

Verslag lijn 3A

Taal en niveau: Spaans vmbo GL/TL en Frans vmbo GL/TL

Vaardigheden: mediation, spreekvaardigheid

Uitvoering: Rosa Bartman; Eva Cornuijt; Danai Psathas

Redactie: Rosa Bartman; Marian Marchal; Danai Psathas; Noëlle Sommer

Inhoudsopgave

1. Aanleiding onderzoek	5
2. Eerder onderzoek.....	5
3. Onderzoeksvragen	6
4. Toetsoperationalisatie	7
5. Conclusies.....	20
6. Aanbevelingen	22
Referenties	24

1. Aanleiding onderzoek

Dit onderzoek is onderdeel van lijn 3 van het onderzoeksproject 'Talenexamens van de toekomst' en richt zich op doelgericht communiceren, wat we hebben vertolkt in taken met een *scenario-based* karakter. Dit betekent dat de ontwikkelde taken een betekenisvolle context bieden, waarin de leerling een opdracht uitvoert. Het doel van dit onderzoeksproject was o.a. het adviseren van vakvernieuwingscommissie MVT en een overtuigend idee geven van taken die wel of niet haalbaar zijn in een vernieuwd eindexamen.

2. Eerder onderzoek

In een eerder stadium van het onderzoeksproject (2021) is voor Spaans een soortgelijke taak ontwikkeld. De belangrijkste aandachtspunten¹ uit deze taak zijn hieronder opgesomd:

- Voor het opnemen van spreekvaardigheid is aan de leerling gevraagd om de eigen mobiele telefoon te gebruiken. Dit is in een highstake toets geen optie, dus het is van belang om een goede applicatie te vinden of ontwikkelen waarin de kandidaten hun spraak kunnen opnemen, die vervolgens toegankelijk is voor de beoordelaars.
- Het bleek dat leerlingen veel moeite hadden met spreekvaardigheid, omdat hier in de les weinig mee wordt geoefend. Het oefenen van spreekvaardigheid is dus van belang om een dergelijke taak te laten slagen. Wanneer besloten zou worden om spreekvaardigheid op te nemen in het eindexamen, zullen de lessen daar dus ook op aangepast moeten worden.
- In de taak werden de leerlingen gevraagd het gesprek te vervolgen en werden een aantal sturzinnen (in het Nederlands) en chunks (in het Spaans) gegeven om spraak uit te lokken. Hierdoor dreigde de taak echter meer vertalen dan spreekvaardigheid te toetsen. Er moet dus worden nagedacht over een opdrachtvorm die authentieker is en daadwerkelijk spreekvaardigheid uitlokt.
- De spraakoutput van de kandidaten werd paarsgewijs beoordeeld. Hiervoor is handmatig alle output gesorteerd en klaargezet voor de beoordelaars, wat erg tijdrovend is. Idealiter zou een geautomatiseerd systeem ingezet worden om de output paarsgewijs te beoordelen.
- De mediation-opdracht werd beoordeeld door simpelweg aan de beoordelaars te vragen of de leerling volgens hen de opdracht wel of niet behaald had. Hoewel de beoordelaarsovereenstemming zeer hoog was, kan er t.a.v. transparantie worden onderzocht of een gedetailleerder beoordelingsmodel ook mogelijk is om mediation te beoordelen.

¹ Zie onder de paragrafen 4.2 en 4.4 hoe deze aandachtspunten zijn verwerkt in de huidige taak.

3. Onderzoeksvragen

De hoofdvraag voor trial 3A is als volgt:

- Op welke manier kunnen de constructen spreekvaardigheid en mediation afgenomen en beoordeeld worden in een highstake toets (examen)?

Deze vraag wordt beantwoord aan de hand van verschillende deelvragen, die verschillende aspecten van het toetsen van spreekvaardigheid en mediation betreffen. Ze zijn opgedeeld in onderzoeksvragen wat betreft de toetsoperationalisatie en de beoordeling.

Toetsoperationalisatie

- Hoe kan mediation getoetst worden en wat voegt dit construct toe ten opzichte van luistervaardigheid?
- Hoe kan de kandidaat zo goed mogelijk in de situatie gepositioneerd worden om een zo veel mogelijk 'directe reactie' uit te lokken?
- In hoeverre kiezen kandidaten in de taak een scenario dat past bij hun profiel?
- Hoe waarden kandidaten de taken?

Beoordeling

- Hoe kan mediation beoordeeld worden met beoordelingskaders?
- Hoe kan spreekvaardigheid paarsgewijs beoordeeld worden?
- In hoeverre is er draagvlak onder docenten voor een paarsgewijze beoordeling van spreekvaardigheid?
- Hoeveel beoordelingen zijn er per spraakopname nodig om tot een betrouwbaar oordeel te komen?
- Welke invloed hebben de verschillende scenario's waaruit kandidaten kunnen kiezen op de beoordeling?

4. Toetsoperationalisatie

4.1 Overzicht taak

Trial 1: Spaans

In de taak konden leerlingen kiezen uit een van de volgende vier scenario's: 'in de souvenirwinkel', 'op de manege', 'autopech' of 'bij de dokter'. De taak bestond steeds uit twee opdrachten, ongeacht het gekozen scenario. In de eerste opdracht werd mediation getoetst en in de tweede opdracht spreekvaardigheid.

In de eerste opdracht luisterden de leerlingen naar een gesprek in het Spaans tussen een Nederlandse vriend of familielid en een Spanjaard. Het gesprek was gesitueerd in het door de leerling gekozen scenario. De Nederlandse vriend of het Nederlandse familielid begrijpt niet goed wat de Spanjaard tegen hem of haar zegt en de leerling wordt middels een tekstballon bij een afbeelding gevraagd in het Nederlands uit te leggen wat er in het Spaans is gezegd.

In de tweede opdracht kreeg de leerling acht vragen in het Spaans te zien, wederom aan de hand van tekstballonnen bij een afbeelding. De leerling moest in het Spaans antwoord geven op minstens vijf van deze vragen. De vragen hadden opnieuw betrekking op het gekozen scenario en waren zo geformuleerd dat het leek alsof het gesprek uit de eerste opdracht werd vervolgd.

en la tienda de recuerdos in de souvenirwinkel

Je bent samen met een Nederlandse vriend op vakantie in Spanje. Jullie staan in een souvenirwinkel om cadeautjes voor jullie familie en vrienden uit te zoeken. Hieronder hoor je een gesprek tussen de Spaanse winkeleigenaar en je Nederlandse vriend.

Luister naar het gesprek.



Je vriend begrijpt niet goed wat de winkeleigenaar tegen hem zegt. Jij biedt je hulp aan.

Maak een spraakopname van 20-30 seconden waarin je in het Nederlands antwoord geeft op de vraag van je vriend.

en la tienda de recuerdos in de souvenirwinkel

Je hebt net aan je vriend uitgelegd wat de winkeleigenaar heeft gezegd. Je gaat nu in het Spaans het gesprek met de winkeleigenaar vervolgen door te antwoorden op minimaal vijf van de onderstaande vragen. Maak daarbij volledige zinnen.



Neem ongeveer 10 minuten de tijd om je voor te bereiden. Je mag hierbij kladpapier gebruiken. Maak vervolgens een opname van 30-60 seconden waarin je in het Spaans in hele zinnen antwoord geeft op minimaal vijf van bovenstaande vragen.

Figuur 1. Voorbeeldschermen uit de afnameomgeving van opdracht 1 (links) en opdracht 2 (rechts) voor trial 1.

Trial 2: Frans

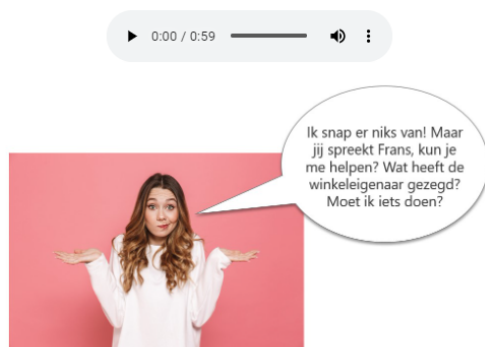
In trial 2 is een naar het Frans vertaalde versie van de taak van trial 1 ingezet. Deze taak was grotendeels identiek aan de taak die in trial 1 is ingezet. Bij het vertalen van de inhoud van deze taak is er echter ook extra aandacht besteed aan de moeilijkheidsgraad van de taak, omdat veel leerlingen in trial 1 aangaven de taak te moeilijk te vinden (zie sectie 4.3). In trial 2 is er daarom gelet op het spreektempo van de twee sprekers van het Frans in de eerste opdracht en de formulering van de vragen die de leerlingen in de tweede opdracht moesten beantwoorden. Daarnaast kregen de leerlingen in trial 2 bij de tweede opdracht een woordenlijst aangereikt die ze bij de uitvoering van deze opdracht konden gebruiken.

dans une boutique de souvenirs - in de souvenirwinkel

Je bent samen met een Nederlandse vriendin op vakantie in Frankrijk. Jullie staan in een souvenirwinkel om cadeautjes voor jullie familie en vrienden uit te zoeken. Hieronder hoor je een gesprek tussen de Franse winkelleigenaar en je Nederlandse vriendin. De onderstaande woorden kunnen je helpen het gesprek beter te begrijpen.

la boutique de souvenirs = de souvenirwinkel
le/la propriétaire = de eigenaar / de eigenaresse
la pause déjeuner / la pause de midi = de lunch-/middagpauze
rouvrir = weer opengaan
plaître = bevallen / leuk vinden
acheter = kopen
payer = betalen

Luister naar het gesprek.



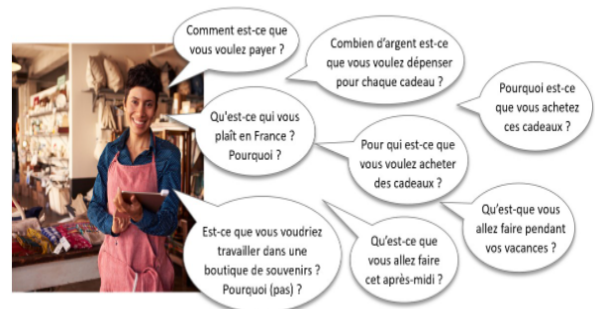
Je vriendin begrijpt niet goed wat de winkelleigenaar tegen haar zegt. Jij biedt je hulp aan.

Maak een spraakopname van 20-30 seconden waarin je in het Nederlands antwoord geeft op de vraag van je vriendin.

dans une boutique de souvenirs - in de souvenirwinkel

Nadat je je vriendin hebt geholpen, raak je in gesprek met de eigenaresse van de souvenirwinkel. Zij stelt jou een aantal vragen, die je beantwoordt. Geef in het Frans antwoord op minimaal vijf van de onderstaande vragen. Maak daarbij volledige zinnen. Je kunt onderstaande woorden gebruiken om je te helpen bij het maken van je zinnen.

la boutique de souvenirs = de souvenirwinkel
le/la propriétaire = de eigenaar / de eigenaresse
la pause déjeuner / la pause de midi = de lunch-/middagpauze
rouvrir = weer opengaan
plaître = bevallen / leuk vinden
acheter = kopen
payer = betalen



Neem ongeveer 10 minuten de tijd om je voor te bereiden. Je mag hierbij kladpapier gebruiken. Maak vervolgens een opname van 30-60 seconden waarin je in het Frans in hele zinnen antwoord geeft op minimaal vijf van de bovenstaande vragen.

Figuur 2. Voorbeeldschermen uit de afnameomgeving van opdracht 1 (links) en opdracht 2 (rechts) voor trial 2.

4.2 Ontwikkeling

Voor beide trials is de taak ontwikkeld voor vmbo-gt/mavo 4-kandidaten. Hiervoor zijn de descriptoren uit het ERK (Europees Referentiekader; Nederlandse Taalunie, 2022) aangehouden die voor de verschillende vaardigheden op dit niveau gelden. Deze worden weergegeven in Tabel 1.

Tabel 1. Gebruikte ERK-descriptoren in trials 1 en 2.

