RCEC
Het expertisecentrum voor de kwaliteit van examinering

College voor Toetsen en Examens

Onderzoek naar de inhoudsvaliditeit, betrouwbaarheid en normering van het centraal examen mbo Nederlandse taal 3F 2018-2019

Arnold Brouwer, Piet Sanders,
Bernard Veldkamp

College voor Toetsen en Examens

Onderzoek naar de inhoudsvaliditeit, betrouwbaarheid en normering van het centraal examen mbo Nederlandse taal 3F 2018-2019

Arnold Brouwer, Piet Sanders,
Bernard Veldkamp

Apeldoorn, augustus 2020

RCEC

RCEC, *Research Center voor Examinering en Certificering*, is het expertisecentrum voor het borgen en bevorderen van de kwaliteit van examinering.

Als auditorganisatie beoordelen wij de kwaliteit van studietoetsen en examens. Wij zijn betrokken bij de beoordeling en certificering van grote Nederlandse examenstelsels. Zowel binnen het reguliere onderwijs ressorterend onder het ministerie van Onderwijs, Cultuur en Wetenschap, als voor het niet-reguliere onderwijs, zoals de door de ministeries ingestelde zelfstandige bestuursorganen (ZBO's).

Verder zijn wij een volledig onafhankelijke adviesorganisatie. Wij ontwikkelen geen eigen instrumenten, maar zijn er exclusief om instellingen in het regulier onderwijs en de beroepspraktijk te ondersteunen met integraal toetsdeskundig advies. Met praktijkgericht onderzoek, data-gedreven innovatie en psychometrische dienstverlening richten wij ons op de volledige toetscyclus. Van constructie en afname tot beoordeling en analyse.

De kennis en ervaring die wij in ons bestaan hebben opgedaan, en nog steeds opdoen, delen we via onze academie. Wij bieden verschillende vormen van opleiding, gericht op kwalitatieve examinering en toegepaste psychometrie. Dit doen we via open inschrijving, in-company maatwerkopleidingen en online via ons learning management systeem. Zo doen wij recht aan beslissingen door, over en voor talent.

RCEC B.V. • Villa de Regentes • Regentesselaan 5 • 7316 AA Apeldoorn
Postbus 71 • 8170 AB Vaassen
T +31 (0)55 - 303 31 24 • **I** www.rcec.nl • **E** info@rcec.nl

Dependance:
Universiteit Twente • Faculteit BMS – OMD • Gebouw Cubicus • Kamer B323
Drienerlolaan 5 • 7522 NB Enschede

Inhoudsopgave

1. Aanleiding blz. 1
   - 1.1. Aanleiding en probleemstelling 1
   - 1.2. Centrale onderzoeksvraag 1
   - 1.3. Aanvraag en proces 2

2. Onderzoekskader 3
   - 2.1. Onderzoekskader centrale examens mbo 3
   - 2.2. Inhoudsvaliditeit 3
   - 2.3. Betrouwbaarheid 5
   - 2.4. Normering 5

3. Onderzoek inhoudsvaliditeit, betrouwbaarheid en normering 7
   - 3.1. Onderzoek inhoudsvaliditeit van het centraal examen mbo Nederlandse taal 3F 2018-2019 7
   - 3.2. Onderzoek betrouwbaarheid van het centraal examen mbo Nederlandse taal 3F 2018-2019 9
   - 3.3. Onderzoek normering van het centraal examen mbo Nederlandse taal 3F 2018-2019 9

4. Onderzoeksresultaten 11
   - 4.1. Inhoudsvaliditeit van het centraal examen mbo Nederlandse taal 3F 2018-2019 11
   - 4.2. Domein definitie 11
     - 4.2.1.Inhoud van het examen 11
     - 4.2.2.Toetsspecificaties 12
   - 4.3. Domein representatie 13
   - 4.4. Domein relevantie 14
   - 4.5. Adequate toetsconstructieprocedure 15
   - 4.6. Betrouwbaarheid van het centraal examen mbo Nederlandse taal 3F 2018-2019 16
   - 4.7. Standaardbepaling en normhandhaving van het centraal examen mbo Nederlandse taal 3F 2018-2019 20

5. Conclusie 21
   - 5.1. Conclusies inhoudsvaliditeit 21
   - 5.2. Conclusies betrouwbaarheid 21
   - 5.3. Conclusies standaardbepaling en normhandhaving 22

6. Geraadpleegde literatuur 23

1. Inleiding

1.1. Aanleiding en probleemstelling

Het College voor Toetsen en Examens (CvTE) is een zelfstandig bestuursorgaan. Sinds 2009 is het de regisseur van centrale toetsing en examinering in ons land voor NT2, po, vo, staatsexamens vo en mbo. Krachtens de 'Wet College voor Toetsen en Examens' heeft het CvTE een regie-voerende verantwoordelijkheid. Het CvTE heeft als taak om namens de overheid de kwaliteit en het niveau van centrale toetsen en examens te waarborgen en te bevorderen dat scholen en instellingen in staat gesteld worden de afname van centrale toetsen en examens vlekkeloos te laten verlopen. Het CvTE adviseert en/of besluit over beleid ten aanzien van de uitvoering van centrale toetsen en examens, regelingen en alle andere zaken die voortvloeien uit de wettelijke taken van het CvTE.

Twee keer eerder heeft RCEC een onderzoek uitgevoerd naar de vraag of de procedures die het CvTE en Cito bij de borging van de inhoudsvaliditeit van de centrale examens voortgezet onderwijs hanteren, voldoen aan (inter)nationale kwaliteitsstandaarden. In het jaar 2017 deed RCEC onderzoek naar de inhoudsvaliditeit van de centrale examens biologie vmbo GL/TL 2016 en Duits havo 2016. In deze zelfde studie werd tevens onderzoek gedaan naar de afhandeling van onvolkomenheden bij de centrale examens. In het jaar 2018 deed RCEC onderzoek naar de inhoudsvaliditeit van de twee centrale examens Engels vwo 2017 en wiskunde vmbo GL/TL 2017.

Het CvTE heeft, vanuit zijn verantwoordelijkheid en zorgplicht, in navolging op de eerder uitgevoerde onderzoeken aan RCEC gevraagd om onderzoek te doen naar het centraal examen mbo Nederlandse taal 3F 2018-2019. In deze studie wordt naast de inhoudsvaliditeit ook de betrouwbaarheid en de normering (standaardbepaling en normhandhaving) van het betreffende examen, zijnde variant P2-5, onderzocht. Aanvullend wordt ook de betrouwbaarheid van de vier varianten: P2-5, P2-6, P4-8 en P4-9 van het betreffende examen onderzocht.

1.2. Centrale onderzoeksvraag

Het door RCEC uit te voeren onderzoek heeft als doel antwoord te geven op de volgende centrale onderzoeksvraag:

Leiden de procedures die het CvTE en Cito bij de ontwikkeling van de centrale examens mbo hanteren tot examenproducten die voldoen aan (inter)nationale kwaliteitsstandaarden met betrekking tot inhoudsvaliditeit, betrouwbaarheid en normering?

Ten behoeve van het onderzoek, is door RCEC een onderzoekskader ontwikkeld (zie hoofdstuk 2) bestaande uit: (1) de vier centrale aspecten van de inhoudsvaliditeit voor centrale examens: domein definitie, domein representatie, domein relevantie en adequate toetsconstructieprocedure, (2) de betrouwbaarheid, en (3) de normering (standaardbepaling en normhandhaving).

1.3. Aanvraag en proces

Op 20 mei 2020 heeft er een eerste gesprek plaatsgevonden tussen de heer Marcel Claessens, Sectormanager po en mbo en de heer Arnold Brouwer, directeur RCEC. Tijdens dit gesprek is de onderzoeksvraag besproken en toegelicht. Volgend op dit gesprek heeft RCEC op 27 mei 2020 een onderzoeksvoorstel aangeboden aan het CvTE. Op 3 juni 2020 heeft het CvTE ingestemd met dit voorstel. Voor de uitvoering van het onderzoeksvoorstel heeft RCEC een onderzoeksteam samengesteld. De leden van dit onderzoeksteam zijn de heer dr. Arnold Brouwer en de heer dr. Piet Sanders. Beide onderzoekers hebben de opdracht via het vier-ogen-principe uitgevoerd onder supervisie van de heer prof.dr.ir. Bernard Veldkamp. De bevindingen zijn in dit rapport dd. 31 augustus 2020 uiteengezet.

