Dorst, J.P. van (Manon), s1005881

De superioriteit van het paarsgewijs beoordelen nader onderzocht: verschillen tussen paarsgewijs en individueel beoordelen bij taalvak- en niet-taalvakdocenten

Bachelorwerkstuk Taalbeheersing – LET-NTCB300TB

Begeleider: prof. dr. Spooren, W.P.M.S. (Wilbert)
25-5-2020

Inhoudsopgave

Inleiding 2
Manieren van beoordelen 2
Paarsgewijze beoordeling 3
Welke afwegingen maken docenten bij paarsgewijs beoordelen? 3
Vergelijking paarsgewijze en individuele beoordelingsmethode 5
Onderzoeksvragen 7
Hypotheses 7
Methode 8
Materiaal 8
Participanten 8
Instrumentatie 9
Ontwerp 9
Procedure 9
Verwerking gegevens 10
Resultaten 12
Cijfer 12
Spreiding 14
Motieven voor beoordeling 15
Conclusie en discussie 18
Literatuuropgave 22
Bijlage A 24
Bijlage B 28

Inleiding

Het beoordelen van tekstkwaliteit: het blijft subjectief. En het is problematisch op het middelbaar onderwijs. Want kun je als docent uitleggen waarom de ene leerling een 8 heeft voor zijn essay, terwijl de andere een 7,5 kreeg? Misschien was de desbetreffende docent in een wat betere bui toen hij een 8 toekende aan de ene leerling, terwijl hij op een ander tijdstip of een andere dag wat minder vrolijk was en daarom de andere leerling een 7,5 gaf. De gemoedstoestand van de docent kan van invloed zijn op de beoordeling van een tekst (Wright & Bower, 1992). Vooral het individueel beoordelen van teksten van leerlingen is subjectief (Goossens & Maeyer, 2018). Daarom is er een andere methode op de markt gekomen wat betreft het beoordelen van teksten: *comparative judgement* (ook wel: vergelijkende beoordeling of paarsgewijze beoordeling). Deze methode vergelijkt steeds twee teksten met elkaar waarvan er één als beste tekst wordt uitgekozen. Hierdoor ontstaat er een soort *ranking* van alle teksten waar uiteindelijk een cijfer aan gekoppeld wordt (Bramley, 2008). In deze studie worden onder andere deze twee methodes (paarsgewijze vs. individuele beoordeling) met elkaar vergeleken.

Manieren van beoordelen

Er bestaan veel varianten van beoordelingsmethoden, maar er is een onderscheid te maken in ruwweg twee dimensies. De eerste dimensie is het onderscheid tussen holistisch en analytisch beoordelen. De holistische beoordeling ziet het schrijven als een "alomvattende competentie" waarin de "onderliggende deelvaardigheden zo sterk samenhangen dat het niet gepast is deze apart te beoordelen" (Coertjens, Lesterhuis, Verhavert, Van Gasse, & De Maeyer, 2017, p. 283), terwijl analytisch beoordelen zegt dat tekstkwaliteit uit "verschillende dimensies" (Coertjens et al., 2017, p. 283) bestaat.

De tweede dimensie gaat over absolute vs. vergelijkende beoordelingen. Bij absolute beoordelingen beoordeelt de beoordelaar elke tekst op zichzelf aan de hand van bijvoorbeeld criterialijsten of *rubrics*[1]. Bij de vergelijkende methode gaat de beoordelaar, zoals het woord zelf al zegt, vergelijkend te werk (Coertjens et al., 2017): de beoordelaar vergelijkt steeds twee teksten met elkaar. Paarsgewijze beoordeling is holistisch en vergelijkend van aard (Pollitt, 2012) en de individuele beoordelingen zijn analytisch en absoluut (Goossens & Maeyer, 2018).

[1] *Rubric* = een beoordelingsschema; een document dat de verwachtingen van een opdracht uitdrukt aan de hand van verschillende criteria. De kwaliteitslevels van de verschillende criteria lopen van uitstekend tot onvoldoende (Reddy & Andrade, 2010). Aspecten waarop docenten een tekst kunnen beoordelen zijn bijvoorbeeld: grammatica, organisatie, argumentatie etc.

Paarsgewijze beoordeling

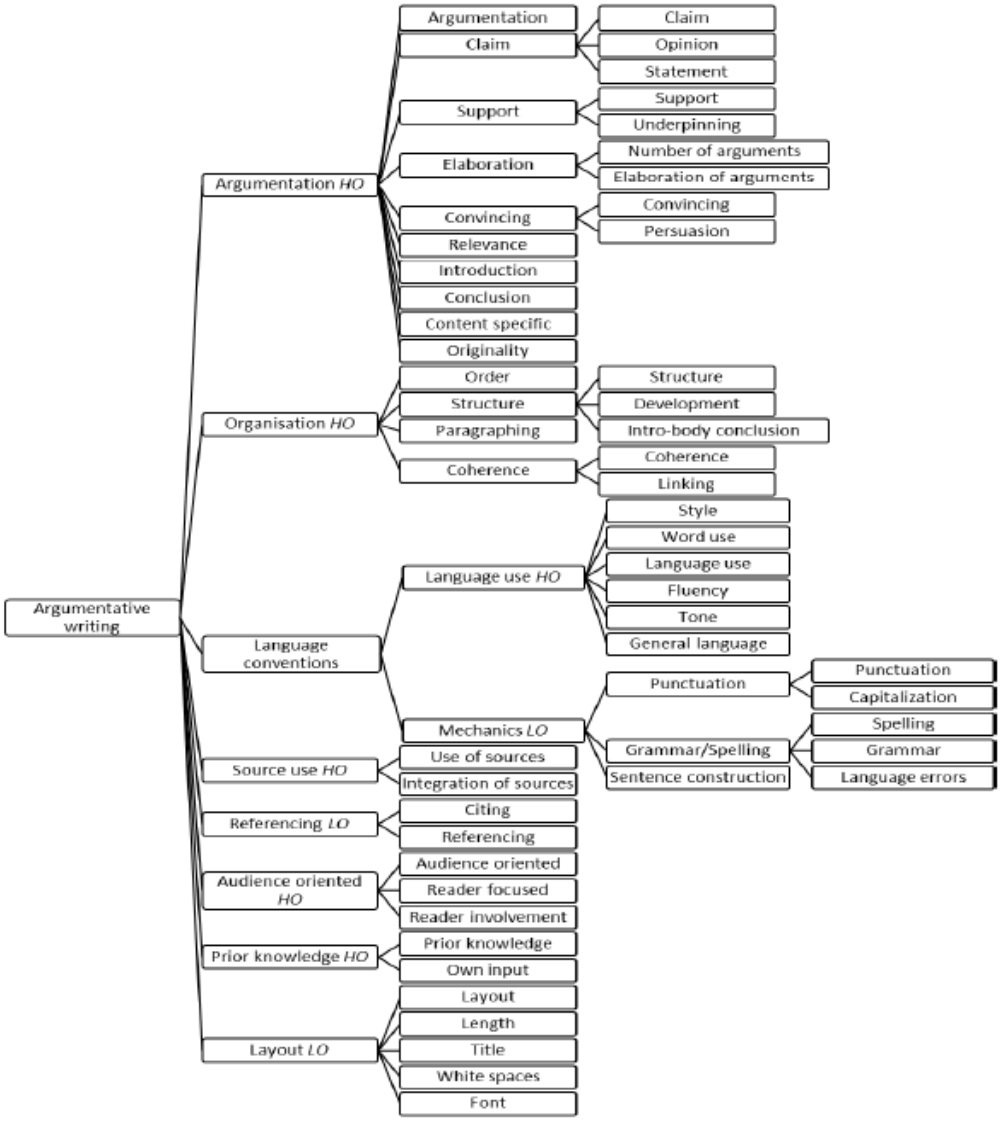
De paarsgewijze beoordeling is geïntroduceerd door de psycholoog Louis Leon Thurstone (Pollitt, 2012). Het paarsgewijs beoordelen van teksten gaat als volgt te werk: de beoordelaars krijgen steeds twee willekeurige teksten te zien waarvan ze de beste moeten kiezen. De teksten hebben een vergelijkbare opdracht (betoog, beschouwing, uiteenzetting etc.) en de beoordelaars mogen steeds zelf de criteria hanteren waarop ze een tekst beoordelen; zij het op spelling, dan wel inhoud, formulering, een combinatie van verschillende criteria etc. Beoordelaars die verstand hebben van beoordelen, bijvoorbeeld docenten, weten natuurlijk welke criteria ze zwaarder moeten laten wegen dan andere. Als ze een keuze hebben gemaakt, moeten ze opnieuw willekeurig twee teksten vergelijken. Op die manier wordt elke tekst meerdere keren met steeds andere teksten vergeleken. Aan de hand van het Bradley-Terry-Lucemodel (vanaf nu: BTL-model) kunnen de vergelijkingen uiteindelijk worden verdeeld in een soort rangorde van teksten van hoge kwaliteit tot teksten van lage kwaliteit (Verhavert, Bouwer, Donche, & Maeyer, 2019). Een BTL-model schat de waarschijnlijkheid dat een willekeurige tekst het wint – dat wil zeggen, beter gevonden wordt – van een andere tekst op grond van een reeks ‘wedstrijden’ tussen alle teksten die vergeleken worden. Het model werd oorspronkelijk gebruikt om rangordeningen van sportteams op grond van sportuitslagen in te schatten (Firth, 2005). De tekst die het vaakst als beste tekst is gekozen tussen alle vergelijkingen, is de tekst van de beste kwaliteit. Omgekeerd is de tekst die het minst als beste tekst is gekozen de tekst van de laagste kwaliteit. Om een betrouwbaarheid van .70 te krijgen moet een tekst 10 tot 14 keer vergeleken worden, 19 tot 20 keer voor een betrouwbaarheid van .80 en 26 tot 37 keer voor een betrouwbaarheid van .90 (Verhavert et al., 2019).

Paarsgewijze beoordeling wordt niet alleen gebruikt voor het beoordelen van teksten van middelbare scholieren, maar wordt ook ingezet om bijvoorbeeld zelfreflectie te beoordelen (Coertjens, Lesterhuis, Goossens, De Maeyer, De Winter & Michels, 2018), om curricila vitae te beoordelen (Mortier, Bouwer, Coertjens, Volckaert, Vrijdag, Van Gasse, Vlerick & De Maeyer, 2019), video’s (Roose, Goossens, Vanderlinde, Vantieghem, & Avermaet, 2018) en instructies van simulatiegames (Settembri, Van Gasse, Coertjens, & De Maeyer, 2018).

Welke afwegingen maken docenten bij paarsgewijs beoordelen?