Categorie	Niveau	Descriptor
Mondeling specifieke informatie doorgeven	A2	- Kan (in taal B) het punt overbrengen dat wordt gemaakt in duidelijke gesproken mededelingen (in taal A) met betrekking tot vertrouwde alledaagse onderwerpen, al zal hij of zij de boodschap moeten vereenvoudigen en naar woorden moeten zoeken. - Kan (in taal B) het punt overbrengen dat wordt gemaakt in korte, heldere, eenvoudige berichten, instructies en mededelingen, mits deze langzaam en duidelijk zijn uitgedrukt in eenvoudige taal (in taal A). - Kan (in taal B) op een eenvoudige manier een reeks korte, eenvoudige instructies overbrengen, mits de originele informatie (in taal A) duidelijk en langzaam gearticuleerd is uitgesproken.
Mondeling tekst verwerken	A2	Kan (in taal B) het belangrijkste punt of de belangrijkste punten in eenvoudige, korte informatieve teksten (in taal A) over vertrouwde onderwerpen samenvatten.
Als intermediair optreden in informele situaties	A2	Kan (in taal B) de algemene boodschap overbrengen van wat er is gezegd (in taal A) in alledaagse situaties en daarbij de elementaire culturele conventies volgen en de essentiële informatie overbrengen, mits de sprekers duidelijk articuleren en hij of zij om herhaling of verduidelijking kan vragen.
Mondelinge productie in het algemeen	A2	Kan een eenvoudige beschrijving of presentatie geven van mensen, woon- of werkomstandigheden, dagelijkse routines, voorkeuren en afkeuren en dergelijke in een korte reeks simpele frasen en zinnen die als in een lijst met elkaar verbonden zijn.

Monoloog	A2	- Kan alledaagse aspecten beschrijven van zijn of haar omgeving, zoals mensen, plaatsen en werk- of studie-ervaring. - Kan korte, elementaire beschrijvingen geven van gebeurtenissen en activiteiten. - Kan zijn of haar gezinssituatie, woonomstandigheden, onderwijsachtergrond en huidige of laatste werk beschrijven. - Kan in het kort iets vertellen over wat hij of zij van plan is in het weekend of in de vakantie te gaan doen.
----------	----	---

Bij de ontwikkeling van de taak is rekening gehouden met het feit dat de opdrachten aangepast moeten kunnen worden naar andere talen en andere niveaus. Zo kunnen de fragmenten die als input dienen zo gemaakt of gekozen worden dat ze passen bij het niveau van de doelgroep. Ook is het mogelijk om ervoor te zorgen dat de vragen bij de tweede deelopdracht betrekking hebben op ERK-descriptoren van andere niveaus.

Hoewel in deze taak ook luistervaardigheid wordt getoetst, is er gekozen voor een mediation- insteek om de taak authentieker te maken. Luistervaardigheid wordt traditioneel getoetst aan de hand van meerkeuzevragen, terwijl deze manier van toetsen weinig valide is wanneer het gaat om de vraag of een leerling succesvol luisterbegrip kan opbouwen in alledaagse situaties. Meerkeuzevragen lijken namelijk vooral *shallow listening*, een oppervlakkige manier van luisteren zonder diepgaand begrip van de input, en het gebruik van toetsstrategieën door leerlingen zoals *keyword matching* uit te lokken (Cito, 2023). Tegelijkertijd blijft het belangrijk om te realiseren dat mediation duidelijk een ander construct is dan luistervaardigheid.

Ook bij de tweede deelopdracht is gepoogd de taak authentieker te maken door in plaats van met sturzinnen en *chunks* te werken met vragen in tekstballonnen. Naar verwachting zou dit twee voordelen bieden, namelijk (a) dat de opdracht meer de natuurlijke vorm van een gesprek aanhoudt (vragen en antwoorden in plaats van een monoloog) en (b) dat de leerling op authentieke wijze het vocabulaire dat ze in het geven van een antwoord nodig hebben al in de vraag aangeboden krijgen.

Afnameomgeving

Voor het afnemen van de taak is gebruikgemaakt van PCIBex, een open-source software die vaak gebruikt wordt voor onderzoeksdoeleinden. Binnen de afnameomgeving konden de leerlingen zowel de fragmenten beluisteren, de opdrachtinstructies lezen als hun eigen spraak opnemen. Op deze manier kon de situatie omzeild worden waarin leerlingen hun eigen mobiele telefoon of andere externe apparatuur moesten gebruiken om hun spraak op te nemen. De geluidsopnames werden direct geüpload op een server van Cito.

4.3 Afname

Trial 1

In trial 1 is de taak afgenomen bij drie klassen op twee verschillende scholen, één in Amsterdam en één in Lekkerkerk. De leerlingen zaten allemaal in het derde leerjaar van vmbo GL/TL. Hoewel de taak ontworpen was voor leerlingen van het vierde leerjaar, is vanuit praktische overwegingen gekozen om de taak af te nemen bij leerlingen in het derde leerjaar. Voor afname zijn er van alle leerlingen toestemmingsverklaringen verzameld. Voor de leerlingen die jonger waren dan 16 jaar hebben de ouders toestemming gegeven. De leerlingen werden zoveel mogelijk uit elkaar gezet en werden aangemoedigd een koptelefoon met microfoon te gebruiken. Er waren echter niet genoeg koptelefoons beschikbaar voor elke leerling. Er was veel verschil tussen leerlingen in hoe lang ze met de taak bezig waren, maar over het algemeen duurde de afname tussen de 15 en 20 minuten.

Tijdens de afname bleek dat de leerlingen de taak erg moeilijk vonden en bovendien onzeker waren over het opnemen van hun spraak. Net zoals bij de eerder afgenomen toetstaak was er dus een effect te zien van het feit dat spreekvaardigheid in de les nauwelijks geoefend wordt: sommige leerlingen gaven aan nog bijna nooit in het Spaans gesproken te hebben. Bij de tweede afname hebben we daarom geprobeerd leerlingen extra gerust te stellen door hen eraan te herinneren dat de taak bedoeld is voor examenleerlingen en dat ze geen cijfer krijgen. Daarnaast probeerde een aantal leerlingen de vragen voor de tweede opdracht te vertalen via internet en vroegen anderen of ze een woordenboek mochten gebruiken. Ten slotte waren er ook enkele onduidelijkheden over de taak. Leerlingen vroegen of ze de opname in het Spaans of in het Nederlands moesten maken en of ze bij de tweede opdracht de vragen moesten herhalen. De authenticiteit die in de tweede opdracht gepoogd was te creëren kwam dus niet helemaal goed uit de verf en blijft een heikel punt. Ook begrepen veel leerlingen niet wat er in het feedbackformulier met de term 'profiel' bedoeld werd.

Er zijn in totaal 28 opnames per opdracht verzameld. Dit betrof 25 bruikbare (verstaanbare, opdracht-gerelateerde) opnames voor opdracht 1 en 16 voor opdracht 2. Er was geen duidelijke relatie tussen het gebruik van een koptelefoon met microfoon en de hoeveelheid achtergrondgeluid, maar voor deze analyse was ook weinig data beschikbaar.

Scenariokeuze

Verreweg de meeste leerlingen kozen bij deze afnames voor het scenario 'bij de dokter'. Er was verder geen enkele leerling die het scenario 'op de manege' koos. In Tabel 2 is per opdracht aangegeven hoeveel bruikbare opnames er waren per opdracht en scenario.

Tabel 2. Scenariokeuze trial 1.

Scenario	Opdracht 1	Opdracht 2
Autopech	6	3
Bij de dokter	13	8
Souvenirwinkel	6	5
Op de manege	-	-
Totaal	25	16

De meeste leerlingen (n = 13) deden het profiel Economie. Zij kozen heel verschillende scenario's. Alle leerlingen met het profiel Zorg kozen het scenario 'bij de dokter', met uitzondering van één die voor het scenario 'in de souvenirwinkel' koos. De ene leerling met het profiel Techniek koos ook voor het scenario 'bij de dokter'. Er was dus geen duidelijk verband tussen het profiel van de leerlingen en het gekozen scenario.

Evaluatie leerlingen

In totaal hebben 22 leerlingen het evaluatieformulier ingevuld. Er was geen verschil in hoe leuk of moeilijk de leerlingen de eerste en de tweede opdracht vonden, en in hoeverre ze een soortgelijke opdracht al eens hadden geoefend. Wel waren er bij de open vragen meer leerlingen (n = 6) die aangaven dat ze nog niet vaak met spreken hadden geoefend dan met luisteren (n = 2). Ook gaven veel leerlingen (n = 13) aan dat ze de vragen waarop ze bij opdracht 2 antwoord moesten geven niet begrepen. Desondanks waren leerlingen wel positief over dat ze iets mochten inspreken en dat deze taak anders was dan anders, al vonden ze het niveau te hoog. De meeste leerlingen gaven de taak in zijn geheel 3 van de 5 sterren.

Trial 2

In trial 2 is de taak afgenomen onder 19 leerlingen van twee verschillende scholen, een school in Hellevoetsluis (havo 3, n = 10 leerlingen) en een school in Zwolle (mavo 4, n = 9 leerlingen). Voor afname zijn er van alle leerlingen toestemmingsverklaringen verzameld. Voor de leerlingen die jonger waren dan 16 jaar hebben de ouders toestemming gegeven. De leerlingen werden tijdens de afname zoveel mogelijk uit elkaar gezet, zodat zij de opdracht individueel zouden uitvoeren. Iedere leerling heeft een koptelefoon met microfoon gebruikt om de opdracht uit te voeren. Tijdens de afname bleek, net als in trial 1, dat de leerlingen de taak erg moeilijk vonden, ondanks de aanpassingen die voor trial 2 aan de taak zijn gedaan om de taak iets makkelijker te maken. Bovendien waren de leerlingen erg onzeker over het opnemen van hun spraak.

Hoewel er in totaal 19 leerlingen hebben deelgenomen aan deze taak, zijn er in totaal slechts 18 opnames per opdracht verzameld. Eén leerling ondervond namelijk technische problemen tijdens de afname en kon de taak daarom niet afronden. Van de in totaal 36 opnames waren 32 opnames bruikbaar (verstaanbaar en gerelateerd aan de opdracht): 18 voor opdracht 1 en 14 voor opdracht 2.

Scenariokeuze

De meeste leerlingen kozen voor de scenario's 'bij de dokter' en 'souvenirwinkel'. De scenario's 'autopech' en 'op de manege' zijn beide slechts één keer gekozen. In Tabel 3 is per scenario aangegeven hoeveel bruikbare opnames er waren per opdracht.

Tabel 3. Scenariokeuze trial 2.

Scenario	Opdracht 1	Opdracht 2
Autopech	1	1
Bij de dokter	9	6
Souvenirwinkel	7	6
Op de manege	1	1
Totaal	18	14

De keuze van het scenario bleek voor de 9 mavo 4-leerlingen niet samen te hangen met hun profielkeuze. Van deze leerlingen deden 8 leerlingen het profiel Economie en één leerling deed het profiel Zorg en welzijn. Zij kozen allemaal heel verschillende scenario's. Zo waren alle scenario's vertegenwoordigd onder de keuzes van de 8 leerlingen met het profiel Economie en koos de leerling met het profiel Zorg en welzijn voor het scenario 'souvenirwinkel'. De leerlingen in havo 3 hebben nog geen profiel; pas vanaf leerjaar 4 hebben deze leerlingen een profiel.

Evaluatie leerlingen

In totaal hebben 17 leerlingen het evaluatieformulier ingevuld. Er was geen verschil in hoe leuk of moeilijk de leerlingen de eerste en de tweede opdracht vonden: de leerlingen gaven bij de stellingen in de vragenlijst aan beide opdrachten niet zo leuk en van gemiddelde moeilijkheid te vinden, al bleek uit de open vragen dat de leerlingen het vooral moeilijk vonden om zelf in het Frans te spreken ($n = 5$). Slechts twee leerlingen gaven in de open vragen aan het luisteren naar een gesprek in het Frans moeilijk te vinden. Ook gaven de leerlingen aan dat ze met beide opdrachten nog niet hadden geoefend op school. Op een schaal van 1 (zeer negatief oordeel) tot 5 sterren (zeer positief oordeel) beoordeelden de leerlingen de taak in zijn geheel met gemiddeld een 2,4 ($SD = 1,18$). Het vaakst gegeven oordeel was 1 van de 5 sterren ($n = 5$ leerlingen). Ondanks deze lage score voor de taak als geheel gaven sommige leerlingen aan het

spreekgedeelte leuk te hebben gevonden. Eén leerling gaf aan het leuk te vinden dat deze taak “een keer iets anders” is. Nog een leerling gaf aan de taak leuk te vinden, omdat het leerzaam was.