Namens RCEC,

Dr. Arnold J. Brouwer
Apeldoorn, 31 augustus 2020

2. Onderzoekskader

2.1. Onderzoekskader centrale examens mbo

Het onderzoekskader centrale examens mbo bestaat uit drie onderdelen. Het eerste onderdeel (paragraaf 2.2) onderzoekt de kwaliteit van de procedures voor het borgen van de inhoudsvaliditeit van een toets of een examen en is ontleend aan de Standards for Educational and Psychological Testing (AERA, 2014). De Standards for Educational and Psychological Testing (AERA, 2014) zijn een set van breed geaccepteerde standaarden die opgesteld zijn door de American Educational Research Association (AERA), de American Psychological Association (APA) en de National Council on Measurement in Education (NCME).

Het tweede en derde onderdeel van het onderzoekskader onderzoeken de betrouwbaarheid (paragraaf 2.3) en de normering (paragraaf 2.4) van het examen en zijn ontleend aan het RCEC beoordelingssysteem voor de kwaliteit van studietoetsen en (praktijk)examens (Sanders, Brouwer, Eggen, & Veldkamp, 2018). Het RCEC beoordelingssysteem onderzoekt de kwaliteit van een toets of (praktijk)examen aan de hand van zes criteria, zijnde het doel en gebruik van een toets of examen, het toets- en examenmateriaal, de representativiteit/validiteit, de betrouwbaarheid, de standaardbepaling en normhandhaving, en de afname en beveiliging. Gezamenlijk geven deze criteria een gedetailleerd beeld van de inhoudelijke, organisatorische en toetstechnische kwaliteitsaspecten van een toets of een examen.

De drie onderdelen van het onderzoekskader gezamenlijk geven een antwoord op de in paragraaf 1.2 geformuleerde centrale onderzoeksvraag.

2.2. Inhoudsvaliditeit

Het eerste hoofdstuk van de Standards for Educational and Psychological Testing (AERA, 2014) betreft de validiteit en begint met: ‘V*alidity refers to the degree to which evidence and theory support the interpretations for proposed uses of tests*.’ Bij het toepassen van deze definitie op het valideren van toetsscores wordt onderscheid gemaakt tussen de interpretatie van toetsscores (welke betekenis heeft een score?) en het gebruik van toetsscores (welke beslissingen worden op grond van de toetsscores genomen?). Hoe dit van toepassing is in het middelbaar beroepsonderwijs, kan worden toegelicht aan de hand van het centraal examen mbo Nederlandse taal 3F.

- De score die een student op het centraal examen mbo Nederlandse taal 3F behaalt, kan men *interpreteren* als een maat voor wat de vaardigheid Nederlandse taal wordt genoemd.
- Die behaalde score kan men vervolgens *gebruiken* om te beslissen of de student een voldoende of onvoldoende krijgt voor het centraal examen mbo Nederlandse taal 3F.

In navolging van voornoemde Standards onderscheidt en beschrijft Wools (2013) verschillende soorten bewijzen die gebruikt kunnen / moeten worden om de validiteit van toetsscores aan te tonen. Voor examens zijn vooral inhoudsbewijzen van belang. Bij inhoudsbewijzen gaat het volgens Wools ‘*om de keuzes die gemaakt worden ten aanzien van de onderwerpen of onderdelen uit de leerstof die in de toets opgenomen worden. Deze keuzes bepalen voor een groot deel of de inhoud van een toets*

representatief is voor het leerstofdomein of de vaardigheden waarover uitspraken gedaan moeten worden.' In het huidige onderzoek wordt in plaats van de term inhoudsbewijzen de 'historische' term inhoudsvaliditeit gebruikt en wordt deze gedefinieerd als: '*de mate waarin de inhoud van het examen overeenstemt met het doel van het examen*'. Dat (gebruiks)doel is bij de centrale examens het certificeren van studenten.

Voor het beoordelen van de inhoudsvaliditeit van examens wordt gebruik gemaakt van een onderzoekskader dat gebaseerd is op publicaties van Sireci (1998a, 1998b) en Sireci en Faulkner-Bond (2014). In lijn met deze studies, wordt in het onderzoekskader voor het evalueren van de inhoudsvaliditeit van een toets of examen een onderscheid gemaakt tussen:

1. domein definitie
2. domein representatie
3. domein relevantie
4. adequate toetsconstructieprocedure

Domein definitie verwijst naar hoe het 'construct', bijvoorbeeld de vaardigheid Nederlandse taal 3F, operationeel gedefinieerd wordt. Wat betreft de examens uit het middelbaar beroepsonderwijs komt die operationele definitie neer op het geven van (a) gedetailleerde beschrijvingen van de inhoud van het domein en de cognitieve vaardigheden waar een beroep op wordt gedaan, en (b) toetsspecificaties die zowel de specifieke inhoudscategorieën of inhoudsgebieden als de cognitieve niveaus benoemen. Evaluatie van de domein definitie houdt in dat de overeenstemming onderzocht wordt tussen de operationele definitie van het examen en de heersende opvattingen over het domein van vakdeskundigen uit het werkveld. Voor dit onderzoek worden veelal onafhankelijke vakdeskundigen ingezet bij de ontwikkeling en evaluatie van de toetsspecificaties. De mate waarin belangrijke aspecten van het construct of het curriculum niet voorkomen in de toetsspecificaties is een belangrijk criterium voor de evaluatie van de domein definitie.

Domein representatie verwijst naar de mate waarin een toets het domein zoals dat gedefinieerd is door de toetsspecificaties adequaat representeert en meet. Evaluatie van de domein representatie vereist de inzet van vakdeskundigen die de items/vragen/opdrachten bekijken en beoordelen. Het is de taak van de vakdeskundigen om te bepalen of de items in voldoende mate het beoogde domein representeren.

Domein relevantie betreft de mate waarin elk item van een examen relevant is voor het beoogde domein. Items die belangrijke aspecten van het domein meten, zouden hoge beoordelingen voor domein representatie moeten krijgen en items die minder belangrijke aspecten meten lage beoordelingen. Ook hier zou men aan vakdeskundigen moeten vragen de relevantie van items voor bepaalde toetsspecificaties te evalueren en deze beoordelingen zou men dan binnen elke inhoudscategorie kunnen aggregeren om de domein representatie te bepalen.

Adequate toetsconstructieprocedures zullen de inhoudsvaliditeit van examens bevorderen door bij alle onderscheiden stappen van het toetsconstructieproces kwaliteitscontrole te doen plaatsvinden. In de loop der jaren zijn er vele toetsconstructieprocedures voorgesteld die veelal grote overeenkomsten vertonen. Een van de bekendste toetsconstructieprocedures is het twaalf stappenplan dat beschreven staat in het eerste hoofdstuk van Downing en Haladyna (2006). In

navolging hierop introduceerde Veldkamp (2016) een tien stappenplan ten behoeve van de (centrale) examens zoals wij die in Nederland kennen. In het onderhavige onderzoek wordt nagegaan of de wijze waarop de centrale examens geconstrueerd worden, overeenkomt met de tien stappen uit deze toetsconstructieprocedure.

2.3. Betrouwbaarheid

Afhankelijk van het psychometrisch model dat bij de analyse van toetsscores gebruikt wordt, worden verschillende coëfficiënten berekend om de betrouwbaarheid van de toets of het examen mee aan te geven. In het voorgenomen onderzoek naar de betrouwbaarheid van het centraal examen mbo Nederlandse taal 3F 2018-2019 wordt ervan uitgegaan dat de toetsscores geanalyseerd zijn met de toets- en item analyse uit de klassieke testtheorie die 'klassieke' betrouwbaarheidscoëfficiënten oplevert.

Bij de beoordeling van de betrouwbaarheid gaat het om de vraag of de betrouwbaarheid voldoende is gezien de beslissingen die met het examen genomen worden. Voor het antwoord op deze vraag hanteert het RCEC standaarden zoals die bijvoorbeeld ook door de COTAN gehanteerd worden. Aangeraden wordt om behalve de betrouwbaarheidscoëfficiënt ook het percentage misclassificaties (ten onrechte geslaagd/gezakt) te vermelden. Deze maat geeft een meer begrijpelijke interpretatie van de betrouwbaarheid dan de betrouwbaarheidscoëfficiënt. Daarnaast kan het ook zinvol zijn om het effect van individuele items op de betrouwbaarheid van het examen vast te stellen. Met behulp van een item-/toetsinformatietabel waarbij de item-/toetsinformatie van de items bij bepaalde vaardigheidsscores, met name ook bij de cesuur, berekend wordt, kan dit nagegaan worden.

2.4. Normering

Het onderzoek naar de normering richt zich op de aspecten standaardbepaling en normhandhaving. Standaardbepaling is het bepalen van de regels om prestaties op examens in cijfers of waarderingen om te zetten. Een belangrijk onderdeel hierbij is het vaststellen van de cesuur, de grens tussen voldoende of onvoldoende.

