Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering

L. Coertjens[1], M. Lesterhuis[1], S. Verhavert, R. Van Gasse en S. De Maeyer

Samenvatting

Tekstkwaliteit betrouwbaar beoordelen zonder daar veel tijd aan te besteden is cruciaal voor zowel schrijfonderzoekers als de onderwijspraktijk. In deze studie namen we twee beoordelingsmethoden onder de loep: criterialijsten, die analytisch en absoluut van insteek zijn, en paarsgewijze vergelijking, een methode met een holistische en vergelijkende opzet. Voor beide methoden brachten we in kaart hoe lang een beoordeling per tekst duurde en hoe de betrouwbaarheid veranderde naarmate de groep van beoordelaars meer tijd investeerde in het beoordelen. Uit de resultaten bleek dat voor beide methoden de benodigde tijd afnam naarmate een beoordelaar al (meerdere) beoordelingen had gemaakt. De resultaten lieten ook zien dat wanneer betrouwbaarheid opgevat wordt als een maat voor de stabiliteit van de rangorde, beide methoden een vergelijkbare tijdsinvestering vragen. Vervolgonderzoek moet uitwijzen welke methode meer tijd vraagt wanneer rekening gehouden wordt met de tijd die nodig is om een criterialijst te ontwikkelen of om een evaluatie met behulp van paarsgewijze vergelijking op te zetten. Daarnaast moet toekomstig onderzoek uitwijzen of de conclusies uit dit onderzoek ook gelden voor andere teksten en andere criterialijsten.

Kernwoorden: tekstbeoordeling, paarsgewijze vergelijking, criterialijsten, betrouwbaarheid, tijdsinvestering

1. Inleiding

Het beoordelen van de kwaliteit van een tekst is geen sinecure. Cruciaal is dat de beoordelingen voldoende betrouwbaar zijn. We willen immers dat conclusies over de tekstkwaliteit niet afhankelijk zijn van de beoordelaars die de teksten beoordeeld hebben. Verschillende beoordelingsmethoden zijn voorhanden die het betrouwbaar beoordelen van tekstkwaliteit ondersteunen. De keuze voor een bepaalde beoordelingsmethode is echter niet vanzelfsprekend, vooral omdat de tijdsinvestering (de totale tijd die de groep beoordelaars nodig heeft om alle teksten te beoordelen) om tot betrouwbare beoordelingen te komen erg kan verschillen. Deze studie beoogt meer inzicht te geven in de tijd die nodig is om tot een betrouwbaar oordeel van de kwaliteit van teksten te komen, door de relatie tussen de betrouwbaarheid en de tijdsinvestering van verschillende beoordelingsmethoden na te gaan.

Hoewel er veel varianten bestaan, kunnen we beoordelingsmethoden ruwweg indelen aan de hand van twee dimensies. De eerste dimensie onderscheidt analytisch van holistisch beoordelen (Bacha, 2001; Weigle, 2002). Analytische methoden vertrekken vanuit het idee dat tekstkwaliteit uit verschillende dimensies bestaat. Door deze vooraf vast te leggen in een criterialijst, beoogt men de beoordelingen van diverse beoordelaars gelijk te trekken (Hamp-Lyons, 2002). Holistische methoden gaan uit van het idee dat schrijven een alomvattende competentie is, en dat onderliggende deelvaardigheden zo sterk samenhangen dat het niet gepast is deze apart te beoordelen (Sadler, 2009)

De tweede dimensie onderscheidt absolute beoordelingen van vergelijkende beoordelingen. Bij absolute beoordelingen beoordeelt men elke tekst op zichzelf met behulp van een competentieomschrijving of criterialijst. De drijfveer achter vergelijkende methoden is dat absolute beoordelingen moeilijk zijn voor beoordelaars (Bouwer & Koster, 2016; Pollitt, 2012b; Yeates, O'Neill, Mann, & Eva, 2013). Deze methoden bouwen voort op onze intuïtie om te vergelijken wanneer we kwaliteit beoordelen (Crisp, 2013; Greatorex, 2007).

Tot nu toe is er geen onderzoek gedaan naar de verschillen in betrouwbaarheid en tijdsinvestering van beoordelingen van teksten via analytisch en absoluut beoordelen, noch naar verschillen tussen het beoordelen op holistische en vergelijkende wijze. Hierdoor is nog onvoldoende duidelijk hoeveel beoordelingen per tekst nodig zijn om een betrouwbare beoordeling te krijgen en hoeveel tijd een beoordeling kost voor de verschillende beoordelingsmethoden. Beide aspecten zijn belangrijk voor zowel schrijfonderzoekers als de onderwijspraktijk.