4.4 Beoordeling

Trial 1

Mediation

Aangezien in de eerder afgenomen toetstaak (in 2021) het construct mediation holistisch is beoordeeld, is voor de eerste deelopdracht van de huidige taak gepoogd een beoordelingsmodel te creëren met meer gestructureerde beoordelingscriteria. De beoordeling is gebaseerd op de ERK-descriptoren voor mediation op A2-niveau, die worden weergegeven in Tabel 4.

Tabel 4. ERK-descriptoren voor mediation die zijn gebruikt in de beoordeling.

Categorie	Descriptor
Specifieke informatie doorgeven	Kan (in taal B) specifieke, relevante informatie overbrengen uit korte, eenvoudige teksten, etiketten en kennisgevingen (in taal A) over vertrouwde onderwerpen.
Tekst verwerken	Kan (in taal B) het belangrijkste punt of de belangrijkste punten in eenvoudige, korte informatieve teksten (in taal A) over vertrouwde onderwerpen samenvatten.

Naar aanleiding hiervan zijn per scenario de hoofdboodschap en overige relevante informatie bepaald (zie Tabel 5). Voor het overbrengen van de hoofdboodschap werden 2 punten toegekend als deze volledig was, en 1 punt als de hoofdboodschap deels was overgebracht. De overige relevante informatie bestond altijd uit twee of drie boodschappen, waarbij per boodschap één punt werd toegekend. Hierbij waren geen deelscores te behalen. De beoordeling werd getest in een pilot-beoordeling, waarna enkele aanpassingen zijn gedaan aan het beoordelingsmodel. De opdrachten zijn vervolgens beoordeeld door vier toetsdeskundigen, die allemaal aangaven dit een prettige en eenduidige manier van beoordelen te vinden. Er was ook een redelijke overeenstemming tussen de beoordelaars. De onenigheid werd met name veroorzaakt doordat leerlingen parafrases gebruikten waarover nog geen overeenstemming was bereikt, sommige irrelevante informatie impliciet bleef omdat het duidelijk werd uit andere informatie, of doordat de hoofdboodschap of overige relevante informatie uit meerdere delen bestond. Er is daarom naar aanleiding van de beoordeling een verbeterde versie van het beoordelingsmodel gemaakt.

Tabel 5. Hoofdboodschap en overige relevante informatie per scenario in trial 1.

	Hoofdboodschap	Overige relevante informatie			Irrelevante informatie
Souvenirwinkel	De winkel sluit over tien minuten (vanwege de siesta).	Reden voor sluiting is lunch en siesta	Om vijf uur gaat de winkel weer open.	Weet je al wat je wil kopen?	Heb je ergens hulp bij nodig?
Op de manege	Willen je kinderen helpen bij het voeren van de paarden straks om drie uur? (Dan moeten ze om drie uur bij de paardenhokken zijn.)	De paarden worden elke dag om drie uur gevoerd.	De paarden eten groente en fruit, vooral appels vinden ze lekker.	Het is bijna drie uur.	Wat vinden jullie van de manege? Vinden jullie de paarden en ezels leuk?
Autopech	De auto moet naar de garage en is morgen klaar.	Je kunt in een half uur met de bus naar de garage	De motor van de auto is kapot.	De garage is in Cobeña, een dorp hier in de buurt.	Waar hebben jullie hulp bij nodig?
Bij de dokter	Blijf uit de zon (<i>of</i> kijk uit), rust uit (<i>of</i> ga naar huis) en drink veel (water).	De zon is sterk, te veel zon is gevaarlijk.	Als je je (morgen nog) niet beter voelt, neem dan opnieuw contact op met de dokterspraktijk.	-	Waar hebben jullie hulp bij nodig?

Spreekvaardigheid

De tweede deelopdracht werd paarsgewijs beoordeeld. Om het proces te vergemakkelijken n.a.v. de eerder afgenomen toetstaak werd gebruik gemaakt van software die ontwikkeld is binnen Cito. Beoordelaars zagen de *waveform* van elk geluidsfragment en konden zelf op afspelen/pauzeren klikken. Nadat ze aangaven welk geluidsfragment ze beter vonden, werd de volgende vergelijking getoond. De te vergelijken geluidsfragmenten werden adaptief geselecteerd. Dit vermindert het aantal benodigde vergelijkingen, doordat vergelijkingen die duidelijk zijn (bijvoorbeeld omdat het niveau ver uit elkaar ligt) niet nogmaals vergeleken hoeven te worden.

In een instructiesessie voor beoordelaars deden de beoordelaars eerst een aantal oefenvergelijkingen met geluidsopnames die in een eerder stadium van het onderzoeksproject (2021) waren verzameld, waarbij de beoordelaars eerst intuïtief beoordeelden. Daarna werd besproken waarop de beoordelaars gelet hadden, aan de hand van de geluidsfragmenten. Docenten gaven aan te letten op vloeiendheid, uitspraak, complexiteit en gevarieerdheid van de woordenschat en zinsbouw. Een tweede groep docenten gaf aan dat ze vloeiendheid minder belangrijk vonden in het doelniveau (A2). Daarnaast letten de docenten erg op het communicatieve aspect, namelijk of de boodschap was overgebracht en of het begrijpelijk was. Dit laatste vonden de docenten lastig, deels omdat de taak niet bekend was, maar ook omdat ze vragen hadden bij het nut van de inhoud. In een vervolgtask moet het doel van de taak dus duidelijk zijn voor de docent tijdens het beoordelen van de opdracht. Daarnaast moet er volgens de docenten, onafhankelijk van de beoordelingswijze, op de één of andere manier ondervangen worden of aan de taak is voldaan. Een ander aspect dat de beoordeling volgens de docenten lastig maakte was het feit dat leerlingen verschilden in hoeverre de spraak spontaan of voorbereid was. Hierbij is het van belang dat de randvoorwaarden wat betreft de beschikbare middelen (woordenboek, Google Translate) gelijk zijn bij alle afnames. Daarnaast was het voor de docenten lastig om geluidsfragmenten van verschillende scenario's met elkaar te vergelijken. Als er gebruik gemaakt wordt van keuzemogelijkheden zou de output op dezelfde opdrachten gegroepeerd kunnen worden, om zo alleen deze geluidsfragmenten met elkaar te vergelijken. Hierbij gaven de docenten ook aan dat ze, aan de hand van de context, probeerden in te vullen wat de leerling probeerden te zeggen. Twee docenten zeiden hierbij dat het goed en fijn was dat ze de leerlingen niet kenden. Dit pleit voor een geanonimiseerde beoordeling. Ook zei één docent "hoe vaker je het hoort, hoe meer ik eruit kan halen wat de leerling probeert te zeggen". Dit suggereert dat fragmenten die vaker beluisterd worden, bijvoorbeeld omdat deze vaker vergeleken wordt, een hogere score krijgen. Ten slotte gaven docenten aan hun twijfels te hebben bij holistische paarsgewijze beoordeling, omdat ze terugvielen in de verschillende aspecten van analytische beoordeling of met name dachten in termen van voldoende en onvoldoende.

De geluidsfragmenten die tijdens de afname waren verzameld werden voor beide trials thuis beoordeeld door toetsdeskundigen en constructiegroepleden. Vervolgens vulde elke beoordelaar een vragenlijst in, waarin verschillende aspecten van het beoordelingsproces werden bevraagd. Er waren in totaal 10 beoordelaars, waarvan 8 beoordelaars ook de vragenlijst hebben ingevuld. Uit de vragenlijst bleek dat de beoordelaars uiteenlopende standpunten hadden over paarsgewijze beoordeling. De meerderheid van de beoordelaars vond het beoordelingsproces prettig en intuïtief, hoewel sommigen het soms lastig vonden om te moeten kiezen tussen twee fragmenten van hetzelfde niveau en om holistisch en 'op gevoel' te beoordelen, alhoewel dit laatste volgens enkele beoordelaars wel steeds meer wende naarmate er meerdere vergelijkingen werden gedaan. Meerdere beoordelaars gaven echter aan dat ze het belangrijk vonden om binnen bepaalde kaders te blijven beoordelen. Wellicht zou het daarom wenselijk kunnen zijn om toch (paarsgewijs) te blijven beoordelen op bepaalde beoordelingsaspecten (zoals inhoud, woordenschat en uitspraak). Enkele suggesties voor verbetering van de tool waren het toevoegen van een notitieveld per vergelijking en het opnemen van een aparte knop voor het selecteren van het 'betere' fragment. Daarnaast bleek

uit de vragenlijst het belang van gelijkwaardige en bruikbare fragmenten: sommige fragmenten waren te kort, te veel 'voorgelezen' of te slecht te verstaan om goed te kunnen beoordelen of vergelijken. Een ander inzicht uit de vragenlijst was dat bepaalde beoordelingsaspecten, zoals verstaanbaarheid, effectiviteit (de boodschap komt over), vloeiendheid, woordenschat(fouten), als belangrijker werden beschouwd dan andere, zoals lengte, intonatie, voldaan aan taak en niveau van detail. Er was verdeeldheid onder de beoordelaars over de geschiktheid van paarsgewijze beoordeling voor een centraal examen. Hoewel sommigen het als passend beschouwden, werden er zorgen geuit over de subjectiviteit en praktische uitdagingen, zoals het trainen van (grote aantal benodigde) beoordelaars en het bepalen van de cesuur.

Na het verzamelen van de vergelijkingen is er per beoordelaar een ranking gemaakt van alle kandidaten: van de kandidaat die het best beoordeeld is tot de kandidaat die volgens die beoordelaar het minst goed gepresteerd heeft. De rankings van alle beoordelaars zijn vervolgens met elkaar vergeleken. Hieruit bleek dat er relatief weinig beoordelaarsovereenstemming was: een gewogen Cohen's kappa van 0,43 staat volgens Landis en Koch (1977) voor matige overeenstemming. Dit kan te maken hebben gehad met het feit dat de fragmenten, om verschillende redenen, slecht met elkaar te vergelijken waren. Hier kan bijvoorbeeld de kwaliteit van het geluidsbestand of het feit dat de fragmenten gebaseerd waren op verschillende opdrachten invloed op hebben gehad.

Trial 2

De opdrachten zijn in trial 2 op dezelfde manier beoordeeld als in trial 1. De eerste opdracht, de mediation-opdracht, is wederom middels een beoordelingsformulier beoordeeld, waarin leerlingen punten konden verdienen voor het gedeeltelijk of volledig overbrengen van de hoofdboodschap van het Franse luisterfragment en voor het noemen van overige relevante informatie. De inhoud van de hoofdboodschappen en de overige relevante informatie is voorafgaand aan de beoordeling vastgelegd (zie Tabel 6). Opdracht 1 is beoordeeld door drie toetsdeskundigen die ook in trial 1 hebben deelgenomen aan de beoordeling van deze opdracht.

Tabel 6. Hoofdboodschap en overige relevante informatie per scenario in trial 2.

	Hoofdboodschap	Overige relevante informatie			Irrelevante informatie
Souvenirwinkel	De winkel sluit over tien minuten (vanwege de lunchpauze).	Reden voor sluiting is lunch-/middagpauze	Om drie uur gaat de winkel weer open.	Weet je al wat je wil kopen?	Heb je ergens hulp bij nodig?
Op de manege	Willen je kinderen helpen bij het voeren van de paarden straks om drie uur? (Dan moeten ze om drie uur bij de paardenhokken zijn.)	De paarden worden elke dag om drie uur gevoerd.	Vandaag zijn er veel wortels en appels (en daar zijn de paarden dol op).	Het is bijna drie uur.	Wat vinden jullie van de manege? Vinden jullie de paarden en pony's leuk?
Autopech	De auto moet naar de garage (of: moet gemaakt worden) en is morgen klaar.	Je kunt in een half uur met de bus naar de garage	De motor van de auto is kapot.	De garage is in Aix-en-Provence, een dorp hier in de buurt.	Waar hebben jullie hulp bij nodig?
Bij de dokter	Blijf binnen (of: thuis, of: uit de zon), rust uit en drink veel (water).	Je moet voorzichtig zijn (of: de zon is erg sterk)	Als je je (morgen nog) niet beter voelt, neem dan opnieuw contact op met de dokterspraktijk.	-	Waar hebben jullie hulp bij nodig?