Lesterhuis, Van Daal, Van Gasse, Coertjens, Donche en De Maeyer (2018) onderzochten de afwegingen die docenten maakten bij het paarsgewijs beoordelen van betogen van middelbareschoolleerlingen. Na elke vergelijking werd de vraag gesteld of de participant kort zijn/haar beslissing kon toelichten. Lesterhuis et al. (2018) maakten bij de toelichting van de

docenten onderscheid tussen complexe aspecten van een hogere orde (zoals de argumentatie en organisatie van de tekst) en aspecten aan de hand van regels die van een lagere orde zijn (zoals spelling en lay-out). Zie voor een overzicht van de indeling van alle aspecten figuur 1 hieronder.



Figuur 1. Ordening van de verschillende aspecten. HO = aspecten van een hogere orde, complex; LO = aspecten van een lagere orde, aan de hand van regels. Uit: Lesterhuis, Van Daal, Van Gasse, Coertjens, Donche & De Maeyer (2018).

Uit het onderzoek bleek dat de participanten een heel uitgebreid spectrum van aspecten hanteerden als ze teksten vergeleken. Dit laat zien dat de beoordeling van teksten aan de hand van de paarsgewijze methode multidimensionaal is. De participanten beoordeelden de betogen met name op de complexere aspecten van teksten (50,17%) of op een combinatie van de complexere aspecten en de aspecten van een lagere orde (46,48%). Niet alle aspecten van de

hogere orde bleken even belangrijk: er werd bijvoorbeeld vooral gelet op de argumentatie en organisatie van de teksten en er werd weinig aandacht besteed aan publiek-georiënteerd schrijven en voorkennis (slechts in 2% van alle beoordelingen), terwijl deze vier aspecten wel allemaal onder de aspecten van een hogere orde vallen. Het kan zo zijn, zo zeggen de auteurs, dat de participanten het publiek-georiënteerd schrijven verwarren met andere componenten van het schrijven – zoals stijl en taalgebruik – en dat de participanten het aspect 'voorkennis' nogal ambigu vinden: verwijst het naar de voorkennis over het onderwerp, over de grammatica of over verwijzingsregels? Om deze redenen nemen participanten aspecten als 'publiek-georiënteerd schrijven' en 'voorkennis' vaak niet mee in de beoordeling.

Lesterhuis et al. (2018) stellen dat er meer training nodig is bij middelbareschooldocenten als het aankomt op de aspecten van voorkennis en publiek-georiënteerd schrijven. Maar de paarsgewijze beoordeling is wel een goede methode om tekstkwaliteit te beoordelen, omdat docenten wel alle aspecten van tekstkwaliteit meenemen in hun beoordeling. Het zou niet goed zijn als de resultaten lieten zien dat de beoordeling van teksten slechts op één aspect is gebaseerd. Overigens is het onderzoek van Lesterhuis et al. (2018) wel erg specifiek: de resultaten gelden alleen voor het beoordelen van argumentatieve teksten.

Vergelijking paarsgewijze en individuele beoordelingsmethode

In deze paragraaf worden twee artikelen besproken die een vergelijking maken tussen paarsgewijs (*comparative*) en individueel beoordelen. Het artikel van Goossens en De Maeyer (2018) gaat over de kwaliteit van beoordelingen en het artikel van Coertjens et al. (2017) bespreekt de verschillen op het gebied van tijdsinvestering.

In de studie van Goossens en De Maeyer (2018) werden twaalf teksten beoordeeld door zes beoordelaars. De beoordelaars waren allemaal leraren in opleiding (master studenten), er werd niet vermeld voor welk vak de participanten aan het studeren waren. Eerst moesten de beoordelaars alle teksten individueel beoordelen aan de hand van een *rubric*. In totaal waren er dus 72 beoordelingen. Drie weken later kregen dezelfde zes beoordelaars de taak om dezelfde twaalf teksten te beoordelen, maar dan via paarsgewijze beoordeling. Elke beoordelaar maakte hierbij twintig vergelijkingen. In totaal waren er dus 120 vergelijkingen. Uit de resultaten bleek dat voor de individuele beoordelingen de beoordelaars ver uit elkaar lagen wat betreft de kwaliteit van de teksten: de standaarddeviatie liep op tot 3.01[2] en de hoogste positieve correlatie

[2] NB: bij dit onderzoek hebben de auteurs voor de beoordelingen een schaal van 1 tot 20 gehanteerd. Dit betekent dat een standaarddeviatie van 1 staat voor een verschil van twee punten.

tussen de scores was .42 (40% van de correlaties was zelfs negatief, wat laat zien dat een tekst door bepaalde beoordelaars heel hoog wordt beoordeeld, terwijl anderen dezelfde tekst als heel laag beoordelen). Bij de paarsgewijze beoordelingen aan de andere hand, werden veel kleinere verschillen gevonden tussen de scores: alle beoordelingen bleven tussen een bepaalde grens (2SD) en alle beoordelaars beoordeelden de teksten min of meer op dezelfde manier. De interbeoordelaarsbetrouwbaarheid voor de individuele beoordelingen was .30, terwijl deze voor de paarsgewijze beoordelingen .84 was. De paarsgewijze beoordelingsmethode is daarom betrouwbaarder om teksten mee te beoordelen dan de individuele methode. Wel was de *spearman rank order* correlatie tussen de scores van de individuele beoordelingen en de scores van de paarsgewijze beoordelingen redelijk hoog: .78. Dit betekent dat de uiteindelijke gemiddelde scores van de individuele beoordelingen en de paarsgewijze beoordelingen niet ver uit elkaar liggen.

Uit het artikel van Coertjens et al. (2017) over tijdsinvestering bleek dat voor beide methodes de benodigde tijd afnam naarmate een beoordelaar al meerdere beoordelingen had gemaakt, maar de afname in benodigde tijd was sterker in de conditie van het individueel beoordelen dan in de conditie van het paarsgewijs beoordelen. Dit houdt in dat er bij het individueel beoordelen minder teksten nodig waren om een afname in de tijdsinvestering te bereiken. Ook bleek uit het onderzoek dat er bij het individueel beoordelen maar twee beoordelingen per tekst nodig zijn om een betrouwbaarheid van .67 te krijgen. Als een tekst vijf beoordelingen kreeg, steeg de betrouwbaarheid naar .85. Bij paarsgewijze beoordeling waren er twaalf beoordelingen per tekst nodig om tot een betrouwbaarheid van .70 te komen en zeventien beoordelingen per tekst voor een betrouwbaarheid van .80. Ondanks dat er meer beoordelingen per tekst nodig waren bij de paarsgewijze methode dan bij de individuele methode om tot dezelfde betrouwbaarheid te komen, hebben de twee methodes wel een vergelijkbare tijdsinvestering. Bij de individuele methode was de tijdsinvestering namelijk 28 minuten en 56 seconden als een tekst vijf keer werd beoordeeld, en dus een betrouwbaarheid van .85 had. Bij de paarsgewijze methode deden participanten gemiddeld 26 minuten en 29 seconden over 25 oordelen per tekst. Dit leverde een betrouwbaarheid van .87 op. Om tot dezelfde betrouwbaarheid te komen, moeten er dus meer beoordelingen per tekst worden gemaakt bij de paarsgewijze methode dan bij de individuele, maar dit levert geen verschil op in tijdsinvestering.

Onderzoeksvragen

Vaak worden onderzoeken naar paarsgewijze en individuele beoordelingen gedaan op het gebied van het verschil in expertise van de beoordelaar (Paré & Joordens, 2008; Verhavert et al., 2019). Veel onderzoek gaat ook over of *peerassessment* (Falchikov, 1995; Topping, 1998; Topping, Smith, Swanson, & Elliot, 2000). Dit onderzoek zal gaan over hoe experts beoordelen. Experts zijn gedefinieerd als "*judges who were experts (in assessment) in the field, and/or who received specific training in the particular (CJ) assessment task*" (Verhavert et al., 2019, p. 548). Hier vallen middelbareschooldocenten ook onder. Het weinige onderzoek dat al ging over experts en beoordelen was vaak met alleen docenten van een en hetzelfde vak (Jones & Alcock, 2014; Paré & Joordens, 2008). Dit onderzoek neemt niet alleen docenten van een en hetzelfde vak mee, maar docenten van verschillende vakken. Niet-taalvakdocenten komen minder in aanraking met het beoordelen van teksten dan taalvakdocenten, maar het is wel van belang dat ook zij goed in staat zijn om een tekst te kunnen beoordelen (denk aan profielwerkstuk, verslagen); dat hoort bij hun didactisch pakket.

De hoofdvraag van dit onderzoek gaat over de verschillen in betrouwbaarheid tussen de paarsgewijze en individuele beoordelingsmethode, met docenten (experts) van verschillende vakken als participanten. Deze vraag luidt:

1. *In hoeverre zijn er verschillen te vinden in de betrouwbaarheid tussen paarsgewijze beoordelingen en individuele beoordelingen?*

De subvragen van het onderzoek zijn:

2. *In hoeverre verschillen taaldocenten en niet-taaldocenten van elkaar in de toekenning van cijfers?*
3. *In hoeverre verschillen taaldocenten en niet-taaldocenten van elkaar in de spreiding van de cijfers?*
4. *In hoeverre verschillen taaldocenten en niet-taaldocenten van elkaar in de motieven voor beoordelen?*

Hypotheses

Bij de hoofdvraag kan de volgende hypothese worden opgesteld:

> *Paarsgewijze beoordeling is betrouwbaarder dan individuele beoordeling. Dat wil zeggen: de variabiliteit zal bij de paarsgewijze beoordelingsmethode lager zijn dan bij de individuele beoordelingsmethode. Deze verwachting gaat terug op het onderzoek van Goossens & De Maeyer (2018).*

Voor de subvragen kunnen er geen hypotheses opgesteld worden, omdat daar nog geen literatuur over bekend is. Wel kan er enigszins een verwachting opgesteld worden voor wat betreft de motieven voor beoordelen:

> *Doordat niet-taalvakdocenten minder in aanraking komen met het beoordelen van teksten dan taalvakdocenten zullen zij meer letten op wat Lesterhuis et al. (2018) lagere-ordecriteria noemen dan taalvakdocenten.*

Methode

Materiaal

In totaal waren er tien te beoordelen teksten. Al deze teksten waren afkomstig uit eerder onderzoek van Verheijen (2018) en waren gemiddeld 152 woorden lang. Vijf van deze teksten zijn geschreven door leerlingen uit 3 atheneum en de andere vijf teksten zijn geschreven door leerlingen uit 3 havo. Er is gekozen voor een klein niveauverschil tussen de teksten, zodat het voor de docenten misschien iets makkelijker werd om de teksten te (vergelijken en te) beoordelen, maar tegelijkertijd dat het verschil in niveau niet overduidelijk merkbaar was en het beoordelen weer te simpel werd. Alle tien de teksten zijn opgenomen in bijlage A.