Normeren kan worden onderscheiden in absoluut normeren en relatief normeren. Kenmerkend voor absoluut normeren is dat de norm of standaard voor afname van de toets wordt bepaald. De absolute norm is gebaseerd op een minimaal acceptabel beheersingsniveau. Bij relatief normeren wordt de norm na afname van de toets bepaald. De relatieve norm is gebaseerd op een onderlinge vergelijking van de toetsprestaties van de studenten.

Methoden voor standaardbepaling kunnen onderscheiden worden in methoden die gebaseerd zijn op de beoordeling van vragen/opgaven/opdrachten van een toets of examen, bijvoorbeeld de absolute cesuurmethode van Angoff, en methoden die gebaseerd zijn op de beoordeling van studenten die een examen maken, bijvoorbeeld de relatieve cesuurmethode van contrasterende groepen.

De experts met vakdeskundigheid die de standaarden vaststellen dienen kennis te hebben van het vakgebied en de eisen waaraan de studenten moeten voldoen. Verder is het noodzakelijk dat de

experts getraind zijn in het uitvoeren van de gebruikte standaardbepalingsmethode en het beoordelen van vragen.

Naast standaardbepaling gaat het ook om het handhaven van eenmaal bepaalde standaarden. De methoden om standaarden/normen te handhaven worden meestal met normhandhaving aangeduid. Idealiter worden aan studenten die verschillende varianten van hetzelfde examen doen dezelfde eisen gesteld. Dat betekent dat de varianten van het examen inhoudelijk gelijkwaardig moeten zijn en ook dat de ene variant niet moeilijker of gemakkelijker is dan de andere.

In het volgende hoofdstuk worden de drie onderdelen van het onderzoekskader – inhoudsvaliditeit, betrouwbaarheid en normering – toegepast op het centraal examen mbo Nederlandse taal 3F 2018-2019.

3. Onderzoek inhoudsvaliditeit, betrouwbaarheid en normering

3.1. Onderzoek inhoudsvaliditeit van het centraal examen mbo Nederlandse taal 3F 2018 - 2019

Voor het onderzoek naar de inhoudsvaliditeit van de centrale examens middelbaar beroepsonderwijs in Nederland spelen de syllabus en de constructieopdracht een cruciale rol. In deze paragraaf bespreken we deze twee instrumenten eerst in algemene zin. In de twee daaropvolgende paragrafen bespreken we de syllabus en de constructieopdracht van het centraal examen mbo Nederlandse taal 3F 2018-2019.

De minister van OCW stelt conform de Wet referentieniveaus Nederlandse taal en rekenen het referentiekader Nederlandse taal 3F vast. In dit referentiekader staat op hoofdlijnen wat de studenten moeten kennen en kunnen. Omdat het referentiekader niet genoeg houvast geeft aan examenconstructeurs, docenten en schoolboekenuitgevers, is het CvTE belast met de taak om de eisen voor het centraal examen mbo Nederlandse taal 3F te specificeren. Dat doet het CvTE door middel van een syllabus die periodiek geactualiseerd wordt op de momenten dat dit om inhoudelijke en/of redactionele redenen noodzakelijk wordt geacht.

Syllabus

Behalve een beschrijving van de exameneisen voor een centraal examen kan een syllabus verdere informatie over het centraal examen bevatten, bijvoorbeeld over een of meer van de volgende onderwerpen: specificaties van examenstof, begrippenlijsten, bekend veronderstelde voorkennis, bijzondere vormen van examinering (zoals computerexamens), voorbeeldopgaven, hulpmiddelen. In geval een nieuwe syllabus of een grondige herziening van een syllabus nodig is, wordt een syllabuscommissie ingesteld. Een syllabuscommissie bestaat uit een voorzitter en enkele leden die ondersteuning krijgen van het Kennisinstituut voor Taalontwikkeling (ITTA) en Cito. Specifieke referentiekaders en syllabi kan men vinden op www.examenbladmbo.nl.

Constructieopdracht

De wijze waarop de examenstof uit de syllabus wordt getoetst in het centraal examen is vastgelegd in de constructieopdracht van het CvTE. Deze constructieopdracht is uitgewerkt in een algemeen gedeelte dat bestaat uit een vijftal productspecificaties:

1. De verdeling van de vragen over de examenstof (toetsmatrijs);
2. De mate waarin het examen vragen met een reproductief en/of productief karakter moet bevatten;
3. De keuze van teksten en contextmateriaal;
4. De toe te passen vraagvormen en vaardigheidsvragen;
5. De regels voor de scoring.

De constructieopdracht bevat daarnaast een gedeelte dat de specifieke afspraken voor het studiejaar 2018-2019 beschrijft, waaronder (1) het aantal opgaven en het aantal varianten, (2) het te hanteren design, en (3) de overlap met andere examens.

Bij de vaststelling van vragen, teksten en fragmenten dient de preambule van CvTE in acht te worden genomen. In de preambule in bijlage 2 van de constructieopdracht centraal examen mbo Nederlandse

taal 3F wordt er uitgebreid op ingegaan dat examenopdrachten niet aanstootgevend mogen zijn voor bepaalde groepen in de samenleving waarbij 'bepaalde groepen' breed wordt opgevat en niet beperkt is tot religieuze stromingen.

De constructieopdracht voor ieder afzonderlijk vak wordt door de inhoudelijk opdrachtgever CvTE verstrekt aan de opdrachtnemer Cito. Het centraal examen mbo Nederlandse taal 3F bestaat uit het aantal varianten conform het vastgestelde design per afnameperiode (bijlage 4 van de constructieopdracht). Elke variant voldoet aan de toetsmatrijs (bijlage 1 van de constructieopdracht). De verschillende teksten in de varianten hebben zoveel mogelijk dezelfde kenmerken, zoals het aantal scorepunten en kennis en vaardigheden die getoetst worden. Daarnaast worden de tekstfragmenten getoetst op hun informatiewaarde, alvorens deze worden hergebruikt in een volgende examenvariant. Bij de oplevering van een variant van het centraal examen legt de opdrachtnemer verantwoording af aan de opdrachtgever over de gevolgde procedure en de gemaakte keuzes.

De vaststellingscommissie van het CvTE beoordeelt of het door Cito gemaakte concept-examen geschikt en passend is als centraal examen en stelt het examen, inclusief de scoringsvoorschriften voor de opgaven vast. Een vaststellingscommissie bestaat uit een voorzitter die meestal afkomstig is uit het middelbaar beroepsonderwijs en enkele docenten die lesgeven in examenklassen. Deze leden worden voorgedragen door de vakvereniging of de onderwijsbonden. Na afloop van de examenperiode evalueert het CvTE de examens met het scholenveld. De resultaten van deze evaluatie worden in een nieuwe examencyclus ingebracht.

Zoals beschreven in de vaststellingsprocedures en de werkwijze vaststellingscommissie uit bijlage 3 van de constructieopdracht worden er samengevat acht fasen onderscheiden:

1. Adviseren over de constructieopdracht. CvTE communiceert hierover met Cito.
2. Vaststellen van teksten en fragmenten. De vaststellingscommissie (VC) beoordeelt eerst of de teksten en programma's geschikt zijn, voordat de constructiegroep (CG) vragen maakt.
3. Vaststellen nieuwe opgaven. Voordat opgaven aan de VC worden aangeboden, heeft Cito de inhoudelijke en redactionele controle van de opgaven/sleutel/label uitgevoerd.
4. Vaststellen van ingevuld design. In deze fase worden de verschillende varianten vastgesteld, conform het geldende design.
5. Vaststellen van basisvarianten (preview exemplaar).
6. Vaststellen afnamevarianten door middel van het trekken van items uit de itembank.
7. Advisering bij de normering. Cito levert het technisch advies: een lijst met te bespreken items en een beargumenteerd voorstel voor het al dan niet neutraliseren van de opgave.
8. Evalueren van de informatiewaarde van de ingezette tekstfragmenten en de individuele items/opgaven.

Daarmee bestaat het constructie- en vaststellingsproces feitelijk uit twee onderdelen: (1) het construeren en vaststellen van tekstfragmenten en bijbehorende opgaven, en (2) het generen van nieuwe varianten van het examen door middel van het trekken van items uit de itembank.