De tweede opdracht, de spreekopdracht, is wederom paarsgewijs beoordeeld. Er hebben 13 beoordelaars deelgenomen aan de beoordeling, die werkzaam zijn als toetsdeskundige Frans of Spaans of constructiegroeplid Frans. In tegenstelling tot in trial 1 is er voor deze beoordelaars geen instructiesessie georganiseerd, omdat het paarsgewijze beoordelen intuïtief is en zich redelijk vanzelf wijst. Wel hebben de beoordelaars een handleiding met daarin een uitleg van de opdracht en de vergelijkingstool gekregen. De beoordelaars zijn na afloop van de beoordeling, net als in trial 1, gevraagd om een vragenlijst in te vullen om deze manier van beoordelen te evalueren. Deze vragenlijst is ingevuld door 9 beoordelaars.

De beoordelaars die de vragenlijst hebben ingevuld waren overwegend positief over deze manier van beoordelen. De meeste beoordelaars vonden de werkwijze prettig en eenvoudig, al bleek ook hier de beoordeling lastiger te worden naarmate de niveaus van de te vergelijken fragmenten dichter bij elkaar lagen. Ook het feit dat sommige fragmenten over andere scenario's gingen of in lengte erg van elkaar verschilden bemoeilijkten soms de beoordeling. Enkele beoordelaars gaven onder andere om deze redenen aan deze manier van beoordelen niet objectief genoeg te vinden voor een centraal examen. Andere beoordelaars waren juist erg positief en gaven aan dat paarsgewijs beoordelen een uitstekend alternatief is voor andere, vaak lastige beoordelingswijzen. Sommige beoordelaars noemden enkele voorwaarden voor het gebruik van paarsgewijze beoordeling in een centraal examen. Er zou bijvoorbeeld een grote groep beoordelaars nodig zijn om tot een betrouwbare beoordeling te komen, en er zou een standaardnorm moeten komen (bijvoorbeeld door fragmenten van voorgaande jaren aan de te beoordelen fragmenten toe te voegen). Daarnaast gaven sommige beoordelaars aan dat er vooraf meer sturing nodig is voor zowel beoordelaars en leerlingen, en vond een enkeling dat deze manier van beoordelen het best zou werken in combinatie met een rubric.

De beoordelingsaspecten waar de beoordelaars tijdens de paarsgewijze beoordeling het meest en het minst op gelet hebben kwamen grotendeels overeen met de resultaten uit trial 1. Verstaanbaarheid, effectiviteit en uitspraak werden het meest belangrijk geacht, met daarop volgend inhoud en vloeiendheid. De aspecten waar beoordelaars tijdens de beoordeling het minst aandacht aan hebben besteed waren grammatica (complexiteit), niveau van detail, intonatie en lengte.

De beoordelaars waren erg positief over de gebruikte software. Eventuele verbeteringspunten die genoemd werden, waren het kunnen teruggaan naar reeds beoordeelde paren en het krijgen van een bevestiging na het aangeven van hun keuze. Daarnaast zouden sommige beoordelaars het fijn vinden om de leerlingopdracht bij de beoordeling in beeld te hebben en om commentaren of feedback te kunnen toevoegen aan een beoordeling.

De uiteindelijke beoordelingen van alle beoordelaars zijn ook bij deze trial met elkaar vergeleken middels de rankings per beoordelaar. De beoordelaarsovereenstemming uitgedrukt in Cohen's kappa was 0,68, wat aanzienlijke overeenstemming aangeeft (Landis & Koch, 1977).

5. Conclusies

Naar aanleiding van de toetsoperationalisatie en de afnamegegevens kunnen de onderzoeksvragen als volgt worden beantwoord.

Hoe kan mediation getoetst worden en wat voegt dit construct toe ten opzichte van luistervaardigheid?

Deze taak geeft een voorbeeld van een manier waarop mediation getoetst kan worden, namelijk door middel van een gesimuleerde situatie waarin de leerling moet optreden als bemiddelaar tussen een spreker van de doeltaal en een spreker van de moedertaal. Hierbij moet wel rekening gehouden worden met het feit dat luistervaardigheid ook getoetst wordt in dergelijke taken, en deze twee vaardigheden niet afzonderlijk van elkaar getoetst kunnen worden. Aan de andere kant is een taak die mediation toetst authentieker dan een taak met meerkeuzevragen, omdat receptie en productie gecombineerd worden, zoals dat ook vaak in de praktijk gebeurt.

Hoe kunnen we de kandidaat zo goed mogelijk positioneren in de situatie om een zo direct mogelijke reactie uit te lokken?

Er is geprobeerd om de taak zo authentiek mogelijk te maken door een context te schetsen en directe vragen te stellen door middel van een afbeelding van een persoon met tekstballonnen. Toch bleek dit niet voldoende: in trial 1 bleven net zoveel leerlingen ($n = 8$) de hij-vorm gebruiken als een directe aanspreekvorm ('jij/je' of 'u'), en in beide trials herhaalden sommige leerlingen hun output, hetgeen laat zien dat ze zich toch niet volledig in de situatie waanden. Enkele beoordelaars gaven ook aan de taak niet authentiek genoeg te vinden, omdat er zonder directe gesprekspartner altijd sprake is van 'eenrichtingsverkeer'.

Kiezen leerlingen in de taak een scenario dat past bij hun profiel?

In beide trials was er geen duidelijk verband tussen het profiel van de leerling en de scenariokeuze. De keuze van de leerlingen voor de scenario's leek vooral gebaseerd op wat klasgenoten deden. In een highstake setting is dit uiteraard niet mogelijk, dus het is mogelijk dat het verband duidelijk zichtbaarder wordt, wanneer deze 'ruis' wegvalt.

Hoe waarden leerlingen de taken?

Uit het evaluatieformulier bleek dat de leerlingen positief waren over het inspreken en dat ze het leuk vonden dat het een andersoortige taak betrof dan ze gewend zijn. Wel vonden ze het niveau te hoog en hadden sommige leerlingen nog niet veel met spreekvaardigheid geoefend. Tijdens de afname bleek ook dat leerlingen het spannend vonden om te beginnen met praten, waarschijnlijk mede doordat meerdere leerlingen zich in hetzelfde klaslokaal bevonden en ze elkaar konden horen praten. In het vervolg zal ervoor gezorgd moeten worden dat het onderwijs en de toetstaak beter op elkaar aansluiten.

Hoe kan mediation beoordeeld worden met beoordelingskaders?

Deze taak geeft een voorbeeld van een manier waarop mediation kan worden beoordeeld met bepaalde beoordelingskaders, namelijk met een beoordelingsmodel waarin punten worden toegekend voor verschillende aspecten die de leerling uit het beluisterde fragment zou moeten overbrengen. Hierbij is in de huidige taak een onderscheid gemaakt tussen een hoofdboodschap (waarvoor de leerlingen maximaal 2 punten konden verdienen) en drie stukken overige relevante informatie (waarvoor de leerlingen telkens 1 punt konden verdienen). Deze manier van beoordelen bleek goed te werken, met een redelijke beoordelaarsovereenstemming tot gevolg. Wel is het van belang om ervoor te zorgen dat de hoofdboodschap en overige relevante informatie in het geluidsfragment eenduidig is, en dat de verschillende geluidsfragmenten ongeveer hetzelfde aantal boodschappen bevatten.

Hoe kan spreekvaardigheid paarsgewijs beoordeeld worden?

Het is goed mogelijk om spreekvaardigheid paarsgewijs te beoordelen met speciaal hiervoor ontworpen software. De gebruikte tool bleek makkelijk en intuïtief te werken en creëerde per beoordelaar een ranking van alle geluidsopnamen, waaruit vervolgens een gemiddelde ranking kon worden gemaakt. In de huidige taak is geen cesuur bepaald en dit is dus nog een aspect dat bij paarsgewijze beoordeling verder moet worden onderzocht. Daarnaast bleek de beoordelaarsovereenstemming matig tot aanzienlijk te zijn in de twee verschillende trials.

In hoeverre is er draagvlak onder docenten voor een paarsgewijze beoordeling van spreekvaardigheid?

De beoordelaars (grotendeels docenten) hadden verschillende meningen over de geschiktheid van paarsgewijze beoordeling voor een centraal examen. Hoewel sommigen het een passende methode vonden, waren anderen bezorgd over de subjectiviteit en de praktische problemen, zoals het trainen van een groot aantal beoordelaars en het vaststellen van de cesuur.

Hoeveel beoordelingen zijn er per spraakopname nodig om tot een betrouwbaar oordeel te komen?

De tool die in deze taak gebruikt is voor de paarsgewijze beoordeling werkte met adaptieve selectie van de te vergelijken fragmenten. Op die manier zijn er in totaal minder vergelijkingen nodig vergeleken met tools waarin vergelijkingen tussen geluidsfragmenten op niet-adaptieve (willekeurige) wijze tot stand komen. De berekening van het aantal te maken vergelijkingen wordt dan ook door de tool zelf gedaan. Afhankelijk van het aantal beschikbare beoordelaars neemt een beoordelingssessie een bepaalde hoeveelheid tijd in beslag. Bij paarsgewijze beoordeling zal dus altijd een afweging moeten worden gemaakt tussen het aantal beoordelaars, de tijd die iedere beoordelaar aan het vergelijken besteedt en de gewenste beoordelaarsovereenstemming.

Welke invloed hebben de verschillende scenario's waaruit kandidaten kunnen kiezen op de beoordeling?

Verschillende beoordelaars gaven aan het lastig te vinden om geluidsopnamen die betrekking hadden op verschillende scenario's met elkaar te vergelijken. Het aanbieden van keuzemogelijkheden aan de leerling maakt de taak wellicht beter passend bij een leerling, maar maakt paarsgewijze beoordeling ook lastiger. Docenten gaven namelijk aan het lastig te vinden om geluidsopnames van verschillende scenario's met elkaar te gebruiken.

6. Aanbevelingen

Afnamecondities

Voor het centraal examineren van spreekvaardigheid moeten passende afnamecondities tot stand komen. Als een taak zoals deze in een centraal examen zou komen, zouden leerlingen bij voorkeur in een individuele afnameomgeving zitten en gebruikmaken van een hoofdtelefoon met microfoon om de kwaliteit van de geluidsopnames te waarborgen. Ook is een beveiligde server nodig waar de geluidsopnames van de leerlingen opgeslagen kunnen worden. Een beveiligde omgeving zal ook nodig zijn voor de beoordeling van de geluidsopnames van leerlingen.

Beoordeling

In de huidige taak vergeleken de beoordelaars geluidsopnames van verschillende scenario's met elkaar. De beoordelaars gaven aan dit lastig te vinden. Om dit te omzeilen zouden de geluidsopnamen van dezelfde opdrachten gegroepeerd kunnen worden om zo enkel deze geluidsfragmenten met elkaar te vergelijken. Ook bleek het lastig om geluidsfragmenten die op andere vlakken van elkaar verschilden (zoals spontane versus voorgelezen spraak) te vergelijken. Het is bij paarsgewijze beoordeling dus van belang om ervoor te zorgen dat de output altijd goed te vergelijken is.

Bij de beoordeling van het construct mediatie is het van belang dat de hoofdboodschap (en overige relevante informatie) eenduidig is en dat alle informatie waarop beoordeeld wordt een bepaalde informatiewaarde bevat, om te voorkomen dat leerlingen bepaalde informatie niet doorgeven omdat deze al uit de context duidelijk is. Daarnaast is het bij de beoordeling van mediatie lastig dat het voor de leerling mogelijk onduidelijk is welke informatie ze precies moeten doorgeven: alleen de hoofdboodschap of ook de overige informatie. Het is dus belangrijk om hierover vooraf duidelijk en transparant te zijn richting de leerling.

Vervolgonderzoek

Ondanks de poging om de taak zo authentiek mogelijk te maken en de leerling in het scenario te positioneren, waren er tijdens de afname aanwijzingen dat veel leerlingen zich niet goed genoeg konden verplaatsen in de geschetste situatie. Er zal dus meer vervolgonderzoek moeten worden gedaan om te achterhalen of het mogelijk is om zonder gesprekspartner een authentiek scenario te creëren. Een mogelijke denkrichting zou kunnen zijn om een gesimuleerde gesprekspartner in te zetten, of andere ICT-toepassingen.