2. Theoretisch kader

2.1 Analytisch en absoluut beoordelen: criterialijsten

In de onderwijspraktijk beoordeelt men tekstkwaliteit vaak met behulp van criterialijsten; een absolute en analytische beoordelingsmethode (Bloxham, den-Outer, Hudson, & Price, 2016; Jonsson & Svingby, 2007; Lane & Stone, 2006). Beoordelaars beoordelen de teksten daarbij op vooraf opgestelde criteria. Het uiteindelijke oordeel over de kwaliteit is gelijk aan de som van de scores voor de verschillende deelcriteria, al dan niet rekening houdend met een bepaalde weging van de criteria. De mate waarin de scores van verschillende beoordelaars overeenstemmen, bepaalt de interbeoordelaarsbetrouwbaarheid (Stemler, 2004). Deze interbeoordelaarsbetrouwbaarheid kan op verschillende manieren benaderd en berekend worden, bijvoorbeeld door te kijken naar ofwel de absolute consensus ofwel de consistentie. De absolute consensus tussen beoordelaars betreft in welke mate de beoordelaars exact dezelfde scores aan teksten geven. De consistentie geeft aan in welke mate de beoordelaars de teksten op dezelfde wijze ordenen of classificeren.

Verschillende studies wijzen erop dat beoordelaars sterk kunnen verschillen in hoe ze de kwaliteit van een tekst beoordelen (Diederich, French, & Carlton, 1961; McColly, 1970). Zo stelden Bloxham et al. (2016) grote verschillen vast in de beoordelingen van vijf teksten door zes beoordelaars: in bijna de helft van de gevallen lagen de beoordelingen voor de deelcriteria verspreid over de hele schaal (1 tot 5). Er kunnen verschillende oorzaken ten grondslag liggen aan de verschillen in scores tussen beoordelaars. Beoordelaars kunnen verschillen in generieke strengheid, wat met name impact heeft op de absolute consensus tussen beoordelaars. Ook kunnen zij verschillende ideeën hebben over wat tekstkwaliteit inhoudt, met als gevolg dat zij de criterialijst verschillend invullen. Dit kan zowel de absolute consensus als de consistentie van de beoordelingen beïnvloeden (Eckes, 2008; Lumley, 2002). De verschillen tussen beoordelaars kunnen niet altijd worden opgelost door hen vooraf te trainen (Rezaei & Lovorn, 2010; Weigle, 1999).

Omdat oordelen over de kwaliteit van teksten sterk onderhevig zijn aan verschillen tussen beoordelaars, dient men bij voorkeur meerdere beoordelaars te betrekken bij het beoordelen van teksten (Bouwer & Koster, 2016; Schoonen, 2005). Zo waren in de studie van Bouwer en Koster (2016) minstens twee beoordelaars per tekst nodig voor een consistentie van .70 en bereikte maar één van de twee onderzochte taken een betrouwbaarheid van .80 wanneer drie beoordelaars de teksten van een score voorzagen. Maar de lengte van de tekst, de lengte en kwaliteit van de criterialijst en het type van beoordelaars hebben ook een invloed op hoe gemakkelijk betrouwbare scores worden verkregen (Breland, 1983). Het is dus moeilijk om een eenduidig antwoord te geven op hoeveel beoordelingen nodig zijn om tot een betrouwbare score te komen. Daarom is het interessant om het aantal analytisch absolute beoordelingen dat nodig is voor betrouwbare scores te vergelijken met het aantal beoordelingen dat hiervoor nodig is wanneer een holistisch comparatieve methode wordt gebruikt.

2.2 Holistisch en comparatief beoordelen: paarsgewijze vergelijking

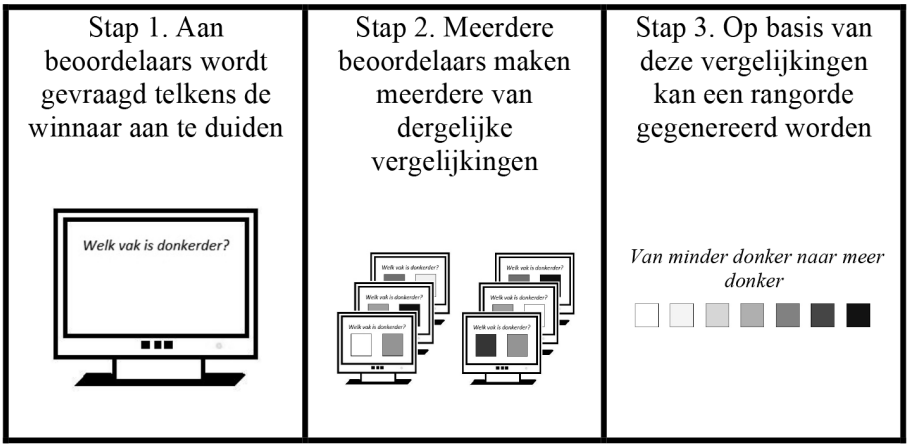
Paarsgewijze vergelijking is een alternatieve beoordelingsmethode, die vooral sterk in opmars is voor het beoordelen van schrijfcompetenties. Deze methode combineert een holistische en een vergelijkende aanpak (Pollitt, 2012a). Zoals Figuur 1 weergeeft, vergelijken beoordelaars telkens twee teksten en geven aan welke de beste tekst is (stap 1).