3.2. Onderzoek betrouwbaarheid van het centraal examen mbo Nederlandse taal 3F 2018-2019

Om de betrouwbaarheid van de verschillende varianten te onderzoeken is gebruik gemaakt van de toets- en item analyses en de analyses met het One Parameter Logistic Model (OPLM), zoals die zijn uitgevoerd door Cito en waarvan de informatie is aangeleverd voor dit onderzoek. Daarbij is gekeken naar de moeilijkheidsgraad van de items, naar het onderscheidend vermogen, naar de betrouwbaarheid van de toetsvarianten, naar de op itemresponstheorie gebaseerde toets informatie functie en naar de standaardfout bij de cesuur. Dit laatste om ook inzicht te krijgen in de lokale betrouwbaarheid, dat wil zeggen de nauwkeurigheid van de onderscheiden vaardigheidsscores van in het bijzonder die van de zak/slaaggrens (cesuur).

De moeilijkheidsgraad is op verschillende manieren onderzocht. Er is gekeken naar de gemiddelde p-waarde, de hoogste p-waarde, de laagste p-waarde, het percentage items met een p-waarde lager dan .20, het percentage items met een p-waarde hoger dan .80 en naar het percentage items met een bij het gehanteerde itemtype passende optimale p-waarde tussen de .20 en de .80. In het RCEC beoordelingssysteem voor de kwaliteit van toetsen en examens (Sanders et al., 2018) wordt aangegeven dat bij een goed examen de gemiddelde p-waarde ligt rond de cesuur en dat het examen weinig items bevat met extreme p-waardes lager dan .20 of hoger dan .80.

Het onderscheidend vermogen is onderzocht met de gemiddelde rit-waarde, de laagste rit-waarde, de hoogste rit waarde, het aantal items met een rit-waarde lager dan 20 en het aantal items met een rit-waarde hoger dan 20. In het RCEC beoordelingssysteem wordt aangegeven dat bij een goed examen de rit-waarde gemiddeld 20 of hoger is.

In het onderzoek van de betrouwbaarheid staat de greatest lower bound (glb) centraal, vanwege het feit dat deze coëfficiënt de betrouwbaarheid schat aan de hand van clusters van items die parallel zijn. Dit past bij de heterogeniteit van het examen, waarin naast lezen ook luisteren en kijken worden getoetst. Daarnaast zijn de Cronbach's alpha en lambda gepresenteerd. Tevens zijn de slagingspercentages en de percentages misclassificaties op basis van de glb in kaart gebracht. In het RCEC beoordelingssysteem wordt voor de betrouwbaarheid van een high-stakes examen aangegeven dat deze .80 of hoger dient te zijn. Het geschatte percentage misclassificaties is daarbij gelijk aan 15% of lager.

De lokale betrouwbaarheid rond de cesuur is onderzocht door op basis van de item parameters de toets informatie functie te berekenen. Met deze functie is de standaardfout voor de verschillende waardes van de vaardigheidsparameter berekend en is in kaart gebracht wat de standaardfout is ter hoogte van de cesuur. Zo is af te leiden of het examen optimaal meet rond de cesuur.

3.3. Onderzoek normering van het centraal examen mbo Nederlandse taal 3F 2018-2019

Het onderzoek naar de normering van het centraal examen mbo Nederlandse taal 3F richt zich op de beide in paragraaf 2.4 genoemde aspecten van de normering, standaardbepaling en normhandhaving.

Bij het centraal examen mbo Nederlandse taal 3F 2018-2019 is er sprake van een absolute standaard/cesuur/norm, de zogenaamde referentiecesuur. Deze is in het jaar 2014 door CvTE

vastgesteld als onderdeel van het project Referentiesets taal (lezen) en rekenen. Het kunnen bepalen of de standaard correct bepaald is, vereist conform het RCEC beoordelingssysteem antwoord op drie vragen:

1. Is de gehanteerde standaardbepalingsmethode op de juiste wijze uitgevoerd?
2. Zijn de experts met vakdeskundigheid die de standaard bepalen naar behoren geselecteerd en getraind?
3. Is er voldoende overeenstemming tussen de experts met vakdeskundigheid?

Zoals beschreven in paragraaf 3.1 is er bij het centraal examen mbo Nederlandse taal 3F 2018-2019 sprake van meerdere varianten van hetzelfde examen. De referentiecesuur wordt daarbij toegepast middels equivalering van de vaardigheidsschaal. Het bepalen of de referentiecesuur op de juiste wijze gehandhaafd wordt, vereist conform het RCEC beoordelingssysteem antwoord op twee vragen:

1. Is de gekozen normhandhavingsmethode van voldoende kwaliteit?
2. Is de normhandhavingsmethode correct uitgevoerd?

In hoofdstuk 4 doen we verslag van een onderzoek naar de inhoudsvaliditeit, betrouwbaarheid en normering van het centraal examen mbo Nederlandse taal 3F.

4. Onderzoeksresultaten

4.1. Inhoudsvaliditeit van het centraal examen mbo Nederlandse taal 3F 2018-2019

In dit hoofdstuk wordt het onderzoekskader centrale examens mbo toegepast op het centraal examen mbo Nederlandse taal 3F 2018-2019. Het onderzoek start met de domein definitie (paragraaf 4.2), uitgewerkt in de inhoud van het examen (paragraaf 4.2.1) en de toetsspecificaties (paragraaf 4.2.2). Hierna volgen domein representatie (paragraaf 4.3), domein relevantie (paragraaf 4.4), en adequate toetsconstructieprocedures (paragraaf 4.5). De vier voornoemde onderdelen geven samen een beeld van de inhoudsvaliditeit van het centraal examen mbo Nederlandse taal 3F 2018-2019.

4.2. Domein definitie

De domein definitie voor het examen mbo Nederlandse taal 3F 2018-2019 betreft een gedetailleerde beschrijving van:

1. De inhoud van het examen en de cognitieve vaardigheden waar een beroep op wordt gedaan;
2. De toetsspecificaties die zowel de specifieke inhoudscategorieën of inhoudsgebieden als de cognitieve niveaus benoemen.

Onderstaande beschrijving van de inhoud van het examen en de toetsspecificaties zijn nagenoeg letterlijk ontleend aan de bespreking van de syllabus ‘Nederlandse taal, referentieniveau 3F’, versie 1 augustus 2018 en aan de constructieopdracht ‘Nederlandse taal mbo 3F, 2018-2019’, versie mei 2018.

4.2.1. Inhoud van het examen

Nederlandse taal is een generiek examenonderdeel, dat wil zeggen een onderdeel voor alle mbo-opleidingen. Het centraal examen mbo Nederlandse taal 3F omvat de onderdelen lezen van zakelijke teksten en luisteren. De focus op zakelijke teksten impliceert dat er bij het centraal examen geen fictionele, narratieve en literaire teksten aan bod komen. Luisteren omvat het beluisteren van audiofragmenten en het beluisteren en bekijken van audiovisuele fragmenten (zogenoemde kijkluisterfragmenten). In de praktijk is er voor gekozen om de verdeling van lees- en luisteropgaven vergelijkbaar te houden: beide ongeveer 50%. Het examen kent een tijdsduur van 120 minuten.

Bij de examinering van de vaardigheden lezen en luisteren wordt zoveel mogelijk gebruik gemaakt van authentieke teksten en luisterfragmenten. De thema’s van de examenteksten zijn niet strikt vastgelegd. Volgens het Referentiekader moeten studenten op niveau 3F in ieder geval kunnen lezen over onderwerpen uit de (beroeps)opleiding en onderwerpen van maatschappelijke aard.

Elk examen biedt meerdere teksten of fragmenten aan, waarbij gestreefd wordt naar diversiteit in onderwerpen. Daarmee kan beter beoordeeld worden of de student met teksten over verschillende onderwerpen uit de voeten kan, zoals op niveau 3F verwacht mag worden. Het 3F-niveau is typerend voor een vaardig taalgebruiker, die zich goed kan redden als zelfstandig beroepsbeoefenaar of als beginnend hbo-student.

Het Referentiekader van lezen beschrijft drie taken die gedefinieerd kunnen worden vanuit de doelen die de schrijver ermee nastreeft: informeren, instrueren dan wel overtuigen. De informatieve teksten in het centraal examen zijn vaak (achtergrond)artikelen uit kranten, tijdschriften en van nieuwswebsites. Ingewikkelde instructies en gebruiksaanwijzingen bij apparaten zijn vaak te (vak)specifiek van aard om in het examen te worden opgenomen. Teksten met een juridisch karakter, zoals de voorwaarden bij een contract of verzekering, vallen wel binnen de criteria voor complexiteit en themakeuze. Betogende teksten op niveau 3F komen meestal uit opiniebladen en (kwaliteits)kranten.