Participanten

In totaal deden er 75 participanten mee aan het onderzoek. Hiervan waren er 32 niet bruikbaar, omdat deze participanten ofwel niet voldeden aan de eisen (sommigen waren bijvoorbeeld basisschooldocent of helemaal geen docent), ofwel de enquête niet volledig hadden afgerond ofwel het onderzoek niet serieus namen. Dit laatste was te zien aan de manier van beoordelen: een iemand deelde bijvoorbeeld alleen maar tienen uit en een andere participant gaf overal maar heel kort nietszeggend commentaar bij. De dataset bestond dus uit 43 bruikbare resultaten in totaal. Alle participanten waren middelbareschooldocenten tussen de 21 en 68 jaar (M = 40) waarvan er 28 vrouw en 15 man waren. 24 participanten hadden een WO opleidingsniveau en 19 participanten hebben hbo gedaan. Het overgrote deel van de participanten had Nederlands als moedertaal (N = 41).

Van het totale aantal participanten kregen 26 participanten de individuele beoordelingsmethode (M_{leeftijd} = 40). Hiervan waren er 12 taalvakdocenten en 14 niet-taalvakdocenten. De andere 17 participanten kregen de paarsgewijze beoordelingsmethode (M_{leeftijd} = 40). Het aantal taalvakdocenten hiervan was 10 en het aantal niet-taalvakdocenten bedroeg 7. Het totaal aantal taalvakdocenten – de meesten waren docenten Nederlands – was dus 22 en het totaal aantal niet-taalvakdocenten was 21.

Instrumentatie

De individuele beoordelingen verliepen via het programma Qualtrics. Hierin werd gemeten wat voor cijfers de participanten toekenden aan de individuele teksten en op wat voor criteria de participanten de teksten hebben beoordeeld. De paarsgewijze beoordelingen verliepen via het programma Cesar-Lingo. Hierin werd gemeten welke van de twee teksten participanten steeds het beste vonden en op wat voor criteria ze dat hebben gebaseerd. In bijlage B zijn van beide tests screenshots en linkjes opgenomen. Aan de hand van het BTL-model werd er gemeten welke teksten als kwalitatief hoog werden beoordeeld en welke teksten als kwalitatief laag werden beoordeeld. Hieruit kwam een rangorde van lage naar hoge tekstkwaliteit waarin alle tien de teksten zijn opgenomen. Deze relatieve scores zijn omgezet naar een schaal van 1 (extreem laag cijfer) tot 10 (extreem hoog cijfer). Alleen de cijfers van 4 tot 7 werden gebruikt, omdat dit bij de individuele beoordelingen ongeveer het gemiddelde laagste en het gemiddelde hoogste cijfer was.

Om aan de betrouwbaarheid van .70 van Verhavert et al. (2019) te voldoen bij de paarsgewijze methode, kreeg elke participant vijftien vergelijkingen voorgelegd. In totaal waren er dus (17x15=) 255 vergelijkingen.

Voor de individuele beoordeling kreeg iedere participant alle tien de teksten om te beoordelen. Dit leidde tot een totaal van (10x26=) 260 beoordelingen.

Ontwerp

Het experiment had twee onafhankelijke variabelen waarbij er bij allebei sprake was van een *between-subject design*. De eerste onafhankelijke variabele was de methode van beoordelen: er was één groep die de teksten individueel moest beoordelen (groep 1) en de andere groep moest de teksten paarsgewijs beoordelen (groep 2). De tweede onafhankelijke variabele was het "type" docent: aan de ene kant zijn er docenten die een taalvak geven en aan de andere kant zijn er docenten die geen taalvak geven, zoals wiskunde, scheikunde, geschiedenis, biologie etc. Allebei de onafhankelijke variabelen waren op nominaal meetniveau. De drie afhankelijke variabelen waren de scores die docenten toekenden aan de teksten, de bijbehorende spreiding van deze scores en motieven voor beoordelen. Hiervan zijn de eerste twee op ratio meetniveau en de laatste op nominaal meetniveau.

Procedure

De participanten moesten klikken op een link waarna ze random doorverwezen werden naar ofwel het paarsgewijze beoordelingsexperiment ofwel het individuele beoordelingsexperiment. Op deze manier werden de participanten willekeurig ingedeeld bij groep 1 (individuele

beoordelingen) of groep 2 (paarsgewijze beoordelingen). Voorafgaand aan allebei de experimenten kreeg iedere participant een instructie waarin werd vermeld dat de participant op ieder gewenst moment mocht stoppen met deelname aan het experiment en dat de gegevens van de participant anoniem en vertrouwelijk waren.

Eerst kregen de participanten van beide groepen een korte vragenlijst over hun persoonsgegevens, waaronder de vraag stond welk vak de docent gaf. Deze vraag in het bijzonder was van belang, omdat er zo een onderscheid gemaakt kon worden tussen participanten die een taalvak geven en participanten die geen taalvak geven.

Groep 1 kreeg achtereenvolgens in totaal tien teksten te zien. De volgorde van de teksten was per participant gerandomiseerd. De participant moest de teksten individueel beoordelen door aan elke tekst een cijfer toe te kennen van 1 (extreem slechte tekst) tot 10 (extreem goede tekst). Ook werd er na elke tekst van de participant gevraagd om kort een motivatie te geven hoe ze tot hun oordeel waren gekomen (bijv. spelling, argumentatie, stijl, opbouw etc.) en werd er ter controle gevraagd waar de tekst over ging om zeker te weten dat de participant de tekst ook daadwerkelijk las.

Groep 2 kreeg achtereenvolgens vijftien setjes van twee teksten te zien. Per participant was de volgorde van de teksten gerandomiseerd. De participant moest elke keer de beste van de twee uitkiezen. Ook hier werd er bij elk setje teksten van de participant gevraagd om een korte motivatie te geven hoe hij/zij tot zijn/haar oordeel was gekomen en werd ook de controlevraag gesteld waar de tekst over ging.

Het afnemen van het experiment met paarsgewijze beoordeling duurde ongeveer twintig minuten en het experiment met de individuele beoordelingen duurde ongeveer een kwartier. Indien de participanten nog vragen hadden over het experiment, konden ze deze mailen naar j.vandorst@student.ru.nl.

Verwerking gegevens

Uit de resultaten van de paarsgewijze vergelijkingen ontstond uiteindelijk een rangorde: de tekst die het vaakst van de twee willekeurig te vergelijken teksten als beste tekst werd gekozen, stond bovenaan in de rangorde en zo omgekeerd eindigde de tekst die het minst vaak als beste van de twee werd gekozen als de tekst van de minste kwaliteit. Alle overige acht teksten zaten daartussenin. Het gebruikte algoritme leverde een *ability* score en een standaardfout op. Die zijn omgerekend naar een rapportcijfer op een schaal van 4 tot 7 met een bijbehorende standaardfout. De schaal van de cijfers is gebaseerd op de spreiding van de resultaten uit groep 1.

Voor het vaststellen van de verschillen in cijfers en de spreiding van de cijfers, zowel tussen de methodes van beoordelen als tussen de typen docenten, werden er twee *repeated measures* toetsen uitgevoerd en een Pearson correlatie. Bij de eerste *repeated measures* toets werden de cijfers als afhankelijke variabele ingevoerd en het type docent en de beoordelingsmethode als onafhankelijke variabelen. De tweede *repeated measures* toets nam als afhankelijke variabele de spreiding van de cijfers. De onafhankelijke variabelen bleven hetzelfde. Aan de hand hiervan kon er worden onderzocht of er een hoofdeffect van type docent en type beoordelingsmethode was en of er een interactie-effect was tussen die twee variabelen. De Pearson correlatie diende ervoor om te onderzoeken of er een verband was tussen de cijfers van de paarsgewijze en de cijfers van de individuele beoordelingsmethode.

Daarnaast is er per type docent en per type beoordelingsmethode geturfd hoe vaak welke aspecten van beoordelen (hoog, laag, combinatie) werden genoemd. De indeling van deze aspecten is gedaan aan de hand van het schema van het artikel van Lesterhuis et al. (2018) (zie figuur 1 op pagina 4). Sommige commentaren waren makkelijk in te delen, zoals wanneer een docent specifiek bijvoorbeeld spelling, interpunctie of opbouw noemde als toelichting voor het cijfer. Deze staan letterlijk in het schema met bijbehorende orde (laag vs. hoog). Andere commentaren waren lastiger in te delen, bijvoorbeeld wanneer een docent als toelichting gaf dat de tekst (niet) duidelijk was, goed/slecht leesbaar was of wanneer men het had over de formulering. Dit soort commentaren staan niet letterlijk in het schema en waren daarom wat lastiger in te delen. Uiteindelijk zijn dit soort criteria allemaal onderverdeeld in wat Lesterhuis et al. (2018) vloeiendheid, taalgebruik, toon of algemene taal noemen. Deze vallen op hun beurt allemaal onder de aspecten van hoge orde. Wanneer een docent bij een tekst commentaar gaf op zowel aspecten van de lage orde (bijvoorbeeld spelling) als op aspecten van de hoge orde (bijvoorbeeld argumentatie), werd dit gecodeerd als ‘combinatie’.

Deze beoordelingsaspecten zijn met elkaar vergeleken aan de hand van een loglineaire analyse met als factoren type docent, type beoordelingsmethode en type commentaar (hoog, laag, combinatie). Als de drieweginteractie significant bleek, zijn er twee vervolganalyses gedaan. Hierbij werd er voor elk niveau van de variabele type docent een Pearson chi-squaretoets uitgevoerd met als factoren type commentaar en type beoordelingsmethode.

Resultaten

Hieronder zullen de resultaten worden weergegeven voor de drie gemeten variabelen. Achtereenvolgens zijn dat cijfer, spreiding en motieven voor beoordelen.

Cijfer

In tabel 1 hieronder zijn de resultaten voor de cijfers weergegeven. Er zijn geen aanwijzingen dat de data niet homogeen zijn voor zowel de individuele beoordelingsmethode ($F(1,18) = 0.038$, $p = .847$) als voor de paarsgewijze beoordelingsmethode ($F(1,18) = 0.109$, $p = .745$).