Nadat de beoordelaar een keuze heeft gemaakt, krijgt hij/zij een nieuw paar om te vergelijken. Verschillende beoordelaars nemen deel aan het beoordelingsproces en elke beoordelaar maakt verschillende vergelijkingen na elkaar (stap 2). Op basis van al deze vergelijkingen kunnen we vervolgens met behulp van het Bradley-Terry-Luce model (Verhavert, De Maeyer, Donche, & Coertjens, ter perse) een schaal opstellen van de teksten, gerangschikt van lagere tot hogere tekstkwaliteit (stap 3). De schaal is gebaseerd op de consensus van de beoordelaars over wat een goede tekst is (Jones & Alcock, 2014; Pollitt, 2012a; van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2016). Voor meer informatie over het opzetten van een assessment op basis van paarsgewijze vergelijking verwijzen we naar Lesterhuis, Verhavert, Coertjens, Donche en De Maeyer (2016).

In de context van paarsgewijze vergelijking is betrouwbaarheid een maat om uit te drukken hoe stabiel de relatieve positie van teksten (en dus ook de resulterende scores) is op de gegenereerde schaal (Andrich, 1982). Of anders uitgedrukt, de mate waarin de teksten na een extra beoordeling door vergelijkbare beoordelaars dezelfde plaats op de schaal zouden innemen. Over de precieze interpretatie van de Scale Separation Reliability (SSR) wordt gediscussieerd (Bramley, 2015; Pollitt, 2012b). Recent onderzoek toonde aan dat de SRR een goede maat is voor interbeoordelaarsbetrouwbaarheid ofwel de stabiliteit tussen beoordelaars (Verhavert et al., ter perse).

Tot nog toe rapporteerden studies naar paarsgewijze vergelijking voornamelijk hoge betrouwbaarheden (>.80, voor een overzicht, zie Bramley, 2015). Hetzelfde geldt voor studies naar het beoordelen van schrijfcompetenties, met betrouwbaarheden tussen .84 (van Daal et al., 2016) en .98 (Heldsinger & Humphrey, 2010). Om deze betrouwbaarheden te bereiken hoeven niet alle mogelijke vergelijkingen gemaakt te worden. Meerdere vergelijkingen per tekst (een steekproef uit alle mogelijke vergelijkingen) volstaat. Dit aantal vergelijkingen per tekst heeft een impact op de betrouwbaarheid van de rangorde. In de studies van van Daal et al. (2016) en Heldsinger en Humprey (2010) werden respectievelijk 8-16 vergelijkingen en 69 vergelijkingen per tekst gemaakt. Een meta-analyse van 49 verschillende soorten assessments toonde aan dat voor een betrouwbaarheid van .70 gemiddeld 12 vergelijkingen per product moesten worden gemaakt. Wanneer een betrouwbaarheid van .80 werd nagestreefd, waren gemiddeld 17 vergelijkingen per product nodig (Verhavert, Bouwer, Donche, & De Maeyer, 2017).

Het feit dat paarsgewijs vergelijken een eenvoudige opdracht is voor beoordelaars



Figuur 1

Paarsgewijze vergelijking: van vergelijken naar een schaal van teksten (overgenomen uit Lesterhuis et al. (2015))

vormt een verklaring voor de hoge betrouwbaarheid van de resultaten. Beoordelaars moeten namelijk enkel aangeven welke tekst zij beter vinden (Pollitt, 2012a). Bovendien schakelt deze methode de impact van verschillen in strengheid tussen beoordelaars op de betrouwbaarheid uit. Daarnaast kunnen zij verschillen in hun absoluut oordeel over twee teksten, maar zijn ze het vaak wel eens over welke tekst de betere is (Pollitt, 2012b).