Het Referentiekader lezen maakt in de omschrijving van de kenmerken van de taakuitvoering bij lezen een onderscheid tussen de volgende deelvaardigheden: (1) techniek en woordenschat, (2) begrijpen, (3) interpreteren, (4) evalueren, en (5) samenvatten en opzoeken.

Voor de examinering van luisteren wordt audiomateriaal en audiovisueel materiaal geselecteerd uit radio- en televisieprogramma's en van internet. Tekstkenmerken betreffen: (1) lengte – op niveau 3F komen lange teksten van 30 minuten of meer voor, en (2) opbouw – op niveau 3F kan de informatiedichtheid hoog zijn.

Het Referentiekader luisteren benoemt specifieke taken die een taalgebruiker moet kunnen uitvoeren. Er worden drie luistersituaties benoemd: (1) luisteren naar instructies, (2) luisteren als lid van een live publiek, en (3) luisteren naar radio en televisie en naar gesproken tekst op internet.

Het Referentiekader luisteren maakt in de omschrijving van de kenmerken van de taakuitvoering onderscheid tussen verschillende deelvaardigheden: (1) begrijpen, (2) interpreteren, (3) evalueren, en (4) samenvatten.

4.2.2. Toetsspecificaties

In de toetsmatrijs van het examen mbo Nederlandse taal 3F 2018-2019, zoals opgenomen in bijlage 1 van de constructieopdracht, komen de volgende toetsspecificaties aan de orde:

- Lengte van de scoreschaal. Het door de student maximaal te behalen punten bedraagt ongeveer 60 punten (met een marge van 5%);
- Vraagvormen. Er is sprake van twee vraagvormen: (1) meerkeuzevraag met maximaal 4 antwoorden waarvan slechts 1 antwoord correct is, en (2) matrixvraag met bijvoorbeeld een aantal juist/onjuist-beweringen;
- Scoring van de vragen. Bij een meerkeuzevraag met maximaal 4 mogelijkheden wordt 1 punt toegekend wanneer de vraag goed beantwoord wordt en bij een matrixvraag wordt 1 punt toegekend wanneer alle betreffende juist/onjuist beweringen correct zijn beantwoord;
- Normering. Bij de normering van het examen worden scores omgezet in cijfers waarbij rekening wordt gehouden met de verschillen in moeilijkheid van de verschillende examenvarianten;
- Aantal teksten en bronnen. In totaal zijn er bij 3F zes teksten opgenomen per variant. Het betreft bij lezen drie typen tekst: informatief, instructief en betogend of hybride/betogend. Bij luisteren betreft het twee typen tekst: informatief (2x) en overig.

Naast voornoemde specificaties zijn er nog de volgende aandachtspunten:

- Authentieke teksten. Er wordt zoveel mogelijk gewerkt met authentieke teksten. Dat wil zeggen dat teksten niet speciaal voor examens gemaakt worden, maar dat bestaande teksten of (kijk)luisterfragmenten worden geselecteerd uit kranten en tijdschriften en van televisie, radio of internet.
- Bronnen. Vanwege de geheimhouding worden bronteksten voorzien van een nieuwe titel, met uitzondering van de betogende leesteksten. Ook de naam van bijvoorbeeld een televisieprogramma moet worden vermeden.
- Beroep op rekenvaardigheden. In het examen mag geen beroep gedaan worden op de rekenvaardigheden van de student.

4.3. Domein representatie

Domein representatie betreft de mate waarin een examen het domein zoals dat gedefinieerd is door de toetsspecificaties adequaat representeert en meet. Het behoort tot de expertise van vakdeskundigen om te bepalen of de opgaven in voldoende mate het beoogde domein representeren. In paragraaf 3.1 is beschreven hoe de inhoudelijk opdrachtgever CvTE voor ieder afzonderlijk vak een constructieopdracht verstrekt aan de opdrachtnemer Cito.

Hierna bespreken we het design van vier varianten van het examen mbo Nederlandse taal 3F 2018-2019, zijnde variant P2-5 en P2-6 uit afnameperiode 2 (29 oktober t/m 16 december 2018) en variant P4-8 en P4-9 uit afnameperiode 4 (4 maart t/m 21 april 2019), zoals weergegeven in tabel 1.

Tabel 1.

Design van vier varianten van het examen mbo Nederlandse taal 3F 2018-2019

Variant P2-5:	Variant P2-6:
- 5 vragen over leestekst: 'Polisvoorwaarden Woonverzekering' - 12 vragen over leestekst: 'Op het menu van de toekomst staat in ieder geval minder zuivel' - 9 vragen over leestekst: 'Help een land, strooi met geld' - 4 luisterfragmenten en 10 vragen over: 'De hebberigheid van mensen' - 4 kijkfragmenten en 10 vragen over: 'Burgerinspraak' - 4 kijkfragmenten en 10 vragen over 'Gevaarlijke planten en dieren'	- 12 vragen over leestekst: 'Een huis vol sensoren' - 5 vragen over leestekst: 'Voorwaarden lidmaatschap sportschool' - 9 vragen over leestekst: 'Babyboomers botsen met starters' - 4 kijkfragmenten en 10 vragen over: 'Herinrichting' - 3 luisterfragmenten en 10 vragen over: 'Gehoorschade voorkomen' - 5 kijkfragmenten en 10 vragen over: 'Fotonica'

Variant P4-8:

- 9 vragen over leestekst: 'Babyboomers botsen met starters'
- 12 vragen over leestekst: 'Op het menu van de toekomst staat in ieder geval minder zuivel'
- 5 vragen over leestekst: 'Voorwaarden Recreatiebedrijven'
- 3 luisterfragmenten en 9 vragen over: 'Gehoorschade voorkomen'
- 4 kijkfragmenten en 10 vragen over: 'Burgerinspraak'
- 4 kijkfragmenten en 10 vragen over 'Nieuwe medicijnen'

Variant P4-9:

- 5 vragen over leestekst: 'OV-chipkaart'
- 9 vragen over leestekst: 'Hoe seksistisch is het brein?'
- 11 vragen over leestekst: 'Een huis voor sensoren'
- 4 kijkfragmenten en 10 vragen over: 'Herinrichting'
- 4 luisterfragmenten en 10 vragen over: 'Waterzuivering'
- 5 kijkfragmenten en 9 vragen over: 'De BV ik'

De inhoudelijke indeling van de examenvarianten is in lijn met de gewenste verdeling van lees- en luisteropgaven van beide ongeveer 50% en is conform de toetsmatrijs uit bijlage 1 van de constructieopdracht en conform het generiek design uit bijlage 4 van de constructieopdracht.

4.4. Domein relevantie

Domein relevantie betreft de mate waarin elke opgave van een examen relevant is voor het beoogde domein. Ook om de relevantie van opgaven te bepalen, is de inzet van vakdeskundigen (docenten Nederlands binnen het mbo) nodig. Bij de constructie van alle examens, dus ook van het examen mbo Nederlandse taal 3F 2018-2019, zijn altijd twee groepen toets- en vakdeskundigen betrokken: leden van de constructiegroep en leden van de vaststellingscommissie.

Bij de constructie van opgaven en correctievoorschriften zijn docenten als constructiegroeplid betrokken. Deze docenten selecteren de contexten en leveren de vragen daarbij aan bij de Cito-toetsdeskundige. Een constructiegroeplid is altijd een docent met ervaring in examenklassen. Per examen wordt in de regel door een drietal constructiegroepleden materiaal aangeleverd. De Cito-toetsdeskundige stelt het examen samen en bespreekt het concept met de CvTE-vaststellingscommissie. De vaststellingscommissie is verantwoordelijk voor de vaststelling van de opgaven en het scoringsvoorschrift.

Voor de vaststellingsprocedure van het examen mbo Nederlandse taal 3F 2018-2019 wordt verwezen naar paragraaf 3.1 en de in bijlage 3 van de constructieopdracht beschreven vaststellingsprocedures en werkwijze vaststellingscommissie.

In paragraaf 3.1 is de samenstelling en vakdeskundigheid van de vaststellingscommissie besproken. Van de constructiegroepleden en de vaststellingscommissie wordt aangenomen dat zij in staat zijn de relevantie van de opgaven voor het meten van de taalvaardigheid te kunnen beoordelen.