Tabel 1. *Gemiddelde cijfers voor type docent en type beoordelingsmethode (SE tussen haakjes)*

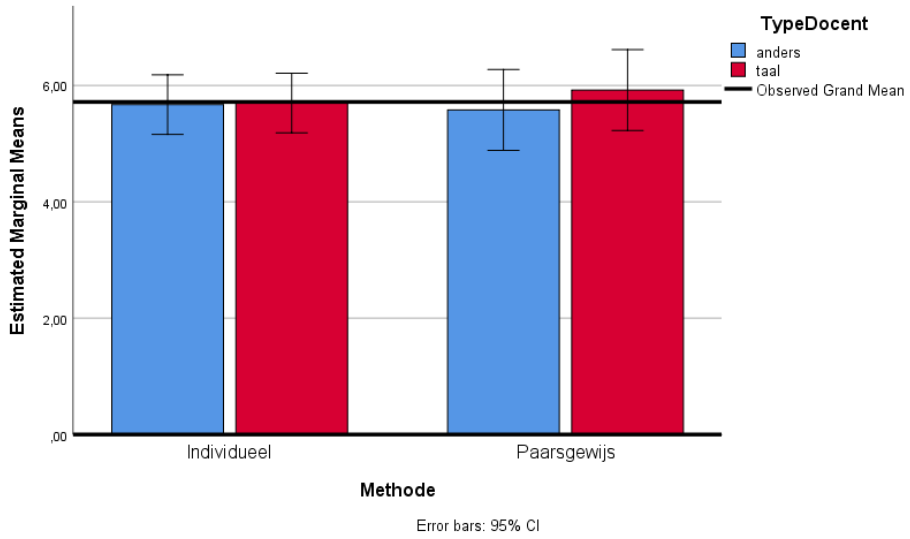
		Methode	
		Individueel	Paarsgewijs
Type docent	Taal	5.70 (0.24)	5.92 (0.33)
	Anders[3]	5.67 (0.24)	5.58 (0.33)

Er is geen significant hoofdeffect gevonden van type docent op het cijfer ($F < 1$). Het gemiddelde cijfer van de niet-taaldocenten was 5.63 ($SE = 0.22$) en het gemiddelde cijfer van de taaldocenten was 5.81 ($SE = 0.22$). Het cijfer valt dus niet hoger of lager uit als een taal- of niet-taaldocent een tekst beoordeelt.

Er is ook geen significant hoofdeffect gevonden van beoordelingsmethode op het cijfer ($F < 1$). Het gemiddelde cijfer van de individuele beoordelingsmethode was 5.69 ($SE = 0.17$) en het gemiddelde cijfer van de paarsgewijze beoordelingsmethode was 5.75 ($SE = 0.23$). Het maakt dus voor het uiteindelijke gemiddelde cijfer niet uit of een docent paarsgewijs of individueel beoordeelt.

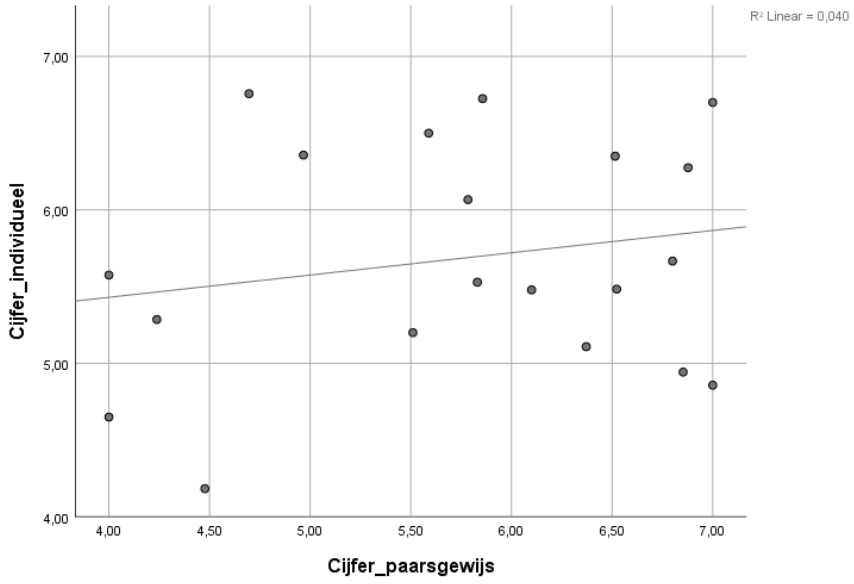
Er is ook geen significant interactie-effect gevonden tussen de methode en het type docent als het gaat om het cijfer ($F < 1$). Het maakt voor het uiteindelijke gemiddelde cijfer niet uit of een taal- of niet-taaldocent een individuele of paarsgewijze beoordeling doet ($M_{\text{taal x ind}} = 5.70$, $SE = 0.24$; $M_{\text{taal x paars.}} = 5.92$, $SE = 0.33$; $M_{\text{niet-taal x ind.}} = 5.67$, $SE = 0.24$; $M_{\text{niet-taal x paars.}} = 5.58$, $SE = 0.33$). De interactie is grafisch weergegeven in figuur 2 op de volgende pagina.

[3] In de gehele resultatensectie worden de niet-taalvakdocenten in de tabellen en figuren aangeduid als ‘anders’.



Figuur 2. Grafische weergave van het interactie-effect van de cijfers tussen het type beoordelingsmethode en het type docent (error bars van 95% CI)

Om te kijken of de twee methodes wel met elkaar correleren is er nog een Pearson correlatie uitgevoerd (zie figuur 3 hieronder). De twee methodes correleren niet goed met elkaar (r (20) = .200, p = .399). Slechts 4% (R^2 = 0.040) van de cijfers van de individuele beoordelingsmethode wordt verklaard door de cijfers van de paarsgewijze beoordelingsmethode. Hoewel de cijfers van de twee verschillende methodes nauwelijks van elkaar verschillen, blijkt er geen samenhang tussen de twee methodes te zijn.



Figuur 3. Scatterplot van de correlatie tussen de paarsgewijze en individuele methode

Spreiding[4]

In tabel 2 hieronder zijn de resultaten voor de spreiding van de cijfers weergegeven. Ook hier zijn er geen aanwijzingen dat de data niet homogeen zijn voor zowel de individuele beoordelingsmethode ($F(1,18) = 0.311$, $p = .584$) als voor de paarsgewijze beoordelingsmethode ($F(1,18) = 0.033$, $p = .859$).

Tabel 2. *Gemiddelde spreiding voor type docent en type beoordelingsmethode (SE tussen haakjes)*

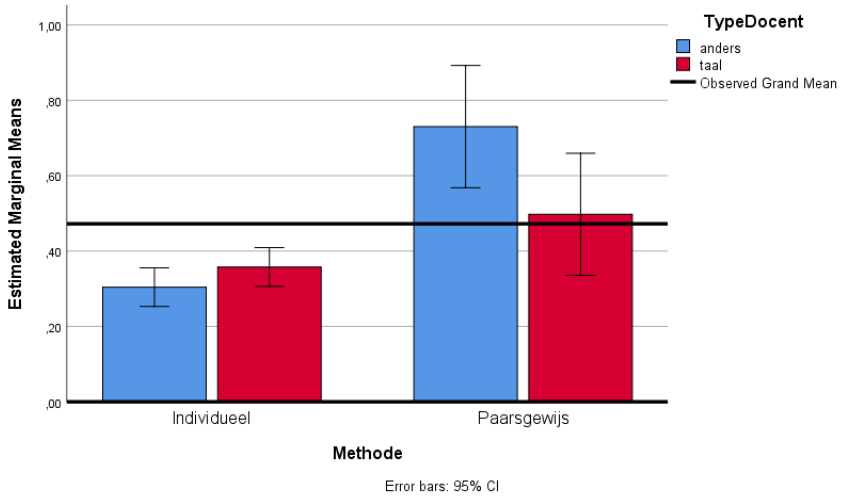
		Methode	
		Individueel	Paarsgewijs
Type docent	Taal	0.36 (0.02)	0.50 (0.08)
	Anders	0.30 (0.02)	0.73 (0.08)

Er is een significant hoofdeffect gevonden van methode op de spreiding van de cijfers ($F(1,18) = 23.479$, $p < .001$, partial $\eta^2 = .566$). Bij de paarsgewijze beoordelingsmethode is er gemiddeld een hogere standaardfout in de cijfers ($M = 0.61$, $SE = 0.06$) dan bij de individuele beoordelingsmethode ($M = 0.33$, $SE = 0.02$).

Er is geen significant hoofdeffect gevonden van het type docent op de spreiding van de cijfers ($F(1) = 2.548$, $p = .128$, partial $\eta^2 = .124$). De standaardfout was bij de niet-taaldocenten 0.52 ($SE = 0.04$) en bij de taaldocenten 0.43 ($SE = 0.04$).

Tot slot is weer wel een significant interactie-effect gevonden tussen de methode en het type docent ($F(1) = 6.017$, $p = .025$, partial $\eta^2 = .251$). Bij de paarsgewijze beoordelingsmethode is er een groot verschil gevonden tussen de standaardfouten van taaldocenten ($M = 0.50$, $SE = 0.08$) en niet-taaldocenten ($M = 0.73$, $SE = 0.08$), terwijl het verschil bij de individuele beoordelingen niet zo groot was ($M_{taal} = 0.36$, $SE = 0.02$; $M_{niet\text{-}taal} = 0.30$, $SE = 0.02$). De interactie is grafisch weergegeven in figuur 4 op de volgende pagina.

[4] Voor de spreiding is overal de standaardfout (SE) gepakt.



Figuur 4. Grafische weergave van het interactie-effect van de spreiding van de cijfers tussen het type beoordelingsmethode en het type docent (error bars van 95% CI)

Motieven voor beoordeling

Om te kijken of er een interactie is tussen type commentaar, type docent en type methode is er een drieweg loglineaire analyse uitgevoerd. De resultaten daarvan zijn te zien in tabel 3 hieronder en op de volgende pagina.

Tabel 3. *Overzicht van de type commentaren per type docent en per type methode met absolute aantallen, percentages en SR-waardes*[5].