Bij paarsgewijze beoordeling kunnen er van de rankings per beoordelaar een gemiddelde, algemene ranking van alle kandidaten worden opgesteld. Mocht deze manier van beoordelen worden ingezet bij een centraal examen, dan moeten cijfers aan deze ranking gekoppeld worden. Hiervoor zouden bepaalde fragmenten vanuit verschillende plekken op de ranking voorgelegd kunnen worden aan docenten, die er een holistisch cijfer aan koppelen. Aan fragmenten tussen deze ankerfragmenten op de ranking kunnen tussenliggende cijfers toegekend worden. Dit biedt ook een optie voor vervolgonderzoek: is paarsgewijze beoordeling betrouwbaarder en eenvoudiger door gebruik te maken van ankerfragmenten? Een mogelijke werkwijze hiervoor zou zijn dat een aantal fragmenten holistisch worden beoordeeld en als ankerfragment dienen. Idealiter zouden deze fragmenten de onder- en bovengrens aangeven van bepaalde categorieën, zoals onvoldoende/voldoende/goed, of een bepaalde cijferschaal. Daarnaast is het belangrijk dat er meerdere geluidsfragmenten zijn voor elk ankerpunt, omdat ankerfragmenten kunnen verschillen in het aspect (vloeiendheid, uitspraak, woordenschat, etc.) waarop ze beter of juist minder goed zijn. Na het bepalen en normeren van deze ankerfragmenten, kunnen de overige fragmenten vergeleken worden met deze ankerfragmenten. Hierbij zouden dan maar enkele vergelijkingen per fragment nodig zijn. De categorieën waaronder het fragment zou kunnen vallen, kunnen namelijk snel bepaald worden, waarna overige vergelijkingen de betrouwbaarheid kunnen waarborgen. Voor meer ideeën over het efficiënter maken van paarsgewijs beoordelen, zie Lesterhuis (2017, p. 130).

Daarnaast bleek de beoordelaarsovereenstemming bij het paarsgewijs beoordelen in beide trials niet hoog genoeg voor een high-stakes toets. Er moet dus nog worden onderzocht of, en zo ja, hoe een hogere beoordelaarsovereenstemming bereikt kan worden. Voorbeelden van mogelijke denkrichtingen zijn het uitgebreid trainen van beoordelaars en/of het aanbieden van meer beoordelingskaders of richtlijnen die gebruikt kunnen worden tijdens het beoordelen. Ook zou er in een vervolgstudie gekeken kunnen worden of de beoordelaarsovereenstemming bij meer vergelijkbare fragmenten groter zou zijn.

Referenties

Cito. (2023). *Wat zijn de moeilijkheden bij het toetsen van luistervaardigheid en kan linguïstische mediation deze problemen (gedeeltelijk) oplossen?* [Intern document]

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. <https://doi.org/10.2307/2529310>

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative Judgement as a Promising Alternative to Score Competences (pp. 119-138). In E. Cano & G. Ion (Eds.), *Innovative Practices for Higher Education Assessment and Measurement*. IGI Global.

Nederlandse Taalunie. (2022). *Gemeenschappelijk Europees referentiekader voor talen: Leren, onderwijzen, beoordelen - Aangevulde uitgave*.
<https://taalunie.org/publicaties/216/gemeenschappelijk-europees-referentiekader-voor-talen-leren-onderwijzen-beoordelen>

Neem voor meer informatie over de specifieke taken en het onderzoek contact op met:
rosa.bartman@cito.nl

Verslag lijn 3B – trial 1

Taal en niveau: Duits havo

Vaardigheden: leesvaardigheid, schrijfvaardigheid, mediation

Uitvoering: Esther Cremers; Noëlle Sommer

Redactie: Rosa Bartman; Danai Psathas; Noëlle Sommer

Inhoudsopgave

1. Aanleiding onderzoek	5
2. Eerder onderzoek.....	5
3. Onderzoeksvragen	6
4. Toetsoperationalisatie	6
5. Conclusies.....	18
6. Aanbevelingen	19
Referenties	21

1. Aanleiding onderzoek

Deze trial is onderdeel van het onderzoeksproject 'Talenexamens van de toekomst'. Het doel van dit onderzoekproject was o.a. het adviseren van vakvernieuwingscommissie MVT en een overtuigend idee geven van taken die wel of niet haalbaar zijn in een vernieuwd eindexamen. Deze trial is onderdeel van lijn 3 en richt zich op doelgericht communiceren in een vreemde taal, wat we hebben vertolkt in taken met een *scenario-based* karakter. Dit betekent dat de ontwikkelde taken een realistische context bieden waarin de leerling een opdracht uitvoert. Terwijl in lijn 3A de vaardigheden mediation, luisteren en spreken centraal staan, staan in lijn 3B naast mediation de vaardigheden lezen en schrijven centraal.

2. Eerder onderzoek

In een eerder stadium (2020-2021) van het onderzoeksproject 'Talenexamens voor de toekomst' zijn voor Duits en Arabisch taken ontwikkeld waarin zowel lees- als schrijfvaardigheid worden getoetst. In de taak die voor Duits is ontwikkeld, lazen leerlingen (doelgroep: havo 5) een literaire tekst, waarbij ze vervolgens vier meerkeuzevragen maakten om hun begrip van de tekst te toetsen. Daarna volgde een schrijfo opdracht, waarin de leerling een Duitse e-mail aan een Duitse leeftijdsgenoot moest schrijven van 150-180 woorden over de gelezen tekst. De leerling moest in deze e-mail vanuit de ik-vorm verslag doen van de gebeurtenissen in het verhaal. De belangrijkste aandachtspunten die uit de taak voor Duits naar voren kwamen en die verbetering behoeven in deze nieuwe taak, zijn hieronder opgesomd.

1. Er moet goed nagedacht worden over hoeveel tijd de leerlingen voor de taak moeten krijgen. Bij de afname in 2020-2021 bleek dat 45 minuten te weinig was om deze taak te voltooien, mede door het feit dat er voorafgaand aan de afname in de deelnemende klassen nauwelijks was geoefend met schrijfvaardigheid (gedeeltelijk i.v.m. corona).
2. De schrijfo opdracht was afhankelijk van de leestekst. Er moet bij gecombineerde lees-schrijfaken goed nagedacht worden over een manier waarop leerlingen de schrijftaak toch kunnen maken, ook als ze de leestekst niet (volledig) hebben begrepen.
3. Bij de schrijfo opdracht was er door het ontwerp van de taak beperkte ruimte voor leerlingen om hun eigen structuur aan te brengen in hun tekst. Er dient een afweging gemaakt te worden tussen het vooraf aanbieden van een structuur die de leerling moet aanhouden, wat de beoordeling kan vergemakkelijken, en de mate van creatieve vrijheid we van de leerling mogen verwachten.
4. Door de beperkte mogelijkheden van Microsoft Forms is de taak op papier afgenomen, terwijl je de leerlingen een e-mail het liefst zou willen laten schrijven op een computer om de setting zo realistisch mogelijk te maken. Bij een volgende afname van een vergelijkbare taak moet er dus gezocht worden naar een geschikte, digitale afnameomgeving.

5. Het globale beoordelingsmodel was bij nader inzien te omslachtig, te uitgebreid en volgens een van de deelnemende beoordelaars te subjectief. Er moet dus nog een passende beoordelingswijze gevonden worden voor dit soort taken.

In secties 4.1, 4.2 en 4.4 wordt toegelicht hoe de eerste vier aandachtspunten die hierboven zijn genoemd, in de huidige taak zijn verwerkt.

3. Onderzoeksvragen

De hoofdvraag voor deze trial is als volgt:

- Hoe kunnen leesvaardigheid, schrijfvaardigheid en mediation getoetst worden in een geïntegreerde toetstaak die aansluit op alledaagse situaties en die toepasbaar is op verschillende niveaus (BB/KB/GLTL/HAVO/VWO)?

Deze vraag wordt beantwoord aan de hand van drie deelvragen die verschillende aspecten van het toetsen van leesvaardigheid, schrijfvaardigheid en mediation betreffen:

1. Is het mogelijk om een geïntegreerde toetstaak zodanig te ontwerpen dat de input geen belemmering vormt tijdens de schrijfo opdracht voor leerlingen die de input niet voldoende hebben begrepen?
2. Is het mogelijk om een beoordelingsvorm te bedenken die werkbaar is voor docenten en waarin zowel leesvaardigheid, schrijfvaardigheid en mediation worden beoordeeld?
3. Wat vinden leerlingen van de geïntegreerde taak die in dit onderzoek is ontwikkeld?

4. Toetsoperationalisatie

4.1 Overzicht taak

De taak bestond uit twee deelopdrachten. In de eerste deelopdracht werden schrijfvaardigheid en leesvaardigheid geïntegreerd getoetst en in de tweede deelopdracht werden mediation en leesvaardigheid geïntegreerd getoetst.

Aan het begin van de taak kregen de leerlingen drie verschillende Duitse vacatureteksten te lezen. Vervolgens kregen de leerlingen in de eerste deelopdracht de instructie om bij één vacature naar keuze in het Duits een sollicitatiemail van 150-170 woorden te schrijven. Hierbij kregen ze een checklist aangereikt van alle elementen die ze in hun sollicitatiemail moesten verwerken. Deze checklist bestond voor iedere vacature uit 10 of 11 elementen. Hiervan hadden 4 of 5 elementen direct betrekking op de inhoud van de leestekst en vereisten zo begrip van de

leestekst. Deze elementen zijn hieronder opgesomd. De overige 5 of 6 elementen hielden daarentegen geen verband met de inhoud van de leestekst en konden zo ook door leerlingen die de leestekst minder goed hadden begrepen, worden verwerkt in de sollicitatiemail. Op deze manier is gepoogd leesvaardigheid en schrijfvaardigheid in één (deel)opdracht te integreren, zonder leerlingen die de leestekst onvoldoende hebben begrepen bij het schrijfgedeelte te benadelen.

Elementen met betrekking op de leestekst:

- Maak duidelijk op welke baan je solliciteert.
- Leg uit dat die baan bij jou past.
- Reageer positief op één taak die volgens de vacature onderdeel is van je werkzaamheden.
- Noem ook iets wat je minder graag doet met betrekking tot de beoogde baan. Zelf verzinnen mag.
- Alleen bij vacature 3: Reageer op de verblijfsmogelijkheid voor de duur van je contract.

Voorafgaand aan de tweede deelopdracht kregen de leerlingen de twee overgebleven vacatures opnieuw te lezen. Na weer een keuze te hebben gemaakt voor een van de vacatures moesten ze in de tweede deelopdracht in het Nederlands een WhatsApp-bericht van 60-80 woorden schrijven aan een vriend(in), waarin ze de gekozen vacature aan de ander aanraadden. Opnieuw kregen ze een checklist, ditmaal met de volgende elementen:

- Vertel van je sollicitatie.
- Enthousiasmeer de ander om ook te solliciteren.
- Maak gebruik van informatie uit de vacaturetekst om te beschrijven dat de baan bij de ander past.

Het eerste en het derde element waren bedoeld om mediation te ontlokken. Bij het eerste element leidt de leerling het onderwerp van het bericht in door aan te geven dat hij/zij zelf ook gesolliciteerd heeft en bij het derde element haalt de leerling relevante informatie uit de vacaturetekst die bij de ander past. Wij zien beide elementen als een vorm van mediation, omdat de leerling betekenis creëert en overbrengt en hierdoor bemiddelt in een communicatieve situatie. Dit gebeurt in dit geval tussen twee verschillende talen (van het Duits naar het Nederlands), wat nog een extra laag creëert in het bemiddelen. Deze vormen van mediation komen terug in de gebruikte ERK-descriptoren voor mediation in Tabel 1 (zie volgende sectie).

4.2 Ontwikkeling

De huidige taak is ontwikkeld voor havo 5-kandidaten. Op dit niveau geldt dat leerlingen over het algemeen het ERK-niveau B2 voor leesvaardigheid behalen en het niveau A2+ voor schrijfvaardigheid (SLO, 2021). Omwille van het vernieuwende karakter van deze taak zijn echter descriptoren van lagere ERK-niveaus aangehouden bij het ontwikkelen van de taak. Voor de vaardigheid mediation is het nog onduidelijk welk niveau leerlingen in havo 5 zouden moeten beheersen. Hier hebben we daarom dezelfde niveaus als voor lees- en schrijfvaardigheid aangehouden. De gebruikte ERK-descriptoren worden weergegeven in Tabel 1.

Tabel 1. ERK-descriptoren die zijn gebruikt bij de ontwikkeling van de taak (Nederlandse Taalunie, 2022).