4.5. Adequate toetsconstructieprocedure

Het gebruik van een adequate toetsconstructieprocedure zal de inhoudsvaliditeit van alle examens in het algemeen ten goede komen. Zoals beschreven in paragraaf 2.1., wordt aan de hand van het tien stappenplan van Veldkamp (2016) de constructieprocedure die bij het examen mbo Nederlandse taal 3F 2018-2019 gehanteerd wordt tegen het licht gehouden. De toetsconstructieprocedure van Veldkamp (2016) onderscheidt de volgende tien stappen:

1. Maak een toetsplan
2. Definieer de inhoud
3. Stel de toetsmatrijs op
4. Ontwikkel de items
5. Stel de toets samen
6. Neem de toets af
7. Scoor de antwoorden
8. Bepaal de zak/slaaggrens
9. Rapporteer de scores
10. Documenteer de voorgaande stappen

Hieronder wordt nagegaan of deze stappen ook bij de constructie van het centraal examen mbo Nederlandse taal 3F 2018-2019 gehanteerd worden.

Deelstap	Beoordeling
Ad 1. Maak een toetsplan Bij het examen mbo Nederlandse taal 3F 2018-2019 is het duidelijk welk 'construct' gemeten wordt, namelijk de vaardigheid lezen-luisteren op niveau 3F. Het examen is bedoeld om studenten te certificeren.	✓
Ad 2. Definieer de inhoud Het definiëren van de inhoud is beschreven in paragraaf 4.2.1 van de domein definitie.	✓
Ad 3.Stel de toetsmatrijs op Voor deze stap zie paragraaf 4.2.2 waar de toetsspecificaties vermeld worden.	✓
Ad 4. Ontwikkel de items Deze stap betreft de constructieopdracht op basis waarvan de opdrachtnemer Cito de opgaven voor een concept-examen construeert.	✓

Ad 5. Stel de toets samen De vaststellingscommissie van het CvTE stelt het examen en de correctievoorschriften vast.	✓
Ad 6. Neem de toets af Het examen mbo Nederlandse taal 3F 2018-2019 wordt onder regie van het CvTE op scholen afgenomen.	✓
Ad 7. Scoor de antwoorden Het examen wordt digitaal afgenomen en bestaat enkel uit geautomatiseerd scoorbare vragen.	✓
Ad 8. Bepaal de zak/slaaggrens In het jaar 2014 is met behulp van de Extended Angoff methode door een panel van 20 vakdeskundigen de referentiecesuur taal 3F bepaald. Met behulp van equivalering wordt voor elke variant van het examen mbo Nederlandse taal 3F uit de periode 2018-2019 de bijpassende vaardigheidsscore vastgesteld. Zo kunnen de verschillen in zwaarte binnen de verschillende varianten gecompenseerd worden.	✓
Ad 9. Rapporteer de scores De studenten ontvangen korte tijd na het afleggen van het examen hun cijfer. Het cijfer wordt vastgesteld door het CvTE.	✓
Ad 10. Documenteer de voorgaande stappen Voorgaande stappen zijn in verschillende documenten beschreven en veelal digitaal beschikbaar via de websites van CvTE en Cito.	✓

4.6. Betrouwbaarheid van het centraal examen mbo Nederlandse taal 3F 2018-2019

Van vier varianten van het centraal examen mbo Nederlandse taal 3F 2018-2019 is een betrouwbaarheidsanalyse uitgevoerd: de varianten P2-5 en P2-6 uit afnameperiode 2 (29 oktober t/m 16 december 2018) en de varianten P4-8 en P4-9 uit afnameperiode 4 (4 maart t/m 21 april 2019).

Tabel 2 geeft een samenvatting van de door Cito uitgevoerde toets- en item analyses op basis van de klassieke testtheorie. Per variant is het totaal aantal studenten (N) en de verdeling van de items over

de verschillende toetsonderdelen gepresenteerd. Vervolgens is de analyse van de moeilijkheidsgraad van de items, uitgedrukt in de p-waarde, gepresenteerd. Er is berekend hoeveel procent van de items een bij het gekozen itemtype passende optimale p-waarde heeft tussen de waarde 20 en 80 op een schaal van 0 - 100. Aanvullend is de analyse van het onderscheidend vermogen (discriminatie-index) van de items, uitgedrukt in de rit-waarde, gepresenteerd. De rit-waarde van een item is een indicatie voor de mate waarin het item onderscheid maakt tussen studenten met hoge toetsscores en studenten met lage toetsscores. Na standaardisering van de rit-waardes met behulp van de Fisher-Z transformatiefunctie is de gemiddelde rit-waarde per variant bepaald en is berekend hoeveel procent van de items een voldoende rit-waarde van 20 of hoger heeft op een schaal van -100 tot +100. De toetsbetrouwbaarheid is gepresenteerd in een drietal coëfficiënten, de greatest lower bound (glb), de Cronbach's alpha en de lambda. Met behulp van de verdeling van de ruwe somscores over de studenten en de omzettingstabellen CE per variant zijn de slagingspercentages berekend. Van daaruit is het percentage niet-consistente beslissingen (de zogenaamde misclassificaties) geschat als functie van het percentage gezakte studenten en de toetsbetrouwbaarheid, uitgedrukt in de glb (Eggen & Sanders, 1993). Tevens is een afleidersanalyse uitgevoerd. Hieruit volgt dat in alle vier de geëvalueerde examenvarianten geen sprake is van items met één of meer niet gekozen afleiders c.q. antwoordalternatieven.

Tabel 2.

Resultaten van de toets- en itemanalyse van de vier examenvarianten

variant	P2-5		P2-6		P4-8		P4-9	
# items								
totaal	56		56		55		54	
lezen instructief	5		5		5		5	
lezen informatief	12		12		12		11	
lezen betogend	9		9		9		9	
luisteren	10		10		9		10	
kijken deel 1	10		10		10		10	
kijken deel 2	10		10		10		9	
moeilijkheidsgraad								
gem. p-waarde	65,59		62,04		66,25		61,65	
laagste p-waarde	15		5		7		15	
hoogste p-waarde	96		95		95		97	
# p-waarde < 20	1	1,79%	3	5,36%	1	1,79%	1	1,79%
# p-waarde > 80	13	23,21%	12	21,43%	11	19,64%	13	23,21%
# p-waarde [20 - 80]	42	75,00%	41	73,21%	43	76,79%	40	71,43%

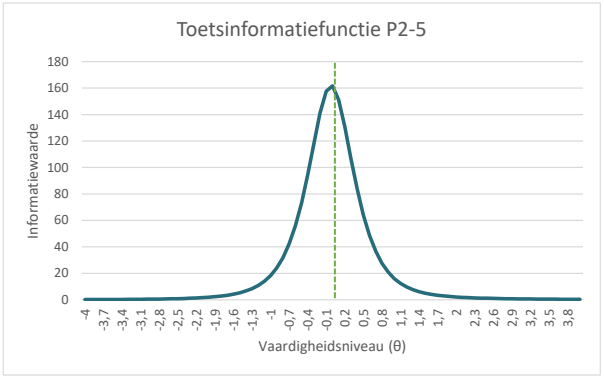
discriminatie-index							
gem. rit-waarde*	16,31		13,62		17,32		16,47	
laagste rit-waarde	-7		-4		-4		1	
hoogste rit-waarde	37		33		32		30	
# rit-waarde < 20	35	62,50%	43	76,79%	33	58,93%	38	67,86%
# rit-waarde >= 20	21	37,50%	13	23,21%	22	39,29%	16	28,57%
betrouwbaarheid								
glb	0,87		0,86		0,79		0,78	
alpha	0,68		0,62		0,69		0,68	
lambda	0,70		0,64		0,70		0,68	
slagings%**	60,21%		45,08%		66,50%		56,80%	
%misclassificaties***	[23 - 24%]		[30 - 31%]		[22 - 23%]		[24 - 25%]	

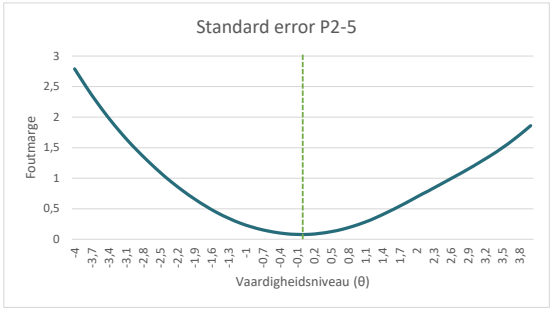
* De gemiddelde rit-waarde is bepaald na toepassing van de Fisher-Z transformation

** Slagingspercentages op basis van de ruwe scores en de omzettingstabellen CE per variant

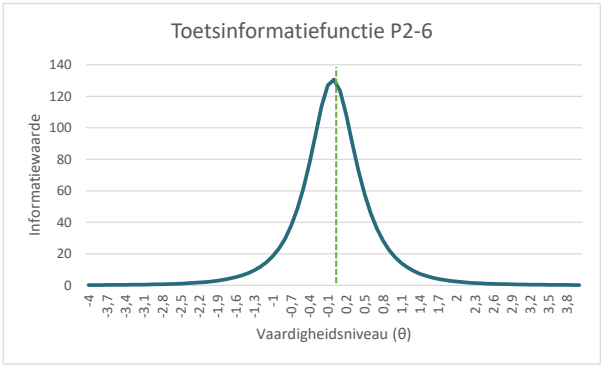
*** Percentage niet-consistente beslissingen als functie van percentage gezakten en betrouwbaarheid (Tabel 3.13 Psychometrie in de Praktijk, Eggen & Sanders, 1993)

