te kunnen constateren dat de data wel geschikt zijn voor het vaststellen van een (verantwoordbare) cesuur.
- Door TNO werd met verwijzing naar hun rapport ten stelligste ontkend dat zij een nieuwe cesuur zouden hebben voorgesteld. Over deze bewering van TNO wil het RCEC het volgende opmerken. In het rapport staat inderdaad geen concreet voorstel voor een cesuur. In dit verband willen we echter opmerken dat in het TNO rapport op pagina 33 op onderzoeksvraag 2a, ‘Vinden we dezelfde zak-/slaaggrens als we de schalingsmethode herhalen met de twee verbeteringen die uit het vooronderzoek volgen’, twee antwoorden gegeven worden: ‘1. Met de verbeteringen vinden we dezelfde afstanden tussen de zak-/slaaggrenzen; 2. De gehele schaal is verankerd volgens een criterium dat ruim één vijfde van de schaal soepeler is dan beoogd.’ Dat het laatste antwoord tot een aanpassing van de cesuur heeft geleid, zal niemand, ook niet TNO, verbazen. Wel dient vermeld te worden dat TNO op pagina 33 nog wel een voorbehoud maakt. Nadat opgemerkt is dat ‘De verankering is nu gedaan ten opzichte van een item met een moeilijkheid die 4 punten lager ligt dan de gemiddelde moeilijkheid’, schrijft TNO vervolgens ’Het ligt volgens TNO meer voor de hand om te kiezen voor verankering ten opzichte van het beoogde criterium, namelijk de gemiddelde itemmoeilijkheid. De keuze is echter niet hard, er kan op basis van goede argumenten ook gekozen worden voor verankeringen ten opzichte van een makkelijker of moeilijker item.’

3. Conclusies en overwegingen

Het RCEC heeft de conclusie van het TNO onderzoek dat de cesuur van de TGN te soepel ingesteld is, bevestigd. Het RCEC en TNO blijven van mening verschillen of met de data die nu voorhanden zijn een methodologisch verantwoordbare cesuur vastgesteld kan worden.
Aanbevolen wordt dan ook nieuw onderzoek naar de verantwoordbaarheid van de TGN cesuur te laten uitvoeren.

Aanzetten voor het op korte termijn doen van nieuw onderzoek zijn voorhanden. Van de data die momenteel in het kader van de herbeoordelingen verzameld worden, moet nagegaan worden of zij voor het onderzoek naar de consistentie van beoordelingen van beoordelaars gebruikt kunnen worden. In een vervolgonderzoek kan dan gericht een vergelijking gemaakt worden tussen het oordeel van getrainde beoordelaars over die kandidaten en hun prestatie op de TGN waarbij gebruik gemaakt wordt van kandidaten die rond de cesuur <A1min/A1min presteren. Daarnaast worden momenteel data verzameld bij kandidaten die voor de TGN geslaagd zijn en na toelating opnieuw de TGN moeten afleggen. De data van deze hertoetsing zouden meer inzicht kunnen geven in de betrouwbaarheid van de TGN. Ten slotte zou nagegaan moeten worden of de resultaten van een eenvoudig uit te voeren procedure voor cesuurbepaling op het onderdeel 'woordenschat' van de TGN geëxtrapoleerd zouden kunnen worden naar de overige onderdelen van de TGN.
Wat betreft nieuw onderzoek dient er wel rekening mee te worden gehouden dat op basis van dat onderzoek wellicht ook geconcludeerd zou kunnen worden dat het met de TGN in zijn huidige opzet niet mogelijk is om de doelgroep van kandidaten voldoende betrouwbaar, dat wil zeggen met relatief geringe percentages ten onrechte gezakte en geslaagde kandidaten, te kunnen beoordelen als zijnde van een lager dan A1min niveau of van een A1min niveau. In dit verband kan de rekentoets die gebruikt wordt bij de selectie van studenten voor de Pabo als voorbeeld gesteld worden. Deze toets is optimaal afgestemd op de doelgroep, heeft een verantwoordbare cesuur terwijl de percentages ten onrechte gezakte en geslaagde kandidaten niet meer dan vier procent bedragen.

In afwachting van de uitkomsten van voornoemd onderzoek wordt voorgesteld de cesuur op de TGN aan te passen overeenkomstig het voorstel dat geformuleerd is in de brief van de minister van Wonen, Wijken en Integratie van 29 mei aan de Tweede Kamer. Bij de nieuwe, hogere cesuur zullen uiteraard meer kandidaten zakken dan bij de oorspronkelijke cesuur maar een verwacht zakpercentage van vijfentwintig procent kan vergeleken met de vele examentoetsen die bijvoorbeeld jaarlijks in het voortgezet onderwijs worden afgenomen zeker niet uitzonderlijk hoog genoemd worden. Met de hogere cesuur correspondeert een beheersingspercentage van iets meer dan dertig procent op onderscheiden onderdelen van de TGN. Voor het onderdeel 'woordenschat' betekent dit bijvoorbeeld dat nu 7 van de 22 gesproken antwoorden het goede antwoord moeten bevatten terwijl dat voorheen 4 goede antwoorden waren. Het nieuwe beheersingspercentage mag als alleszins redelijk beschouwd worden. Met inachtneming van de kanttekeningen die in het TNO rapport hierover op pagina 33 gemaakt worden, houdt aanpassing van de cesuur voor het A1min niveau ook aanpassing van de cesuren voor de opvolgende niveaus in.

Arnhem, 14 november 2007

Dr. P.F. Sanders
Directeur Research Center voor Examinering en Certificering (RCEC)