Categorie	Niveau	Descriptor
<i>Leesvaardigheid</i>		
Lezen ter informatie en argumentatie	A2	Kan de belangrijkste punten begrijpen van korte teksten over alledaagse onderwerpen (bijvoorbeeld levensstijl, hobby's, sport, weer).
Oriënterend lezen	B1	Kan de belangrijke informatie begrijpen in simpele, duidelijk opgestelde advertenties in kranten of tijdschriften, mits er niet te veel afkortingen in staan
<i>Schrijfvaardigheid</i>		
Correspondentie	B1	Kan schriftelijk reageren op een advertentie en om meer informatie vragen over interessante zaken. Kan een elementaire sollicitatiebrief opstellen met beperkte ondersteunende details.
<i>Mediation</i>		
Mediatie van een tekst	A2	Kan relevante informatie overbrengen uit helder gestructureerde, korte, eenvoudige, informatieve teksten, mits de teksten betrekking hebben op concrete, vertrouwde onderwerpen en geformuleerd zijn in eenvoudige alledaagse taal.
Mediatie van een tekst	B1	Kan informatie in heldere, goed gestructureerde informatieve teksten over vertrouwde of actuele onderwerpen of

		onderwerpen binnen de eigen belangstelling overbrengen, ondanks mogelijke lexicale beperkingen bij het formuleren. ¹
Specifieke informatie doorgeven in geschreven taal	B1	Kan (in taal B) schriftelijk specifieke, relevante informatie overbrengen uit directe informatieve teksten (in taal A) over vertrouwde onderwerpen.

Bij de ontwikkeling van de taak is rekening gehouden met het feit dat de opdrachten aangepast moeten kunnen worden naar andere talen en andere niveaus. De teksten die als input dienen (hier: vacatureteksten) kunnen zo worden gekozen dat ze passen bij het niveau van de doelgroep. Ook is het mogelijk om de schrijf- en mediationopdrachten zo aan te passen dat het product van deze opdrachten past bij het niveau van de doelgroep.

Gezien de focus van deze onderzoekslijn op *scenario-based* toetsing is er in deze taak gekozen voor een taakontwerp waarin doelgerichtheid en authenticiteit centraal staan. Het lezen van en reageren op vacatureteksten is iets waar leerlingen ongetwijfeld mee te maken gaan krijgen of wellicht al eens mee te maken hebben gehad. Het is dus een realistisch scenario, waarin de leerling een sollicitatiemail schrijft, en zo dus communiceert met een duidelijk doel. Er is expliciet gekozen voor een mail in plaats van een brief en een WhatsApp-bericht om de huidige, digitale werkelijkheid zo veel mogelijk na te bootsen. In de tweede deelopdracht werd de leerling daarnaast aangemoedigd om het WhatsApp-bericht in eigen taalgebruik te formuleren en om de inhoud naar eigen inzicht in te vullen, om de opdracht zo veel mogelijk te laten aansluiten bij de leefwereld van de leerling.

Ondanks dat Microsoft Forms bij de eerdere toetstaak in 2020-2021 niet geschikt werd bevonden als afnameplatform en er daarom een papieren toetstaak is ontwikkeld, is de huidige toetstaak wel via dit platform afgenomen. Vanwege de nadruk op authenticiteit was het belangrijk om de taak op de computer af te nemen, en hierbij bleek Microsoft Forms toch de beste afnameomgeving te zijn, onder andere doordat in Microsoft Forms goed mogelijk is om verschillende paden te creëren afhankelijk van de keuzes die leerlingen in de opdracht maken. Het grootste nadeel aan het gebruik van Microsoft Forms was het feit dat deze tool geen optie heeft om woorden te tellen, waardoor de leerlingen een ander programma, zoals Microsoft Word, moesten gebruiken om te controleren of ze aan de vereiste woordenaantallen voldeden. De leerlingen hebben bij de afname een volledig lesuur (50-60 minuten) de tijd gekregen om de taak te voltooien. Naar aanleiding van de eerdere toetstaak is er dus meer tijd gegeven, maar meer dan een lesuur werd achtten wij niet nodig, aangezien de huidige taak digitaal is

¹ In deze descriptor wordt ervan uitgegaan dat de taalgebruiker informatie overbrengt en vertaalt vanuit de moedertaal naar een vreemde taal. In deze taak spelen lexicale beperkingen in het mediation-gedeelte (deelopdracht 2) echter geen rol, omdat de leerling informatie moet overbrengen vanuit de vreemde taal, het Duits, naar de moedertaal.

afgenomen in plaats van op papier. Naar verwachting zou deze aanpassing er ook toe leiden dat leerlingen de schrijfpdracht sneller voltooien.

Een ander punt uit de eerdere toetstaak dat verbetering behoefde, was het feit dat er bij de schrijfpdracht weinig ruimte was voor het aanbrengen van eigen structuur, terwijl dit een vaardigheid is die bij schrijfvaardigheid van belang is. In de eerdere taak uit 2020-201 schreven leerlingen hun e-mail in een tabel op een vel papier, waarbij iedere rij in de tabel ruimte bood om iets te schrijven bij een bepaald element dat de leerlingen in hun brieven moesten noemen. In de taak die binnen deze onderzoekslijn is ontwikkeld, hebben leerlingen de mogelijkheid om een eigen structuur aan te brengen en staat in de instructie voor de leerlingen ook aangegeven dat ze de elementen niet per se in de volgorde van de lijst hoeven te noemen in hun mail.

Tijdens het ontwikkelen van de taak is ook nagedacht over de zorg dat er bij geïntegreerde taken een afhankelijkheid bestaat tussen de receptieve en de productieve vaardigheid, en dus dat leerlingen die de input onvoldoende begrepen hebben hun productieve capaciteiten niet (goed) kunnen tonen. Hiervoor zijn twee mogelijke oplossingen geformuleerd, namelijk ofwel het aanbieden van de leestekst op een lager niveau dan de schrijftaak, ofwel het inbouwen van elementen in de schrijftaak die niet afhankelijk zijn van de leestekst. De eerste oplossing bleek ongeschikt, aangezien een van de doelen van deze taak was om ook leesvaardigheid te toetsen en de aangeboden leestekst dus op niveau van de leerling moest zijn. Om die reden is gekozen voor de tweede oplossing. Van de elf elementen die de leerlingen in hun schrijfpdracht moesten verwerken houden er 4 of 5 elementen verband met de inhoud van de leestekst. De overige elementen konden leerlingen zonder enig begrip van de leestekst verwerken in hun schrijfpdracht.

Het oorspronkelijke plan was dan ook om leesvaardigheid en schrijfvaardigheid los van elkaar te beoordelen, door bij de beoordeling van leesvaardigheid te letten op de elementen die verband houden met de leestekst (de input) en bij de beoordeling van schrijfvaardigheid juist te letten op de elementen die leerlingen konden verwerken zonder begrip van de input. De rationale hiervoor was dat leerlingen bij de beoordeling van de productieve vaardigheid niet benadeeld zouden kunnen worden door onvoldoende begrip van de input. Uiteindelijk is er echter gebruik gemaakt van een beoordelingsmodel dat grotendeels overeenkomt met het beoordelingsmodel dat in een eerder stadium van dit onderzoek (2020-2021) is opgesteld voor de beoordeling van de gecombineerde lees- en schrijfpdracht. Om toch een vernieuwend element toe te voegen aan deze trial ten opzichte van de taak die in 2020-2021 is ontwikkeld, is besloten om deelnemende docenten de schrijfpdrachten van leerlingen uit een andere klas na te kijken.

In dit beoordelingsmodel, dat de beoordelaars via Microsoft Forms invulden, werden beide deelopdrachten apart beoordeeld. Bij de beoordeling van deelopdracht 1 werden beoordelaars gevraagd te letten op de inhoud van de door de leerling geschreven tekst (zijn alle elementen die in de e-mail verwerkt moesten worden aanwezig?), de grammaticale correctheid, de correctheid van spelling en interpunctie, het woordgebruik en de structuur van hun tekst. Grammatica, spelling en interpunctie, woordgebruik en tekststructuur werden globaal beoordeeld, terwijl de beoordeling van de inhoud van de tekst analytisch was. In deelopdracht 2 werden de inhoud en de overtuigingskracht van de geschreven tekst op globale wijze beoordeeld.

4.3 Afname

De taak is afgenomen bij 26 leerlingen uit een klas havo 5. Alle leerlingen hebben voorafgaand aan de afname van de taak een toestemmingsformulier ingevuld, waarin ze toestemming gaven dat hun geanonimiseerde uitwerkingen worden gebruikt binnen het huidige onderzoeksproject.

Van de 26 leerlingen hebben 23 leerlingen een complete uitwerking ingeleverd. In Tabel 2 is te zien voor welke vacatureteksten de leerlingen hebben gekozen in de beide deelopdrachten. In deelopdracht 1, het schrijven van een sollicitatiemail, kozen met afstand de meeste leerlingen voor vacaturetekst 1 (*Servicekraft & Aushilfe für Gastrobetrieb*). In deelopdracht 2 kozen de meeste leerlingen voor vacaturetekst 2 (*Aushilfe für Markt am See & im Außenbereich*), maar was het verschil ten opzichte van de andere twee vacatureteksten kleiner dan bij deelopdracht 1.

Tabel 2. Frequenties gekozen vacatures in deelopdrachten 1 en 2.

Vacaturetekst	Aantal keren gekozen in deelopdracht 1 (e-mail schrijven)	Aantal keren gekozen in deelopdracht 2 (appje schrijven)
1: <i>Servicekraft & Aushilfe für Gastrobetrieb</i>	15	7
2: <i>Aushilfe für Markt am See & im Außenbereich</i>	6	10
3: <i>Hilfe beim Zeltaufbau & Putzkraft</i>	2	6

Evaluatie door leerlingen

Na de afname van de taak hebben de leerlingen een vragenlijst ingevuld, waarin zij hun mening over de taak konden geven. De leerlingen werden gevraagd om hun mening te geven over een aantal stellingen op een schaal met de schaalpunten 'Helemaal mee oneens', 'Mee oneens', 'Niet mee eens/niet mee oneens', 'Mee eens' en 'Helemaal mee eens'. Voor de analyse van deze vragenlijst zijn deze schaalpunten hercodeerd naar een numerieke 5-puntsschaal (1 = Helemaal mee oneens, 2 = Mee oneens, 3 = Niet mee eens/niet mee oneens, 4 = Mee eens, 5 = Helemaal mee eens). Na de gesloten vragen volgden een aantal open vragen, waarin de leerlingen werden gevraagd welke onderdelen van de taak ze al hadden geoefend op school, wat ze moeilijk of onduidelijk vonden aan de taak, wat er verbeterd zou kunnen worden aan de taak en wat ze leuk vonden aan de taak. Tot slot konden de leerlingen een oordeel geven over de taak als geheel op een schaal van 1 ster (zeer negatief oordeel) tot 5 sterren (zeer positief oordeel). 21 leerlingen hebben de evaluatie van de taak ingevuld. De tabellen hieronder geven per stelling de gemiddelde oordelen, inclusief standaarddeviatie, van de leerlingen weer voor deelopdracht 1 en deelopdracht 2. De figuren geven daarnaast een beter inzicht in de spreiding van de antwoorden van de leerlingen dan de tabellen.