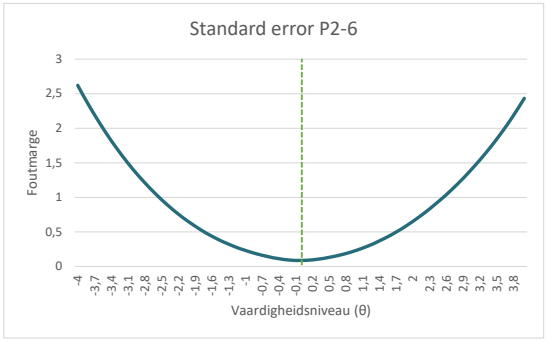
Figuur 1 visualiseert de toetsinformatiefunctie en de standaardfout van de vier varianten, op basis van de met het One Paramater Logistic Model (OPLM) gekalibreerde items. Per item zijn de populatie onafhankelijke moeilijkheidsgraad, uitgedrukt in de geschatte b-parameter, en het onderscheidend vermogen, binnen OPLM uitgedrukt in de geïmputeerde a-parameter, bepaald. De beide itemparameters van de twee varianten uit afnameperiode 4 zijn daarbij gekalibreerd op basis van de totaalkalibratie uit afnameperiode 2. Aan onderstaande grafieken is met de groen gestreepte verticale lijn de referentiecesuur, uitgedrukt in de theta-waarde 0,133 als omzetting van het rapportcijfer 5,5 toegevoegd. Opgemerkt wordt dat, conform de omzettingstabellen CvTE, in variant P2-5 vraag 25 is geneutraliseerd en in variant P2-6 de vragen 12, 45 en 50 zijn geneutraliseerd.



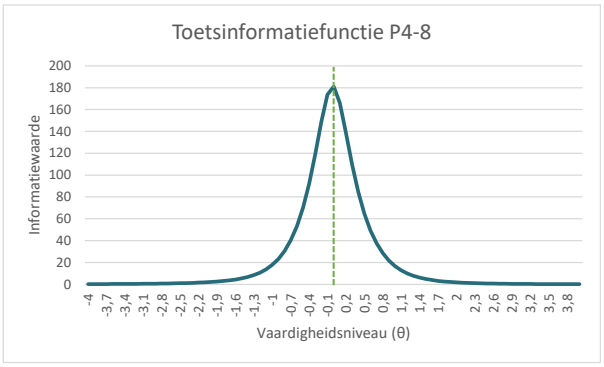


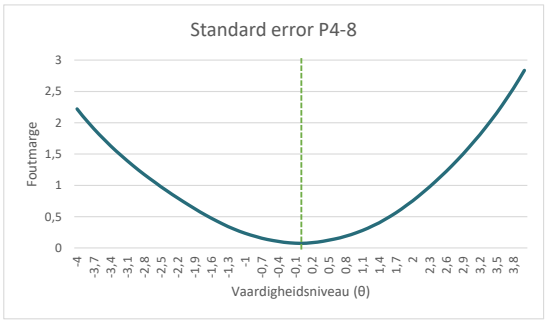
Informatiewaarde / foutmarge P2-5 bij cesuur 0,133: 151,20 / 0,08



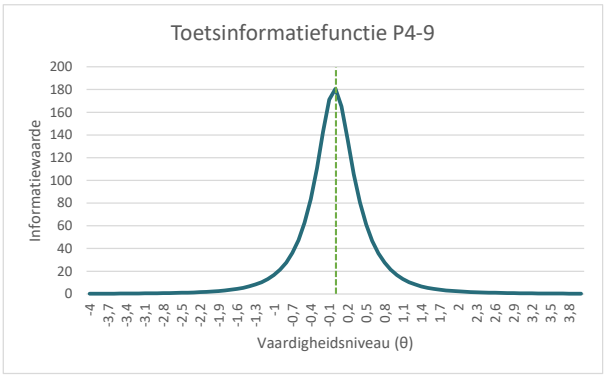


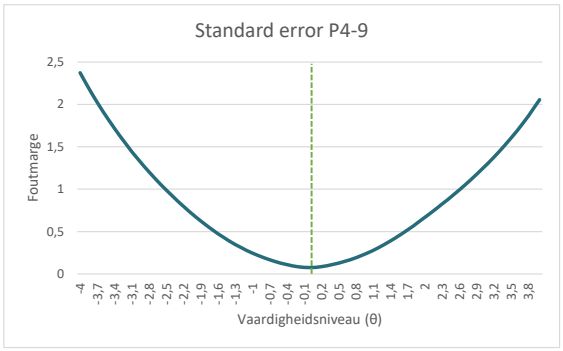
Informatiewaarde / foutmarge P2-6 bij cesuur 0,133: 123,39 / 0,09





Informatiewaarde / foutmarge P4-8 bij cesuur 0,133: 165,92 / 0,08





Informatiewaarde / foutmarge P4-9 bij cesuur 0,133: 165,16 / 0,08

Figuur 1. De toetsinformatiefunctie en standaardfout versus de cesuur van de vier examenvarianten

Uit de verschillende analyses blijkt dat de p-waardes goed verdeeld zijn en nagenoeg optimaal zijn voor de gehanteerde itemtypes. Het onderscheidend vermogen is relatief laag, wat kan duiden op een zekere mate van homogeniteit in de doelgroep en/of heterogeniteit in de onderliggende toetsdomeinen, bijvoorbeeld als gevolg van het combineren van lezen en luisteren/kijken in het examen. De toetsbetrouwbaarheid van de verschillende varianten is conform de glb goed en conform de Cronbach's alpha en lamda matig. Het geschatte percentage misclassificaties op basis van de slagingspercentages en de glb is acceptabel te noemen. De toetsinformatiefunctie en de standaardfout bij de cesuur laten zien dat de lokale toetsbetrouwbaarheid rond de cesuur hoog is.

4.7. Standaardbepaling en normhandhaving van het centraal examen mbo Nederlandse taal 3F 2018-2019

In het najaar 2013 heeft het CvTE het expertpanel 3F, bestaande uit zes deskundigen met vakdeskundigheid, ingezet voor de inhoudsvalidering van de referentieset taal 3F. Aanvullend is door 20 experts met vakdeskundigheid, verdeeld over vier panels, met behulp van de Extended Angoff methode de bijbehorende standaard (referentiecesuur) bepaald. Daarmee wordt aangenomen dat de gehanteerde standaardbepalingsmethode op de juiste wijze is uitgevoerd, dat de betrokken experts met vakdeskundigheid naar behoren zijn geselecteerd en getraind, en dat er voldoende overeenstemming is tussen de experts over de vastgestelde standaard.

Doordat in elke variant van het examen de opgaven uit de referentiesets samen met de reguliere opgaven worden afgenomen is het mogelijk om alle opgaven op één vaardigheidsschaal te plaatsen (door middel van IRT). De standaard op de referentieset wordt daarbij omgezet in een cesuur op de vaardigheidsschaal. Deze standaard wordt vervolgens gebruikt voor de normering van toetsen en examens door middel van de reguliere procedure ‘normering met een vaardigheidsschaal’. Bij het centraal examen mbo Nederlandse taal 3F is er sprake van een eerste afnameperiode van één dag, gevolgd door vier opeenvolgende afnameperiodes van elk zeven weken. Per afnameperiode worden meerdere basisvarianten van het examen geconstrueerd conform het generiek design uit bijlage 4 van de constructieopdracht. De verschillende teksten in de varianten hebben zoveel mogelijk dezelfde toetstechnische kenmerken. Daarnaast worden de tekstfragmenten getoetst op hun informatiewaarde, alvorens deze worden hergebruikt in een nieuwe examenvariant.