				Type commentaar			
Type docent				Hoog	Laag	Combinatie	Totaal
Taal	Type methode	Individueel	Aantal	47	5	68	120
			%	39.1	4.2	56.7	100
			SR	-1.4	-0.9	1.8	
		Paarsgewijs	Aantal	80	11	52	143
			%	55.9	7.7	36.4	100
			SR	1.3	0.8	-1.6	
	Totaal		Aantal	127	16	120	263
			%	48.3	6.1	45.6	100
Anders	Type methode	Individueel	Aantal	34	43	56	133
			%	25.6	32.3	42.1	100
			SR	-2.4	0.3	2.5	
		Paarsgewijs	Aantal	56	29	14	99
			%	56.6	29.3	14.1	100

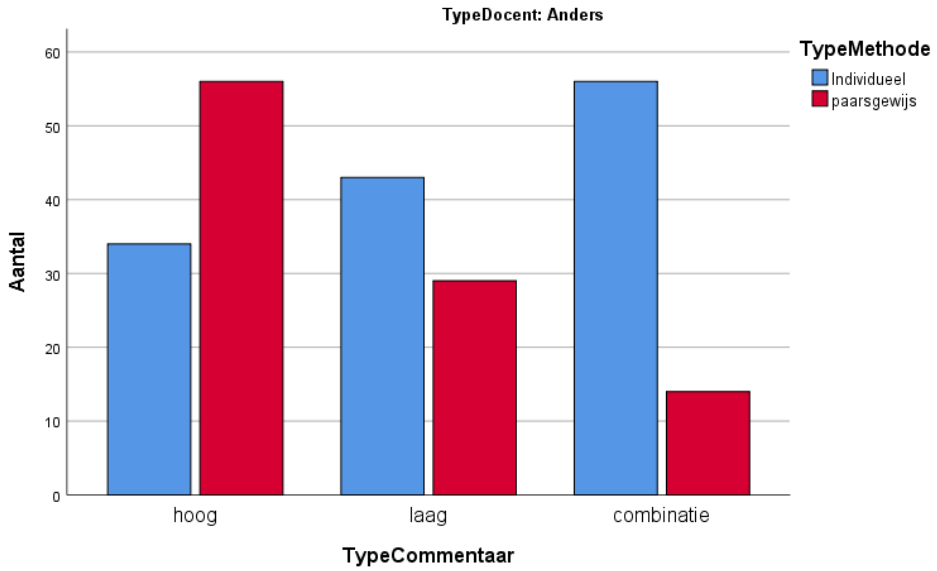
[5] SR = *Standardized Residual*. Geeft de discrepantie aan tussen de geobserveerde en verwachte waarde. Als de SR-waarde (ook wel z-waarde) buiten (-)1.96 ligt, dan draagt de cel significant bij aan de Pearson chi-square ($p < .05$) (Field, 2018, p. 857).

	SR	2.8	-0.3	-2.9	
Totaal	Aantal	90	72	70	232
	%	38.8	31.0	30.2	100

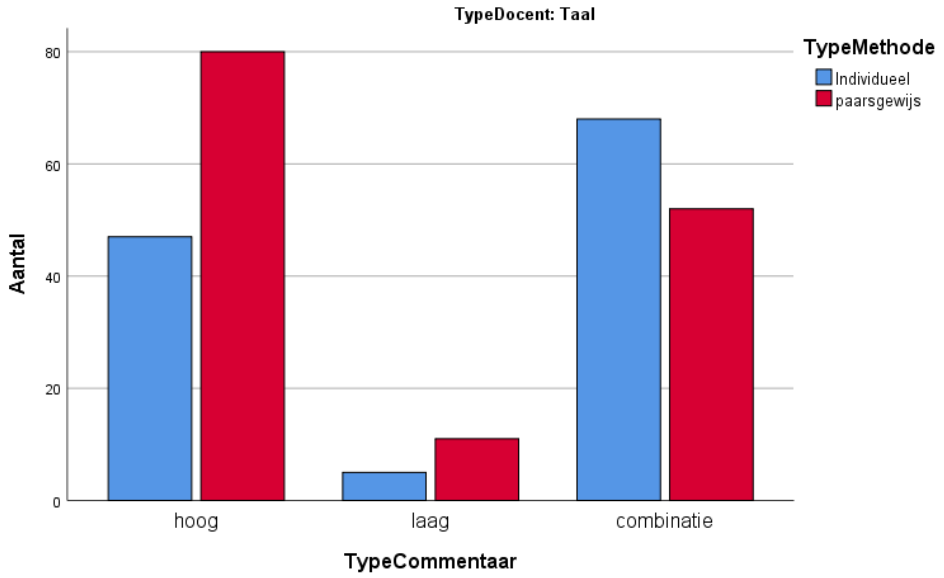
De drieweg loglineaire analyse leverde een eindmodel op dat alle effecten behield. De likelihood ratio van dit model was χ^2 (0) = 0, p = 1. Dit toont aan dat de interactie van de hoogste orde (type docent x type commentaar x type methode) significant was: χ^2 (2) = 7.261, p = .026.

In tabel 3 is te zien dat de verdeling hoog-laag-combinatie bij niet-taaldocenten in het algemeen vrij gelijk is verdeeld (respectievelijk 38.8, 31.0 en 30.2%), terwijl het bij taaldocenten zo is dat de hoge aspecten en de combinatieaspecten ongeveer even vaak werden toegepast (respectievelijk 48.3 en 45.6%) en alleen het lage commentaar relatief minder werd toegepast (6.1%).

Om het effect van de loglineaire analyse verder te analyseren zijn er twee aparte chi-squaretoetsen gedraaid voor het type commentaar en het type methode van enerzijds de taaldocenten en anderzijds de niet-taaldocenten. De grafische weergaves daarvan zijn hieronder in figuur 5 en op de volgende pagina in figuur 6 te zien.



Figuur 5. Grafische weergave van het verband tussen het type commentaar en type methode bij de niet-taaldocenten



Figuur 6. Grafische weergave van het verband tussen het type commentaar en type methode bij de taaldocenten

Bij zowel de taal- als bij de niet-taaldocenten is er een significant verband gevonden tussen het type commentaar en het type methode (taaldocenten: χ^2 (2) = 11.031, p = 0.004; niet-taaldocenten: χ^2 (2) = 28.939, p < .001). De niet-taaldocenten beoordelen bij de individuele methode het minst op alleen hoge aspecten (25.6%) en het meest op een combinatie van hoge en lage aspecten (42.1%), terwijl het patroon bij de paarsgewijze methode juist omgekeerd is: daar maken de niet-taaldocenten het meest gebruik van alleen de hoge aspecten (56.6%) en beoordelen ze het minst op de combinatieaspecten (14.1%). De z-waardes zijn hier ook significant: $z_{ind.xhoog}$ = -2.4; $z_{ind.xcomb.}$ = 2.5; $z_{paars.xhoog}$ = 2.8; $z_{paars.xcomb.}$ = -2.9.

Het verband is minder sterk bij de taaldocenten, maar alsnog wel aanwezig. Bij deze groep is het namelijk ook zo dat er binnen de individuele beoordelingsmethode het meest op een combinatie van de aspecten wordt beoordeeld (56.7%) en minder op alleen de hoge aspecten (39.1%) en dat het patroon omgekeerd is voor de paarsgewijze beoordelingsmethode (hoog: 55.9%; combinatie: 36.4%)[6]. De afzonderlijke z-waardes zijn hier echter allemaal niet significant.

[6]NB: de taaldocenten beoordelen bij beide methodes het minst op de lage aspecten. De hoge/combinatie commentaren worden dus niet het minst gebruikt bij de individuele/paarsgewijze methode, maar beide typen commentaren worden wel *minder* gebruikt dan het type commentaar dat het meest wordt gebruikt.

Conclusie en discussie

De hoofdvraag van dit onderzoek was:

1. *In hoeverre zijn er verschillen te vinden in de betrouwbaarheid tussen paarsgewijze beoordelingen en individuele beoordelingen?*

De subvragen luidden als volgt:

2. *In hoeverre verschillen taaldocenten en niet-taaldocenten van elkaar in de toekenning van cijfers?*
3. *In hoeverre verschillen taaldocenten en niet-taaldocenten van elkaar in de spreiding van de cijfers?*
4. *In hoeverre verschillen taaldocenten en niet-taaldocenten van elkaar in de motieven voor beoordelen?*

Om dat te onderzoeken is aan taalvakdocenten en niet-taalvakdocenten gevraagd om teksten van middelbarescholieren paarsgewijs of individueel te beoordelen en om steeds een korte motivatie te geven voor hun oordeel.

De hypothese bij de hoofdvraag was dat paarsgewijs beoordelen betrouwbaarder zou zijn dan individueel beoordelen. Uit het onderzoek kwam echter het tegenovergestelde: de variabiliteit lag bij de paarsgewijze beoordelingsmethode verder uit elkaar dan bij de individuele methode. Dit effect was significant en hiermee kan de hypothese die was opgesteld voor de hoofdvraag worden verworpen. Het gevonden resultaat gaat tegen het onderzoek van Goossens en De Maeyer (2018) in, die juist stelden dat paarsgewijs beoordelen een betrouwbaardere methode is dan individueel beoordelen. Wel was hun onderzoeksmethode iets anders: waar zij een *within-subjects design* gebruikten – alle participanten moesten dezelfde teksten via zowel de paarsgewijze als de individuele methode beoordelen met drie weken tijdsverschil ertussen – heeft dit onderzoek gebruik gemaakt van een *between-subjects design*: er waren twee aparte groepen waarbij de ene groep de teksten individueel moest beoordelen en de andere groep paarsgewijs. Ook maakten Goossens en De Maeyer (2018) gebruik van *rubrics* voor de individuele beoordelingen, wat in dit onderzoek niet gedaan was. Dit was een bewuste keuze om de participanten niet te ‘beperken’ in hun motivatie bij hun oordeel. Het tegengestelde effect van de spreiding is extra opvallend, omdat het BTL-model altijd aan één soort ‘team’ de *ability* score van 0 toekent. In dit geval is het zo dat bij allebei de groepen docenten er één tekst is die de score 0 heeft. Deze teksten hebben dan ook als standaardfout 0. Bij de paarsgewijze methode zijn er in dit geval dus twee standaardfouten die de waarde 0 hebben, terwijl er bij de

individuele beoordelingen geen enkele tekst is die standaardfout 0 heeft. Dit is opvallend, omdat je zou verwachten dat de gemiddelde spreiding bij de paarsgewijze methode door die nullen juist omlaag gehaald zou worden en dus kleiner zou zijn dan de individuele methode, wat niet het geval is. De berekening van de *ability* scores is wel een beetje een *black box*; we weten niet goed hoe deze wordt uitgevoerd. Misschien dat het verschil in de spreiding bij de paarsgewijze beoordelingen van daaruit te verklaren is.