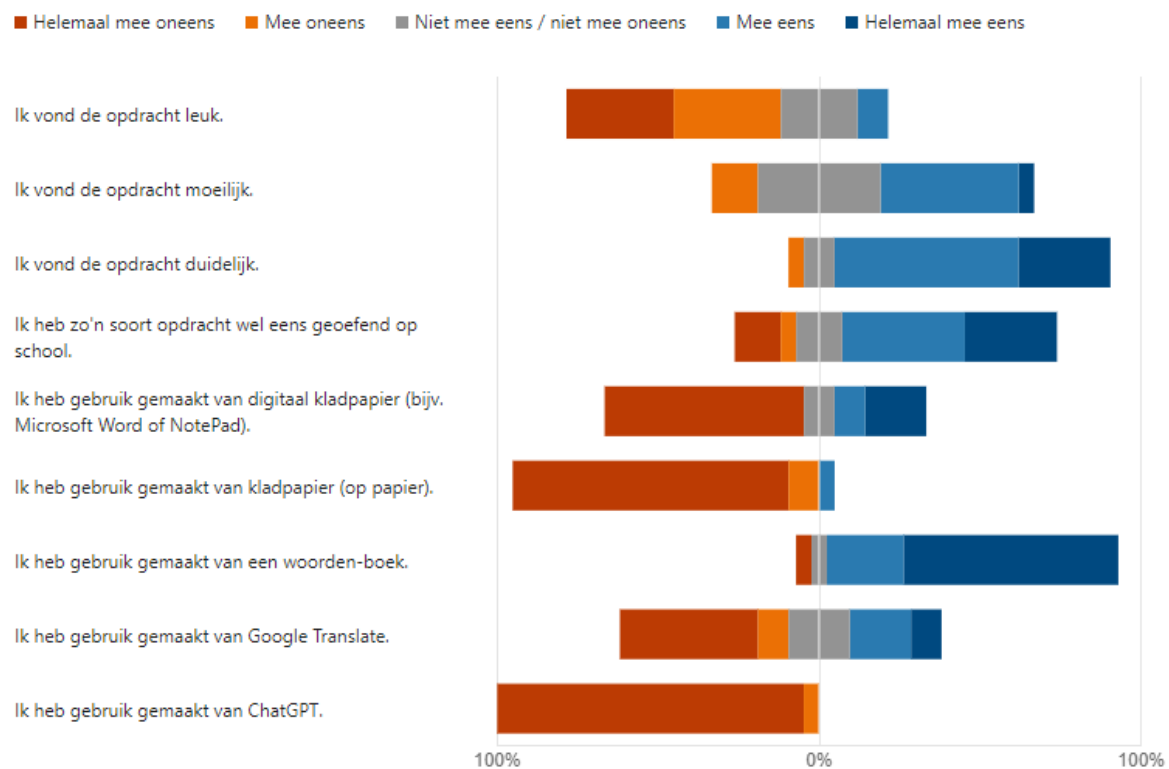
De resultaten in Tabel 3 en Tabel 4 geven aan dat de leerlingen beide opdrachten niet uitzonderlijk leuk vonden om te maken. Wanneer men Figuur 1 en Figuur 2 vergelijkt, wordt echter wel duidelijk dat leerlingen deelopdracht 2 iets positiever evalueerden dan deelopdracht 1. Bovendien vonden de leerlingen deelopdracht 1 moeilijker dan deelopdracht 2, wat waarschijnlijk toe te schrijven is aan het feit dat deelopdracht 1 in de doeltaal (Duits) geschreven moest worden en deelopdracht 2 in het Nederlands. Beide deelopdrachten waren volgens de leerlingen duidelijk. Verder gaven de leerlingen aan weleens op school te hebben geoefend met het schrijven van een sollicitatie (deelopdracht 1), maar niet met het schrijven van een appje (deelopdracht 2).

Op het vlak van hulpmiddelen gaven de meeste leerlingen aan geen kladpapier (fysiek dan wel digitaal) te hebben gebruikt. Een grote groep leerlingen gaf aan voor deelopdracht 1 gebruik te hebben gemaakt van een woordenboek. Ook in deelopdracht 2 gaf een aantal leerlingen aan gebruik te hebben gemaakt van een woordenboek. In deelopdracht 1 maakten ook een aantal leerlingen gebruik van Google Translate, maar dit hulpmiddel werd minder vaak gebruikt dan een woordenboek. In beide deelopdrachten is er nauwelijks gebruikgemaakt van ChatGPT.

De leerlingen gaven de taak als geheel een gemiddeld oordeel van 2,67 sterren ($SD = 1,06$) op een schaal van 1 ster (zeer negatief oordeel) tot 5 sterren (zeer positief oordeel). Hoewel leerlingen de taak als geheel niet erg leuk vonden, gaven enkele leerlingen in de open vragen aan dat ze het leuk vonden dat ze in de tweede deelopdracht een appje moesten schrijven aan een vriend(in). De docent die deze taak heeft afgenomen gaf bovendien de terugkoppeling dat leerlingen het als positief ervoeren dat ze zelf een keuze mochten maken voor een vacature om op te reageren in deelopdracht 1.

Tabel 3. Leerlingevaluatie van deelopdracht 1.

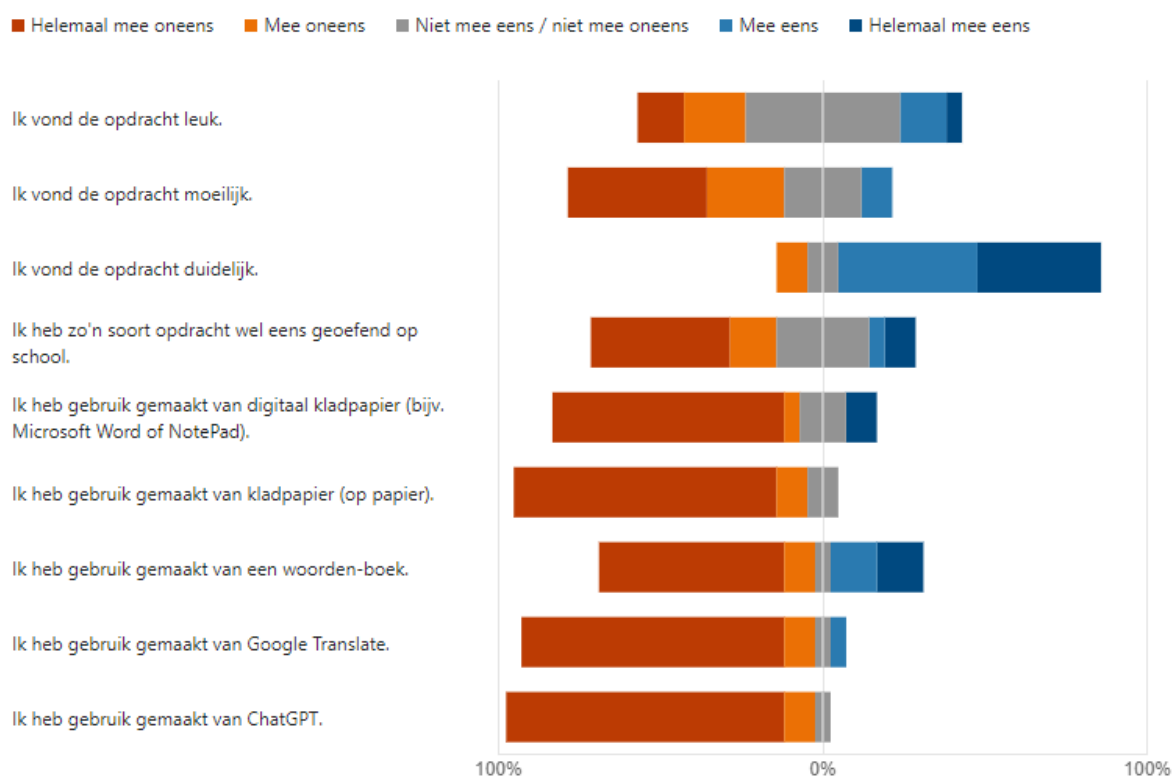
Stelling	Gemiddeld oordeel	Standaarddeviatie
Ik vond de opdracht leuk.	2,10	1,00
Ik vond de opdracht moeilijk.	3,38	0,80
Ik vond de opdracht duidelijk.	4,10	0,77
Ik heb zo'n soort opdracht wel eens geoeffend op school.	3,62	1,36
Ik heb gebruik gemaakt van digitaal kladpapier (bijv. Microsoft Word of NotePad).	2,24	1,70
Ik heb gebruik gemaakt van kladpapier (op papier).	1,24	0,70
Ik heb gebruik gemaakt van een woordenboek.	4,48	0,98
Ik heb gebruik gemaakt van Google Translate.	2,43	1,47
Ik heb gebruik gemaakt van ChatGPT.	1,05	0,22



Figuur 1. Leerlingevaluatie van deelopdracht 1.

Tabel 4. Leerlingevaluatie van deelopdracht 2.

Stelling	Gemiddeld oordeel	Standaarddeviatie
Ik vond de opdracht leuk.	2,76	1,04
Ik vond de opdracht moeilijk.	2	1,05
Ik vond de opdracht duidelijk.	4,10	0,94
Ik heb zo'n soort opdracht wel eens geoefend op school.	2,24	1,34
Ik heb gebruik gemaakt van digitaal kladpapier (bijv. Microsoft Word of NotePad).	1,71	1,31
Ik heb gebruik gemaakt van kladpapier (op papier).	1,29	0,64
Ik heb gebruik gemaakt van een woordenboek.	2,19	1,60
Ik heb gebruik gemaakt van Google Translate.	1,33	0,80
Ik heb gebruik gemaakt van ChatGPT.	1,19	0,51



Figuur 2. Leerlingevaluatie van deelopdracht 2.

4.4 Beoordeling

Voor de beoordeling zijn alle schrijfproducten van de leerlingen geanonimiseerd door identificerende informatie als namen, leeftijden en specifieke informatie (zoals het noemen van bepaalde vakantiebestemmingen waar de leerling heen gaat na de examens) te verwijderen uit de schrijfproducten. Vervolgens zijn de schrijfproducten beoordeeld aan de hand van een beoordelingsmodel door twee beoordelaars, die beiden als constructiegroeplid bij Cito en als docent Duits in de bovenbouw van het voorgezet onderwijs werkzaam zijn. Een van de beoordelaars was ook de docent van de leerlingen die deze opdracht hebben gemaakt. De andere beoordelaar kende de leerlingen niet en geeft enkel les op vwo-niveau. In het beoordelingsmodel waren er punten te verdienen voor zowel de inhoud als de vorm van de schrijfoopdrachten.

Een derde beoordelaar werd ingezet op het moment dat de twee beoordelaars 3 of meer van de in totaal 17 te behalen scorepunten van elkaar afweken. De derde beoordelaar is werkzaam bij Cito als toetsdeskundige Duits en werd als onafhankelijke expert ingezet met als doel om een zo objectief mogelijke beoordeling te realiseren. De derde beoordelaar heeft 8 van de in totaal 23 leerlinguitwerkingen beoordeeld. In één geval viel het oordeel van de derde beoordelaar hoger uit dan dat van de andere beoordelaars. In drie gevallen lag het oordeel van de derde beoordelaar tussen de oordelen van de andere beoordelaars in. In twee gevallen kwam het oordeel van de derde beoordelaar overeen met het oordeel van één andere beoordelaar. In al

deze gevallen werd het eindcijfer van de leerling bepaald door het gemiddelde te nemen van de eindcijfers die iedere beoordelaar met behulp van een omzettingstabel aan de leerling had gegeven. In twee gevallen viel het oordeel van de derde beoordelaar lager uit dan de oordelen van de andere beoordelaars. In deze gevallen is het eindcijfer van de leerling berekend zonder het eindcijfer van de derde beoordelaar daarin mee te nemen, om de leerling niet te benadelen.

De beoordelaarsovereenstemming tussen de eerste en tweede beoordelaar is aan de hand van de intraclass correlation coefficient (ICC) berekend voor beide deelopdrachten apart. De ICC is een getal tussen de 0 en 1, waarbij 0 voor geen overeenstemming staat en 1 voor perfecte overeenstemming. De ICC voor deelopdracht 1 was 0,29, wat voor slechte overeenstemming staat. De slechte overeenstemming bij deelopdracht 1 is voor een groot deel te verklaren vanuit het feit dat de beoordelaars in dit onderzoek niet zijn getraind en gekalibreerd. Zo bereikten de beoordelaars geen overeenstemming over een ogenschijnlijk simpel beoordelingsaspect als 'de leerling maakt duidelijk op welke baan hij of zij reageert'. In dit specifieke geval is de onenigheid tussen de beoordelaars waarschijnlijk te linken aan de manier waarop de vacatureteksten waren opgesteld. Binnen één vacature kon een leerling namelijk altijd kiezen tussen twee verschillende functies. Enkel het noemen van de titel van de vacature was daarom niet voldoende en derhalve kende de ene beoordelaar geen punten toe voor dit aspect, terwijl de andere beoordelaar (ten onrechte) in dit soort gevallen wél een punt toekende voor dit aspect.

Daarnaast gaven de twee beoordelaars op verschillende niveaus les. Terwijl de ene beoordelaar op zowel havo- als vwo-niveau lesgeeft, geeft de andere beoordelaar enkel op vwo-niveau les. In een analyse is bekeken of de vwo-niveau de havo-leerlingen die deze toetstaak hebben gemaakt strenger beoordeeld dan de havo-docent. Dit bleek inderdaad het geval te zijn: in 16 van de 23 gevallen viel de beoordeling van de beoordelaar die op vwo-niveau lesgeeft lager uit dan de beoordeling van de andere beoordelaar. Ook dit zou een verklaring kunnen zijn voor de lage beoordelaarsovereenstemming die is gevonden voor opdracht 1.