Voor de normhandhaving wordt het design van de basisvarianten uit de afnameperiodes 2 (29 oktober t/m 16 december 2018) en 4 (4 maart t/m 21 april 2019) verbonden. Hetzelfde procedé geldt voor de afnameperiodes 3 (7 januari t/m 24 februari 2019) en 5 (13 mei t/m 30 juni 2019). Na de eerste twee weken van afnameperiode 2 vindt een eerste normeringsronde plaats. De vaststellingscommissie bestudeert daarbij alle opgaven en bereidt de normeringsvergadering voor. Op basis van de resultaten uit de toets- en item analyse kan worden besloten om specifieke items te neutraliseren en zo buiten de normering te houden. Hierna worden de examenresultaten uit week 3 en 4 van afnameperiode 2 verzameld, waarmee er voldoende data is om alle items te kalibreren met behulp van het itemresponsmodel OPLM. Derhalve is na de eerste vier weken van afnameperiode 2 de definitieve normering bepaald. Deze definitieve norm geldt eveneens voor de basisvarianten uit afnameperiode 4, welke inhoudelijk en toetstechnisch zijn gelinkt met de basisvarianten uit afnameperiode 2. Een en ander conform de toetsmatrijs (bijlage 1 uit de constructieopdracht) en conform het generiek design (bijlage 4 uit de constructieopdracht). De overlap in tekstfragmenten en bijbehorende items, zoals ook weergegeven in tabel 1, draagt eraan bij dat met behulp van het itemresponsmodel OPLM de scores van de verschillende basisvarianten op één en dezelfde schaal afgebeeld kunnen worden.

Daarmee worden aan studenten die het examen doen dezelfde eisen gesteld, onafhankelijk van de aangeboden basisvariant. Dit betekent dat de basisvarianten van het examen inhoudelijk gelijkwaardig zijn en van eenzelfde moeilijkheidsgraad zijn, wat tevens wordt bevestigd door de inhoudelijke overlap tussen de varianten uit tabel 1 en door de vergelijkbare gemiddelde p-waarde van de vier onderzochte varianten uit tabel 2. Daarmee wordt aangenomen dat de gekozen normhandhavingsmethode van voldoende kwaliteit is en correct is uitgevoerd.

5. Conclusie

5.1. Conclusies inhoudsvaliditeit

In de paragrafen 4.1 tot en met 4.5 zijn voor het centraal examen mbo Nederlandse taal 3F 2018-2019 de vier aspecten van het onderdeel inhoudsvaliditeit van het onderzoekskader centrale examens mbo onderzocht. Op basis van de analyse van de vier aspecten kan geconcludeerd worden dat de procedures voor het realiseren van de inhoudsvaliditeit van het voornoemde examen voldoen aan de eisen die daaraan volgens nationale en internationale richtlijnen gesteld worden.

Voor wat betreft het onderzochte examen is er sprake van een domein definitie die genoegzaam als een volledige beschrijving van beide aspecten (zie paragraaf 2.2) aangemerkt kan worden. De gevolgde procedures resulteren in voldoende informatief met betrekking tot de domein representatie. Het proces van totstandkoming van de examens is beschreven in paragraaf 3.1 en in de daarbij genoemde vaststellingsprocedures en de werkwijze vaststellingscommissie. Dit proces wordt als zorgvuldig gekwalificeerd. Het examen voldoet tevens aan de eisen die aan het aspect domein relevantie worden gesteld. De wijze van construeren van het examen volgt het tien stappenplan (Veldkamp, 2016) voor een adequate toetsconstructieprocedure.

5.2. Conclusies betrouwbaarheid

De betrouwbaarheidsanalyse is uitgevoerd op de varianten P2-5 en P2-6 uit afnameperiode 2 (29 oktober t/m 16 december 2018) en de varianten P4-8 en P4-9 uit afnameperiode 4 (4 maart t/m 21 april 2019).

Uit de analyse volgt dat de moeilijkheidsgraad van de vier varianten, uitgedrukt in de p-waarde, onderling vergelijkbaar is en statistisch optimaal is voor de gehanteerde itemtypes. Het percentage p-waardes lager dan .20 of hoger dan .80 betreft voornamelijk items die om inhoudelijke redenen zijn opgenomen in het examen om zo het volledige domein te kunnen examineren.

Het onderscheidend vermogen is relatief laag, wat kan duiden op een zekere mate van homogeniteit in de doelgroep en/of heterogeniteit in de onderliggende toetsdomeinen als gevolg van het combineren van lezen en luisteren/kijken in het examen.

De betrouwbaarheid is geschat met behulp van de greatest lower bound (glb), wat past bij de heterogeniteit van het examen waarin naast lezen ook luisteren en kijken worden getoetst. Daarnaast zijn de Cronbach's alpha en lambda gepresenteerd. De betrouwbaarheid, uitgedrukt in de glb, en het daarbij geschatte percentage misclassificaties worden gekwalificeerd als goed. De lokale toetsbetrouwbaarheid rond de cesuur is voor de vier varianten hoog. Daarmee voldoet het onderzochte examen aan de criteria voor betrouwbaarheid zoals gesteld in het RCEC beoordelingssysteem (Sanders et al., 2018).

5.3. Conclusies standaardbepaling en normhandhaving

Er kan worden aangenomen dat de gehanteerde standaardbepalingsmethode op de juiste wijze is uitgevoerd, dat de betrokken experts met vakdeskundigheid naar behoren zijn geselecteerd en getraind, en dat er voldoende overeenstemming is tussen de experts over de vastgestelde standaard.

Uit de normhandhavingsprocedure volgt dat de basisvarianten van het examen inhoudelijk gelijkwaardig zijn en van eenzelfde moeilijkheidsgraad zijn. Dit wordt bevestigd door de inhoudelijke overlap tussen de varianten uit tabel 1 en door de vergelijkbare gemiddelde p-waarde van de vier onderzochte varianten uit tabel 2. Daarmee worden aan studenten die het examen doen dezelfde eisen gesteld, onafhankelijk van de aangeboden basisvariant en kan er worden aangenomen dat de gekozen normhandhavingsmethode van voldoende kwaliteit is en correct is uitgevoerd.

Geconcludeerd kan worden dat de procedures die het CvTE en Cito hanteren bij de ontwikkeling van de centrale examens mbo leiden tot examenproducten die voldoen aan (inter)nationale kwaliteitsstandaarden met betrekking tot inhoudsvaliditeit, betrouwbaarheid en normering.

6. Geraadpleegde literatuur

AERA (2014). *Standards for educational and psychological testing. Joint Committee on Standards for Educational and Psychological Testing.* Washington D.C.: American Educational Research Association.

Downing, S.M., & Haladyna, T.M. (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

Dekker, J., Sanders, P.F. (2008). *De kwaliteit van beoordeling in de praktijk*. Ede: Kenniscentrum handel.

Eggen, T.J.H.M., & Sanders, P.F. (1993). *Psychometrie in de praktijk*. Arnhem: Cito.

Gulikers, J.T.M., Bastiaens, T.J., & Kirschner, P. (2005). The Five-Dimensional Framework for Authentic Assessment. *Educational Technology Research and Development,* 52(3), 67-86.

Kuhlemeier, H., Til, A. van, & Sanders, P.F. (Red.) (2012). *Toetsen op School: Voortgezet onderwijs*. Arnhem: Cito. Geraadpleegd 10 juni 2020 van: https://www.rcec.nl/publicaties.

Sanders, P.F., Brouwer, A.J., Eggen, T.J.H.M, & Veldkamp, B.P. (2018). *RCEC beoordelingssysteem voor de kwaliteit van studietoetsen en (praktijk)examens*. Apeldoorn: RCEC.

Sanders, P.F. (Red.) (2013). *Toetsen op School*. Arnhem: Cito. Geraadpleegd 10 juni 2020 van: https://www.rcec.nl/publicaties.

Sanders, P.F., Brouwer A.J., & Veldkamp, B.P. (2017). *Onderzoek naar de inhoudsvaliditeit van de centrale examens en de afhandeling van onvolkomenheden bij de centrale examens*. Enschede: RCEC.

Sanders, P.F., Brouwer, A.J., & Veldkamp, B.P. (2018). *Onderzoek naar de inhoudsvaliditeit van een tweetal centrale examens voortgezet onderwijs 2017*. Enschede: RCEC.

Sireci, S.G. (1998a). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299-321.

Sireci, S.G. (1998b). The construct of content validity. *Social indicators Research*, 45, 83 – 117.

Sireci, S.G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, 26, 100 – 107.

Veldkamp B.P. (2016). De inhoud en constructie van toetsen. In Sanders, P.F. (Red.), *Toetsen op School: Hoger onderwijs* (pp. 21 – 30). Arnhem: Cito. Geraadpleegd 10 juni 2020 van: https://www.rcec.nl/publicaties.

Wools, S. (2013). De validiteit van toetsscores In: Sanders, P.F. (Red.),*Toetsen op School* (pp. 69 – 83). Arnhem: Cito. Geraadpleegd 10 juni 2020 van: https://www.rcec.nl/publicaties.