Een verwachting die was opgesteld bij de subvragen was dat niet-taalvakdocenten meer zouden letten op de aspecten van de lagere orde dan de taalvakdocenten. Uit het onderzoek bleek dit ook zo te zijn: bij de niet-taalvakdocenten was 31.0% van alle oordelen op lage beoordelingscriteria, tegenover 6.1% bij de taalvakdocenten. Er bleek bovendien een significante interactie te zijn tussen type methode, type docent en type commentaar. Dit is verder uit elkaar gehaald met afzonderlijke chi-squaretoetsen en daaruit bleek ook dat er zowel bij taal- als bij niet-taaldocenten een significant verband was tussen het type methode en het type commentaar. Het verband is sterker bij de niet-taaldocenten, maar bij beide groepen is er wel ongeveer hetzelfde patroon gevonden: bij de individuele beoordelingen wordt er minder/het minst beoordeeld op alleen hoge aspecten en het meest op de combinatie van hoge en lage aspecten, terwijl er bij de paarsgewijze beoordelingen minder/het minst op een combinatie van hoge en lage aspecten wordt beoordeeld en het meest op alleen hoge aspecten. Dit laatste is deels in tegenstrijd met het onderzoek van Lesterhuis et al. (2018). Zij vonden namelijk naast dat er bij paarsgewijze beoordelingen veel gelet werd op aspecten van de hogere orde er ook veel gelet werd op een combinatie van de hoge en lage criteria. Dit laatste is niet uit dit onderzoek gebleken; er werd juist weinig op een combinatie van hoge en lage criteria gelet bij de paarsgewijze beoordelingen. Aan de hand hiervan valt er te twisten over de superioriteit van de paarsgewijze beoordelingsmethode, want is het wel goed om alleen op hoge criteria te letten? Als een docent een tekst beoordeelt is het toch ook van belang om op de lage criteria te letten, zoals spelling, verwijzingen en zinsconstructies. En hoe het er nu naar uitziet stimuleert de paarsgewijze methode alleen het geven van hogere-ordecriteria.

Daarnaast bleek er nog een significant interactie-effect te zijn tussen het type docent en het type methode op de spreiding van de cijfers: de standaardfouten verschillen veel tussen taal- en niet-taaldocenten in de paarsgewijze methode, maar bij de individuele methode verschillen ze niet zo veel. Dit laat opnieuw zien dat het paarsgewijs beoordelen misschien toch niet zo'n goede methode is om een oordeel te vellen over teksten.

Tot slot bleek er geen significant hoofdeffect te zijn tussen spreiding en het type docent en waren er helemaal geen significante effecten te vinden voor de toekenning van cijfers, zowel

voor type methode, als voor type docent, als voor de interactie tussen die twee. De kleine verschillen in cijfers zijn niet te verklaren vanuit een correlatie: de twee methodes correleren slecht met elkaar. Ook dit is in tegenspraak met het onderzoek van Goossens en De Maeyer (2018), waaruit bleek dat de correlatie tussen de twee methodes juist heel hoog was. Dat de cijfers zo dicht bij elkaar liggen moet dus met iets anders te maken hebben dan met het type beoordelingsmethode. De berekening van de cijfers bij de paarsgewijze beoordelingen kan een verklaring voor de kleine verschillen zijn. Deze cijfers zijn namelijk gebaseerd op het minimale en maximale cijfergemiddelde van de individuele beoordelingen. Het minimale gemiddelde bij de individuele beoordelingen was 4.18 en voor de berekening van de scores van de paarsgewijze beoordeling is als minimum cijfer 4.0 gepakt. Het maximale gemiddelde bij de individuele beoordelingen was 6.73 en daarvoor is bij de paarsgewijze beoordeling als maximum 7.0 gepakt. De *range* tussen minimum en maximum ligt dus bij beide methodes erg dicht bij elkaar, wat er misschien voor heeft gezorgd dat de uiteindelijke gemiddelde cijfers van beide methodes ook dicht bij elkaar liggen.

Natuurlijk zijn er ook wat haken en ogen aan dit onderzoek. Allereerst zijn er maar weinig verschillende teksten gebruikt (tien in totaal), met allemaal een betogend genre. Dit komt de generaliseerbaarheid van het onderzoek niet ten goede. Vervolgonderzoek zou kunnen uitwijzen in hoeverre de resultaten ook generaliseerbaar zijn naar andere genres, zoals uiteenzettingen.

Ook waren de tekstjes allemaal vrij kort, waardoor het voor de participanten moeilijk kon zijn om er een goed oordeel over te vestigen. Sommige participanten gaven ook expliciet aan dat ze het moeilijk vonden om een oordeel over de tekst te vellen, omdat ze niet wisten waar de opdracht op beoordeeld moest worden of dat het niet duidelijk was wat de criteria waren waaraan de tekst moest voldoen. Dit stond wel in de instructies vermeld, maar misschien had dat nog beter benadrukt moeten worden.

Daarnaast is de verdeling van participanten over de paarsgewijze en individuele beoordelingsmethode niet gelijk; er waren meer participanten bij het experiment met de individuele methode. Hoe dit is gekomen, is niet helemaal duidelijk. Zoals in de methodesectie onder ‘procedure’ is vermeld, werden de participanten random doorverwezen naar ofwel het paarsgewijze ofwel het individuele beoordelingsexperiment. Daarbij is ook ingesteld dat er naar beide experimenten ongeveer evenveel participanten doorverwezen moesten worden. Hier is toch iets bij misgegaan, wat de verdeling oneerlijk maakt. De kans bestaat dat er daarom een enigszins vertekend beeld ontstaat voor de resultaten. In het vervolg zou deze verdeling dus gelijker moeten zijn om dat uit te kunnen sluiten.

Verder is er door maar één persoon een indeling gemaakt in hoge, lage en combinatiecriteria. Dit brengt de betrouwbaarheid van de indeling in het geding. In het vervolg zou er minstens nog één iemand extra moeten kijken naar de indeling van de commentaren.

Desalniettemin zijn er uit dit onderzoek verrassende uitkomsten gekomen die de superioriteit van het paarsgewijze beoordelen op zijn minst ter discussie stellen. Dus voordat we de paarsgewijze beoordelingsmethode helemaal de hemel in prijzen, moet er eerst nog goed wat onderzoek gedaan worden of dit wel echt dé methode is om teksten mee te beoordelen.

Literatuuropgave

Bramley, T. (2008). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). Londen: Qualifications and Curriculum Authority.

Coertjens, L., Lesterhuis, M., Goossens, M., Maeyer, S. de, Winter, B. de, & Michels, N. (2018). Assessing self-reflections in medical education using rubrics or using Comparative Judgement. *Manuscript Submitted to Medical Education.*

Coertjens, L., Lesterhuis, M., Verhavert, S., Gasse, R. van, & Maeyer, S. de. (2017). Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering. *Pedagogische Studiën*, *94*(4), 283–303.

Falchikov, N. (1995). Peer feedback marking: developing peer assessment. *Innovations in Education and Training International*, *32*(2), 175–187. https://doi.org/10.1080/1355800950320212

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Londen: Sage Publications.

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software*, 1–10.

Goossens, M., & Maeyer, S. de. (2018). How to obtain efficient high reliabilities in assessing texts: rubrics vs comparative judgement. In E. Ras & A. E. G. Roldán (Eds.), *Proceedings of Communications in Computer and Information Science: Technology Enhanced Assessment* (pp. 13–25). Cham: Springer.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Lesterhuis, M., Daal, T. van, Gasse, R. van, Coertjens, L., Donche, V., & Maeyer, S. de. (2018). When teachers compare argumentative texts. Decisions informed by multiple complex aspects of text quality. *L1-Educational Studies in Language and Literature*, *18*, 1–22.

Morier, A. V., Bouwer, R., Coertjens, L., Volckaert, E., Vrijdag, A., Gasse, R. van, Vlerick, P., & Maeyer, S. de. (2019). De comparatieve beoordelingsmethode voor een betrouwbare en valide cv screening: een vergelijking tussen experts en studenten. *Gedrag & Organisatie*, *32*(2), 86–107.

Paré, D. E., & Joordens, S. (2008). Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool. *Journal of Computer Assisted Learning*, *24*, 526–540. https://doi.org/10.1111/j.1365-2729.2008.00290.x

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448. https://doi.org/10.1080/02602930902862859

Roose, I., Goossens, M., Vanderlinde, R., Vantieghem, W., & Avermaet, P. van. (2018). Measuring professional vision of inclusive classrooms through video-based comparative judgement: an expert study. *Studies in Educational Evaluation, 56*, 71–84.

Settembri, P., Gasse, R. van, Coertjens, L., & Maeyer, S. de. (2018). Oranges and apples? Using comparative judgement for reliable briefing paper assessment in simulation games. In P. Bursens, V. Donche, D. Gijbels, & P. Spooren (Eds.), *Simulations of Decision-Making as Active Learning Tools. Professional and Practice-based Learning* (pp. 93–108). Cham: Springer.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25*(2), 149–169. https://doi.org/10.1080/713611428

Verhavert, S., Bouwer, R., Donche, V., & Maeyer, S. de. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027

Verheijen, L. (2018). *Is textese a threat to traditional literacy ? Dutch youths' language use in written computer-mediated communication and relations with their school writing*. Utrecht: LOT.

Wright, W. F., & Bower, G. H. (1992). Mood effects on subjective probability assessment. *Organizational Behavior and Human Decision Processes, 52*, 276–291.

Bijlage A

Tekst 1; atheneum 3

Respect voor content creators.

Illegaal downloaden, wie doet/deed het niet? Ik deed het, maar sinds kort ben ik ook zelf bezig met het maken van ‘content’. Hierom vind ik het heel hypocriet om iets te downloaden (niet te betalen) terwijl je wil dat andere voor jou ‘context’ wel betalen.

Vooral in de muziekindustrie is dit een groot probleem omdat je bijna geen albums op CD meer kan verkopen. Dit geld ook voor films, steeds minder mensen gaan nog naar de bioscoop en gebruiken liever Netflix of uTorrent (om het te downloaden)

De derde industrie waarin dit een probleem in is is de videogame industrie, als je bijvoorbeeld ‘GTA 5’ intypt op een torrent site krijg je duizende resultaten, een paar minuten later zit de gebruiker gratis ‘GTA 5’ te spelen.

Er word wel moeite gedaan om bijvoorbeeld illegaal downloaden m.b.v. torrents te stoppen maar dat heeft tot nu toe nog nooit gewerkt. Daarom vind ik ook dat er vanuit de Nederlandse regering of de VN eens wat hieraan gedaan moet worden.

Hierom vind ik dat je niet illegaal moet downloaden maar respect moet hebben voor content creators.