De ICC voor deelopdracht 2 was 0,79, wat voor goede overeenstemming staat. Deze hoge overeenstemming bij deelopdracht is opvallend, omdat mediation een relatief nieuw construct is dat nog weinig voorkomt in het mvt-onderwijs. De verwachting was daarom dat docenten moeite zouden hebben met het beoordelen van dit construct. Dit bleek echter veel minder het geval te zijn dan vooraf werd verwacht. Een mogelijke verklaring hiervoor is dat mediation een minder complexe vaardigheid is dan schrijfvaardigheid. Waar schrijfvaardigheid uit verschillende facetten bestaat, zoals inhoud, woordenschat, spelling, grammatica en structuur, is dit voor mediation mogelijk minder het geval, waardoor deze vaardigheid zich wellicht gemakkelijker laat beoordelen.

Tot slot hebben de beoordelaars een korte vragenlijst ingevuld, waarin zij de gehanteerde beoordelingsmethode konden evalueren. Hieruit bleek dat beide beoordelaars positief zijn over het inzetten van twee beoordelaars bij de beoordeling van schrijfproducten. Toch merkten beide beoordelaars op dat het beoordelingsmodel dat voor deze taak is opgesteld te subjectief is, omdat er niet duidelijk genoeg wordt aangegeven wanneer ergens wel of geen punten voor toegekend moeten worden. Deze subjectiviteit is mogelijk te linken aan de erg lage beoordelaarsovereenstemming bij deelopdracht 1.

5. Conclusies

De hoofd- en deelvragen van dit onderzoek waren als volgt:

Hoofdvraag:

- Hoe kunnen leesvaardigheid, schrijfvaardigheid en mediation getoetst worden in een geïntegreerde toetstaak die aansluit op alledaagse situaties en die toepasbaar is op verschillende niveaus (BB/KB/GLTL/HAVO/VWO)?
 - **Deelvraag 1:** Is het mogelijk om een geïntegreerde toetstaak zodanig te ontwerpen dat de input geen belemmering vormt tijdens de schrijfopdracht voor leerlingen die de input niet voldoende hebben begrepen?
 - **Deelvraag 2:** Is het mogelijk om een beoordelingsvorm te bedenken die werkbaar is voor docenten en waarin zowel leesvaardigheid, schrijfvaardigheid en mediation worden beoordeeld?
 - **Deelvraag 3:** Wat vinden leerlingen van de geïntegreerde taak die in dit onderzoek is ontwikkeld?

In dit verslag hebben wij aan de hand van de toetsoperationalisatie en afnamegegevens een taak gepresenteerd die als voorbeeld kan dienen van een geïntegreerde toetstaak, waarin de vaardigheden lezen en schrijven (deelopdracht 1) en mediation (deelopdracht 2) aan bod komen. De taak sluit aan op alledaagse situaties en is toepasbaar op verschillende onderwijsniveaus wanneer andere inputteksten worden gebruikt en de instructies van de schrijf- en mediation-opdrachten zo worden aangepast dat het verwachte product past bij het niveau van de doelgroep.

De schrijftaak in deelopdracht 1 is zodanig ontworpen dat leerlingen die de input, de vacatureteksten, niet hebben begrepen, alsnog hun schrijfvaardigheid kunnen laten zien. Hiertoe zijn er elementen opgenomen die de leerlingen in hun e-mail moeten verwerken die geen direct verband houden met het begrip van de tekst, zoals “Zeg iets over je talenkennis” en “Vraag naar een verblijfsmogelijkheid voor de duur van je contract”. Op deze manier kon de schrijfvaardigheid van leerlingen die de input niet hebben begrepen toch tot op zekere hoogte beoordeeld worden.

De beoordeling van de taak vond ook op geïntegreerde wijze plaats: in het beoordelingsmodel werd namelijk zeer beperkt onderscheid gemaakt tussen de afzonderlijke vaardigheden. Alleen de vaardigheid mediation kon aan de hand van het beoordelingsmodel onafhankelijk beoordeeld worden, terwijl de vaardigheden lezen en schrijven met elkaar verweven zijn in de beoordeling van de inhoud van de sollicitatiemail (deelopdracht 1). Met een vrij eenvoudige aanpassing zouden lees- en schrijfvaardigheid echter ook onafhankelijk van elkaar beoordeeld kunnen worden, door een onderscheid te maken tussen elementen die direct verband houden met de inhoud van de tekst (bijv. “Reageer positief op één taak die volgens de vacature onderdeel is van je werkzaamheden.”) en elementen die dat niet doen (bijv. “Zeg iets over je talenkennis”).

Leerlingen gaven de opdracht een gemiddeld oordeel van 2,67 sterren op een schaal van 1 ster (zeer negatief) tot 5 sterren (zeer positief). Volgens de vragenlijst en de deelnemende docent vonden ze het vooral leuk om te kunnen kiezen tussen verschillende vacatureteksten en om een appje te schrijven in deelopdracht 2, als onderdeel van de toetsing van het construct mediation.

6. Aanbevelingen

Afnamecondities

Wanneer een geïntegreerde toetstaak, zoals de taak die in dit onderzoek is gepresenteerd, in een centraal examen voor de moderne vreemde talen opgenomen zou worden, zijn veel aanpassingen aan het examenstelsel nodig zijn. We geven hier daarom geen uitputtende lijst van benodigde aanpassingen. In plaats daarvan concentreren we ons op drie aanpassingen die deze specifieke toetstaak oproept.

Gezien het *scenario-based* karakter van de taken in deze onderzoekslijn is het essentieel de afnameomgeving zo authentiek mogelijk te maken. De afnameomgeving die voor deze toetstaak is gebruikt (Microsoft Forms), was hiervoor niet optimaal. Mocht een dergelijke *scenario-based* taak worden afgenomen in een centraal examen, dan zou een e-mailomgeving nagebootst moeten worden. Daarnaast is het voor een afname van toets in deze vorm noodzakelijk dat de leerling zo weinig mogelijk externe software of hulpmiddelen hoeft te gebruiken om de taak te kunnen voltooien. Voor de huidige taak zou een geschikte afnameomgeving daarom minimaal moeten beschikken over een ingebouwde woordenteller en wellicht een ingebouwde spellingcontrole.

Ook is het niet zomaar mogelijk om een dergelijke toetstaak op één centraal moment te examineren, doordat iedere leerling over een eigen laptop of pc moet beschikken om de toets te kunnen maken. Flexibele afnames als bij de digitale en flexibele examens vmbo-bb en vmbo-kb bieden hiervoor een mogelijke oplossing.

Beoordeling

Bij de beoordeling van de eerste deelopdracht is geen expliciet onderscheid gemaakt tussen elementen die samenhangen met begrip van de input (en dus samenhangen met de leesvaardigheid van de leerling) en elementen waarvoor begrip van de input niet van belang is (en die dus samenhangen met de schrijfvaardigheid van de leerling). Om beter zicht te krijgen op de leesvaardigheid en de schrijfvaardigheid afzonderlijk van elkaar zou het bij de beoordeling van een dergelijke geïntegreerde taak daarom nuttig zijn om explicieter onderscheid te maken tussen de elementen die wel en de elementen die geen verband houden met begrip van de inputtekst, al is hiervoor ook vervolgonderzoek benodigd (zie *Vervolgonderzoek* verderop).

In dit onderzoek heeft een derde beoordelaar de schrijfproducten beoordeeld in geval van onenigheid, om zo het effect van de individuele beoordelaars te 'dempen'. Om een zo objectief mogelijke beoordeling te bereiken is enkel het inzetten van een derde beoordelaar echter niet voldoende: de beoordelaarsovereenstemming was bij deelopdracht 1 namelijk erg laag. Het is daarom belangrijk om beoordelaars ook vooraf te trainen in de geselecteerde beoordelingswijze. Dit zou naar verwachting de beoordelaarsovereenstemming vergroten (zie ook *Taakdesign* verderop). Ook is het aan te bevelen om beoordelaars te selecteren die in nauw contact staan met de doelgroep. In dit onderzoek bleek de beoordelaar die enkel op vwo lesgeeft namelijk over het algemeen strenger te beoordelen dan de andere beoordelaar, die zowel op havo als vwo lesgeeft.

Voorbereiding

Bij nieuwe toetstaken is het belangrijk om de leerlingen van tevoren goed voor te bereiden, zodat zij bij het maken van de taak al bekend zijn met het type opdracht. In dit geval was vooral deelopdracht 2 een andersoortige schrijfoopdracht dan leerlingen uit de les gewend zijn, zoals bleek uit de leerlingevaluatie. Naast het feit dat het schrijven van WhatsApp-berichten in de les volgens de deelnemende leerlingen niet gebruikelijk is, is mediation een construct dat op scholen nog weinig aandacht krijgt. Mocht deze vaardigheid een grotere plaats gaan innemen in de centrale examens, dan zullen leerlingen daarop in de les goed voorbereid moeten worden.

Taakdesign

In deze taak konden leerlingen kiezen uit drie vacatures, waarin steeds twee vakantiebanen werden gepresenteerd waarop de leerlingen konden solliciteren. In feite waren er dus zes keuzemogelijkheden in plaats van drie keuzemogelijkheden. Uit de leerlinguitwerkingen bleek dat dat niet voor alle leerlingen duidelijk was, omdat niet alle leerlingen ook daadwerkelijk de naam van de functie noemden waarop ze wilden solliciteren. Tijdens de beoordeling leidde dit tot onenigheid tussen de beoordelaars over het beoordelingsaspect 'de leerling maakt duidelijk op welke baan hij of zij reageert'. De ene beoordelaar kende alleen een punt toe voor dit aspect wanneer de leerling de naam van de functie noemde, terwijl de andere beoordelaar al een punt gaf als de leerling de naam van de vacature noemde, zonder daarbij te specificeren op welke

van de twee functies deze leerling reageerde. Dit had een negatieve invloed op de beoordelaarsovereenstemming in dit onderzoek. Om deze redenen is het aan te raden om bij het gebruik van keuzemogelijkheden altijd duidelijk aan te geven waar de leerling keuzes moet maken, of om het aantal keuzemogelijkheden zoveel mogelijk te beperken.

Vervolgonderzoek

In deze taak zijn lees- en schrijfvaardigheid op geïntegreerde wijze beoordeeld. Dit wil zeggen dat beide vaardigheden aan de hand van het gebruikte beoordelingsmodel niet los van elkaar konden worden beoordeeld. Omdat het in onderwijscontexten juist vaak gebruikelijk is om uitspraken te doen over afzonderlijke vaardigheden, zou in een vervolgonderzoek een aanpassing kunnen worden gedaan aan het beoordelingsmodel waarmee de twee vaardigheden onafhankelijk van elkaar worden beoordeeld. Dit zou kunnen door een onderscheid te maken tussen elementen die direct verband houden met de tekst (bijv. “Reageer positief op één taak die volgens de vacature onderdeel is van je werkzaamheden.”), welke de leesvaardigheid van een leerling meten, en elementen die dat niet doen (bijv. “Zeg iets over je talenkennis”), welke gebruikt kunnen worden om de schrijfvaardigheid van een leerling te meten. In dit vervolgonderzoek zou moeten worden bekeken of het aan de hand van de informatie uit een geïntegreerde toetstaak, gecombineerd met een dergelijk beoordelingsmodel, mogelijk is om een betrouwbare en valide uitspraak te doen over het niveau van de lees- en schrijfvaardigheid van de leerling.

Referenties

Nederlandse Taalunie. (2022). *Gemeenschappelijk Europees referentiekader voor talen: Leren, onderwijzen, beoordelen - Aangevulde uitgave.*

<https://taalunie.org/publicaties/216/gemeenschappelijk-europees-referentiekader-voor-talen-leren-onderwijzen-beoordelen>

SLO. (2021). *Niveaus havo/vwo En Du Fa Sp*. <https://www.slo.nl/thema/vakspecifieke-thema/mvt/erk/leraar-vo/niveaus-vo/niveaus-havo-vwo-du-fa-sp/>

Neem voor meer informatie over de specifieke taken en het onderzoek contact op met:
rosa.bartman@cito.nl