Tekst 2; atheneum 3

de invloed van gewelddadige games

Ik denk dat gewelddadige games geen tot weinig invloed heeft op het gedrag van mensen. Ik vind het maar een fabeltje. Een race spel maakt me geen racer Een spel over politiek maakt me geen advocaat. en een spel over boksen maakt me geen bokser. dus een gewelddadig spel maakt me ook geen gewelddadig persoon. De reden dat ik denk dat mensen denken dat gewelddadige spellen invloed hebben op mensen is omdat mensen misschien beter leren hoe ze voor zichzelf op leren komen en voor zichzel uit durven komen Het kan zijn dat dat als gewelddadig opgevat kan worden, ook al is dat niet zo, ook laten gewelddadige spellen tenminste wat van de echte, harde wereld zien i.p.v. alle leuke en gezellige dingen. Juist daardoor denk ik dat gewelddadige spellen je eigenlijk alleen maar beter voorbereiden op de koude, harde wereld. Het laat je het leven anders inzien en ik denk ook dat je erdoor andere keuzes gaat maken. Ik ben het compleet oneens met het fabeltje van gewelddadige spellen met een gewelddadige invloed.

Tekst 3; atheneum 3

'Alcohol voor je 18e: ok of niet?'

Volgens mij drinken veel jongeren al voor hun 18e. Ik vind alcohol na je 18 een goede regel. Want veel jongeren kennen alle risico's die er zijn als je alcehol drinkt niet. Het is heel slecht voor je hersenen en je houdt er blijvende schade aan over. Ook kun je bijvoorbeeld een alcohol vergiftiging krijgen. Op een jonge leeftijd denk je daar niet goed over na. Ik zelf heb wel eens alcohol gehad. Met vriendinnen hadden we toen allemaal een glaasje gehad. Ik denk dat sommige jongeren ook drinken door groepsdruk. Omdat ze denken dat het 'stoer' is. Je hoort wel eens van die verhalen dat jongeren in het ziekenhuis terecht zijn gekomen door te veel alcohol. Ik vind dat wel echt te ver gaan. Maar opzich vind ik ook een glaasje alcohol op je 16e of 17e niet zo'n groot probleem.

Tekst 4; atheneum 3

Het downloaden van films of muziek?

ik heb gekozen voor illegaal downloaden van muziek en films omdat in deze tijd iedereen films en muziek kijkt of luistert. Ik vind dat het moet kunnen omdat iedereen het zo veel kijkt of luistert en als je daar voor zou moeten betalen zullen aardig wat mensen blut raken. en je kan nu alles online dus niemand gaat meer naar de winkel om een cd of Dvd te kopen. En er zijn wel legale manieren om aan muziek of films te komen zoals spotify of netflix. Maar daar moet je voor betalen terwijl daar nog niet eens alles opstaat. Als je het via een illegale manier doet bijvoorbeeld popcorntime voor films kan je het voor niks downloaden en staan er bijna alle films op. Dus het is voordeliger en je hebt meer. tenzij je een boete krijgt dan kan je wel beter kiezen voor de legale manier.

Tekst 5; atheneum 3

gewelddadige games heeft weinig invloed

Ik vind zelf dat het geen of weinig invloed op me heeft maar op mijn neefje van 3 heb ik dat wel gemerkt want hij speelt nu soms gta en loopt constant te zeggen dat hij je gaat doodschieten of doodmaken of slaan of auto's gaat stelen maar wel alleen met de familie niet op straat, Bij mij heeft het waarschijnlijk ook wel een beetje invloed maar niet zo veel of niet zo duidelijk vind ik. het heeft wel invloed op je maar niet zoveel als mijn ouders zeggen. nou ja op mijn leeftijd dan voor kleine kindjes wel. zoals mijn neefje. ik speel/speelde veel vechtspellen maar in het echt vecht ik nooit dus ja het licht ook wel aan hoe je zelf bent denk ik als je zelf

geweldadig bent wekken die games dat in je op maar als je zelf niet gewelddadig bent dan heeft het weinig invloed.

Tekst 6; havo 3

Alcohol voor je 18e: oké of niet?

Ik vind het wel oke om onder je 18e alcohol te drinken, want het is je eigen keuze. en dan is het ook je eigen schuld als het fout gaat. Maar je moet niet teveel drinken en verslaafd raken want dan is het slecht voor je. je kan dan bijvoorbeeld alcohol vergiftiging krijgen. aan de andere kant vind ik het ook niet oke, want als je dronken bent kun je andere mensen in gevaar brengen en dat is natuurlijk niet oke. maar als je over het algemeen gewoon een paar pilsjes neemt, Lijkt het me gewoon oke om onder je 18e alcohol te drinken. als je er maar niet teveel van drinkt.

Tekst 7; havo 3

'Alcohol voor je 18e: oké of niet?'

Het drinken van alcohol voor je 18e lijkt voor mij geen probleem. Het ligt een beetje aan de hoeveelheden die je drinkt en of het sterke drank is. Mij lijkt het onverantwoord als je voor je 18e al grote hoeveelheden (sterke) drank drinkt. Ik vind ook dat je ouder er dan van op de hoogte moeten zijn. Als je op jonge leeftijd al begint met grote hoeveelheden alcohol drinken, is de kans om er aan verslaafd te raken heel groot. Op een gegeven moment kun je niet meer zonder. Je zult dan heel veel moeite moeten doen om af te kikken.

Dus, alcohol drinken voor je 18e lijkt mij geen probleem onder enkele voorwaarden:
- Je ouders moeten er vanaf weten.
- matig de hoeveelheid, niet te veel alcohol.
- houd de sterke drank in de gate. Met sterke drank kun je wel beter wachten tot je 18e.

Tekst 8; havo 3

De invloed van gewelddadige games

Gewelddadige games zoals GTA of Call of Duty vind ik leuker dan Fifa of een soort gelijk spel. Ik vind dat iedereen het moet kunnen spelen, er hoeft geen leeftijdslimiet op, zolang je het maar niet na gaat doen. Van gewelddadige games leer je ook hoe je problemen NIET moet oplossen, vind ik. Bij gewelddadige games gaat je geduld en reactiesnelheid ook steeds een beetje omhoog. In gewelddadige games doe je meestal dingen die je in de echte wereld nooit zou doen. Ook zijn er momenten die niet normaal zijn en je ze dus ook nooit in de echte wereld ziet. Ik

vind dat iedereen, hoe oud je ook bent, gewelddadige games mag spelen zolang je het niet na doet.

Tekst 9; havo 3

Alcohol voor je 18

Ik vind dat het je eigen keuzen is om te drinken voor je 18. Ik hoor van sommige mensen dat ze wel drinken voor hem 18e. Ook al zou het van hem ouders niet mogen zou ik er niks aan doen.

Ik hoor ook best vaak van anderen dat ze hebben gedronken of gaan drinken. Ik ga waarschijnlijk zelf ook een keer drinken voor mijn 18 maar niet te veel. En vooral als je hoort dat je hersencellen erdoor kapot gaan.

Waarschijnlijk heeft iedereen al gedronken voor zijn 18e.

ik vind het goed dat de overheid het heeft verhoogd naar 18 want de meeste zitten nog op school als ze 16 zijn.

Ik vind niet dat ze alcohol moeten verbieden, want dan zouden mensen illegaal maken (zoals drug) en ik vind het voor mezelf goed als ik soms drink.

Tekst 10; havo 3

Illegaal downloaden van muziek en films

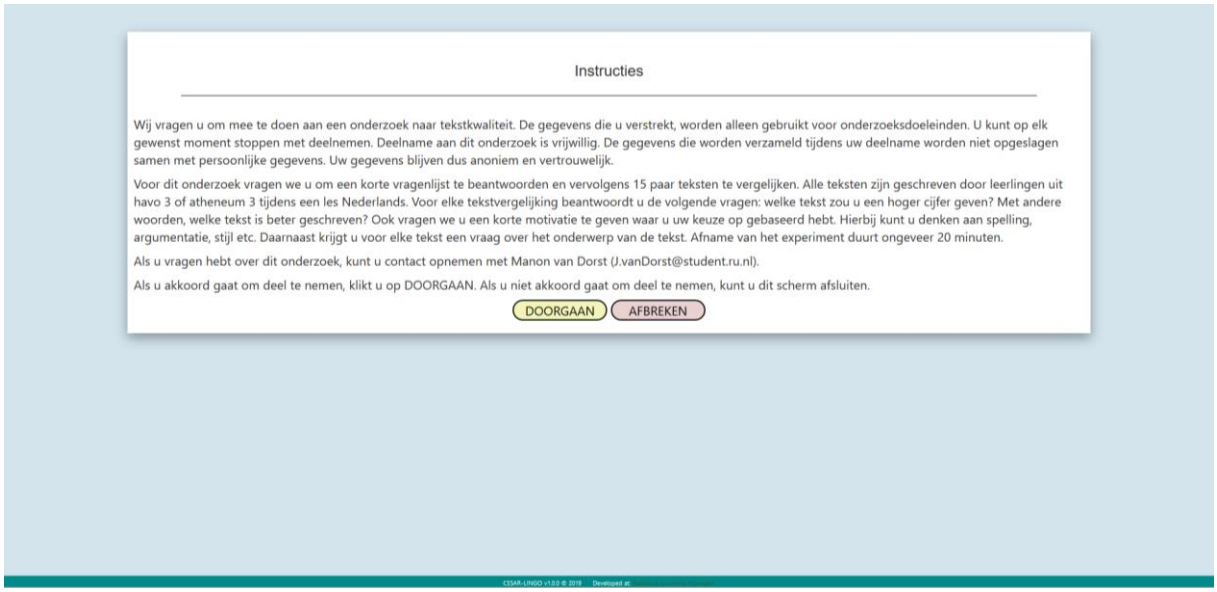
iedereen heeft wel een keer muziek of een film illegaal gedownload, omdat ze het geld er niet voor willen uitgeven of omdat andere zeggen dat ze dat moeten doen. De makers van de film of muziek vinden dat zelf niet zo fijn, omdat ze er veel tijd en geld hebben ingestopt en als je het dan illegaal download krijgen de makers geen geld voor hun harde werk. Andere mensen proberen het illegaal downloaden van films of muziek verbieden maar dat is niet zo makkelijk want mensen blijven het doen. Je kan tegenwoordig boetes krijgen als je illegaal dingen download.

Iedereen ken het programma Netflix wel, waarmee je films/series mee kan kijken. Er is daarvan ook al een illegale versie genaamd Popcorntime. Popcorntime is al wel ‘’kapot gemaakt’’, omdat je het niet meer kan gebruiken zelf heb ik wel wat dingen illegaal gedownload, maar ik ben er al wel mee gestopt.

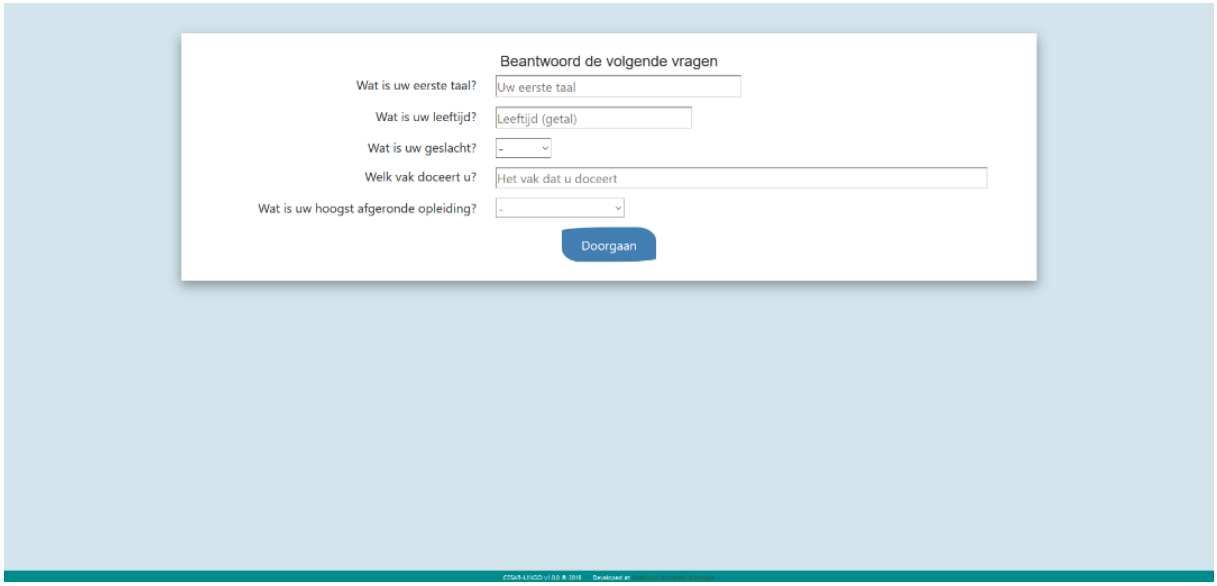
Bijlage B

Cesar-Lingo-experiment: paarsgewijze beoordelingsmethode

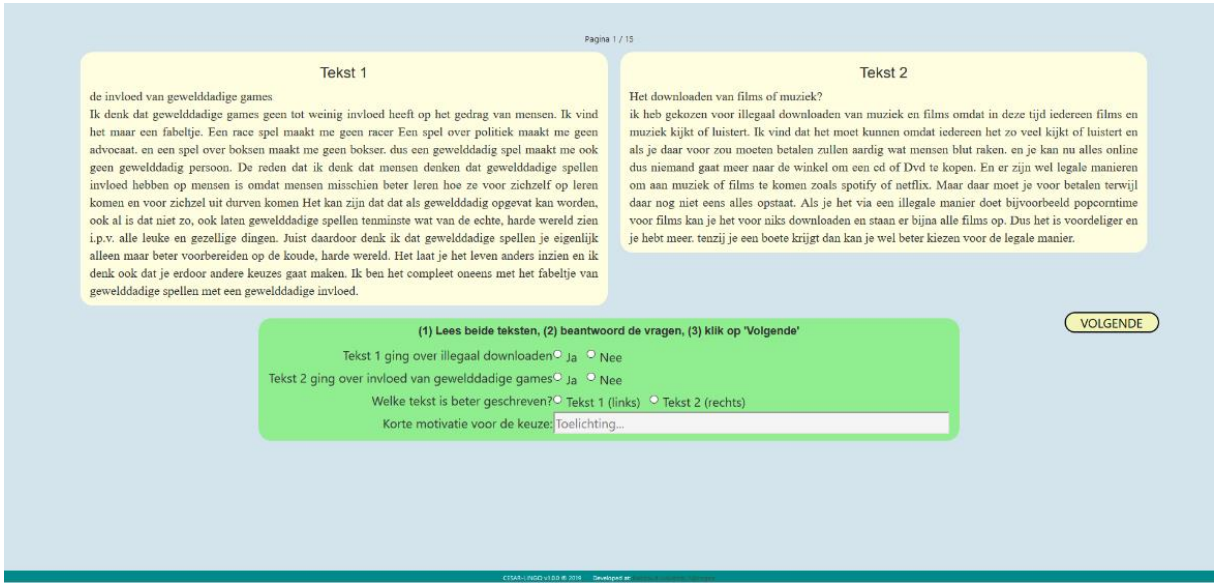
Link: https://cesar.science.ru.nl/lingo/experiment/do/6



Figuur 7. Screenshot van het paarsgewijze beoordelingsexperiment: instructie



Figuur 8. Screenshot van het paarsgewijze beoordelingsexperiment: vragen over de demografische gegevens van de participanten

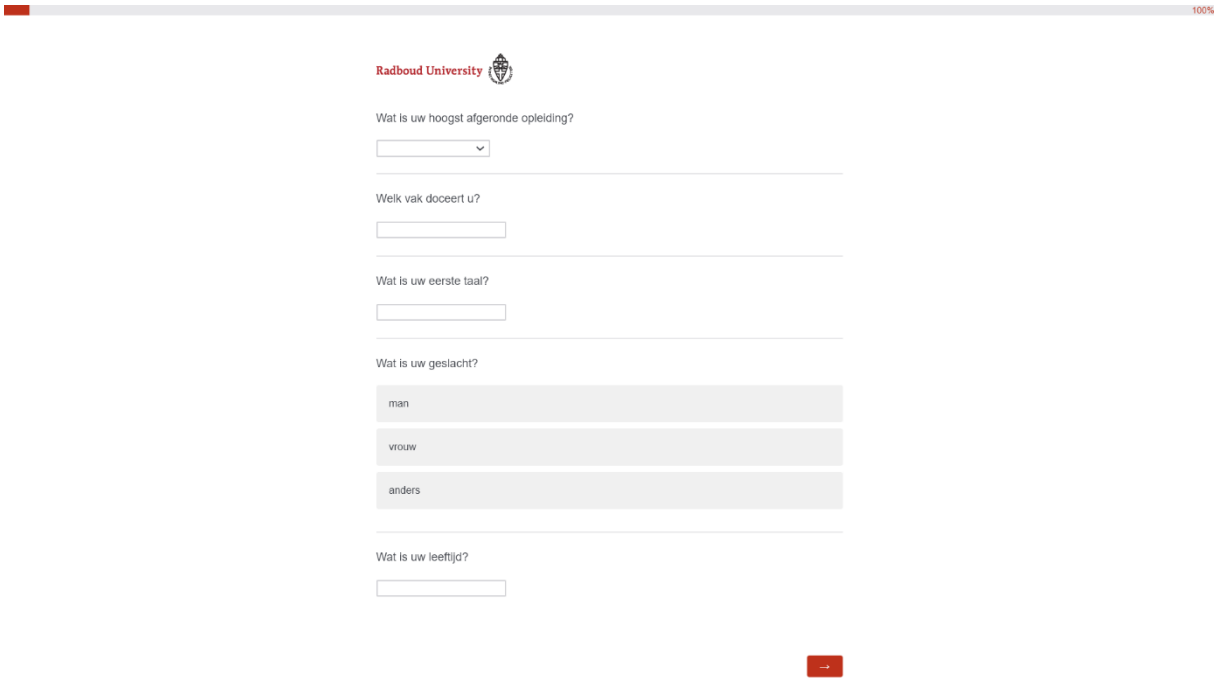


Figuur 9. Screenshot van het paarsgewijze beoordelingsexperiment: voorbeeld beoordelingsvraag

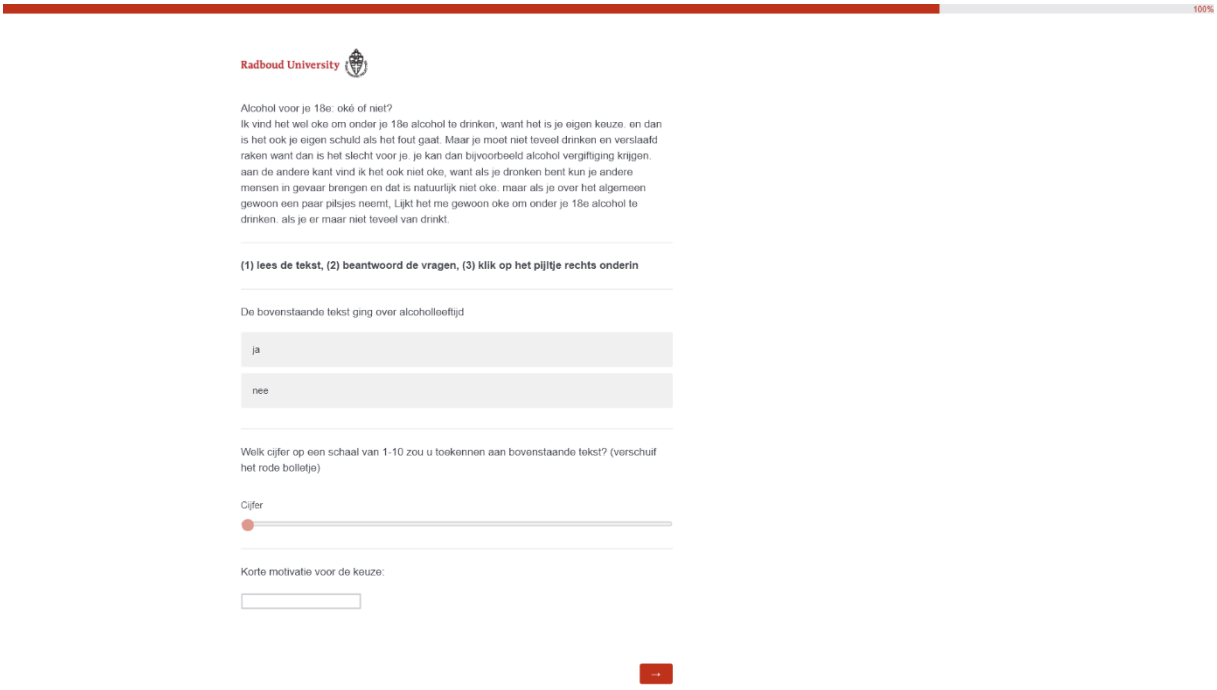
Qualtricsexperiment: individuele beoordelingsmethode

Link: https://radboudletteren.eu.qualtrics.com/jfe/form/SV_dnkgYTXH0li6Xpr

Figuur 10. Screenshot van het individuele beoordelingsexperiment: instructie



Figuur 11. Screenshot van het individuele beoordelingsexperiment: vragen over de demografische gegevens van de participanten



Figuur 12. Screenshot van het individuele beoordelingsexperiment: voorbeeld beoordelingsvraag