Onderzoekers stelden recent vast dat de betrouwbaarheid mogelijk ook afhangt van de wijze waarop de vergelijkingen samengesteld zijn (Bramley, 2015; Lesterhuis et al., 2016). Vergelijkingen kunnen op twee manieren samengesteld worden: willekeurig of adaptief. Bij een willekeurige samenstelling worden de vergelijkingen willekeurig getrokken uit de groep teksten met het kleinste aantal uitgevoerde vergelijkingen. De adaptieve manier houdt rekening met informatie uit voorafgaande vergelijkingen. Bijvoorbeeld, twee teksten die tot dan toe telkens als betere aangeduid werden, gaan samen een nieuwe vergelijking vormen (voor meer uitleg, zie Pollitt, 2012b). Vergelijkingen adaptief samenstellen is efficiënter. Pollitt (2012b) concludeerde dat er 40% tot 50% minder vergelijkingen nodig zijn om tot eenzelfde betrouwbaarheid te komen als met willekeurig samengestelde vergelijkingen.

Bramley (2015) toonde echter aan dat het Zwitsers systeem voor adaptieve samenstelling van vergelijkingen de betrouwbaarheid artificieel verhoogt. De beslissing of een tekst al dan niet wint in de eerste vergelijking heeft immers een grote impact op de samenstelling van de volgende paren. Dit kan ervoor zorgen dat twee teksten met gelijkaardige tekstkwaliteit op erg verschillende plaatsen in de rangorde uitkomen (Bramley, 2015). Er wordt, met andere woorden artificieel spreiding in kwaliteit gecreëerd. Aangezien de betrouwbaarheidsmaat SSR wordt bepaald door deze spreiding, kan deze dus vertekend zijn. Bovendien laten studies van Jones, Swan en Pollitt (2014) en McMahon en Jones (2015) zien dat ook een willekeurige samenstelling van paren tot adequate betrouwbaarheidsniveaus kan leiden (respectievelijk .80 en .87). Om hier meer duidelijkheid over te krijgen willen wij in deze studie nagaan hoe de betrouwbaarheid zich ontwikkelt wanneer we vergelijkingen willekeurig samenstellen.

2.3 Tijdsinvestering in verhouding tot betrouwbaarheid

Zowel de methode van criterialijsten als de methode van paarsgewijze vergelijking vereist meerdere beoordelingen per tekst om tot een betrouwbare score te komen. Het aantal benodigde beoordelingen per tekst lijkt echter sterk te verschillen tussen de beoordelingsmethoden. Om de beide methoden goed te kunnen vergelijken, moet er dus ook meer zicht komen op de tijd die een beoordeling per tekst vergt.

Er zijn twee studies die de tijdsduur van één beoordeling met een criterialijst vergelijken met de tijdsinvestering die nodig is om tot een betrouwbare rangorde te komen met paarsgewijze vergelijking (McMahon & Jones, 2015; Pollitt, 2012b). Pollitt (2012b) beschreef een beoordeling van 1000 tekstbundels. Deze bundels bestonden uit twee teksten van een leerling. 54 beoordelaars vergeleken de bundels paarsgewijs. Elf vergelijkingen per bundel leidde tot een betrouwbaarheid van .90. De benodigde tijdsinvestering per bundel (dit is de tijd die de groep beoordelaars samen nodig heeft om alle bundels te beoordelen, gedeeld door het aantal bundels) om tot deze betrouwbaarheid te komen, bedroeg omgerekend 25 minuten en 30 seconden (A. Pollitt, persoonlijke communicatie, 15 augustus 2016).

Pollitt (2012b) vergeleek deze tijdsinvestering met de benodigde tijd bij het gebruik van criterialijsten. De criterialijst bestond voor het eerste thema uit vier criteria en voor het tweede thema uit twee criteria (A. Pollitt, persoonlijke communicatie, 11 juli 2017). Pollitt (2012b) vermeldde dat de tijd om een bundel te beoordelen met de criterialijsten 15 tot 20 minuten bedroeg. Het is echter onduidelijk of beoordelaars al dan niet sneller werden naarmate ze meerdere beoordelingen hadden gemaakt. Daarnaast is er altijd een tweede beoordeling nodig om zicht te krijgen op de betrouwbaarheid van de beoordeling. Hiermee zou de totale tijdsinvestering bij

paarsgewijze vergelijking (25 min. 30 s.) dus lager liggen dan wanneer een bundel twee keer met behulp van criterialijsten beoordeeld zou zijn geweest (30 tot 40 min).

In zijn studie concludeerde Pollitt (2012b) ook dat er een grote variatie bestond in de tijd die een beoordelaar gemiddeld nodig had om een vergelijking af te ronden. De gemiddelde tijd voor een vergelijking bedroeg 4 min. 40 s.[2]. De snelste beoordelaar had echter gemiddeld maar 1 min. 30 s. nodig per vergelijking terwijl de traagste beoordelaar gemiddeld 9 min. nodig had. De snelheid van een beoordelaar bleek niet samen te hangen met de kwaliteit van de beoordelingen (Pollitt, 2012b). In reactie daarop berekende Pollitt (2012b) de benodigde tijdinvestering opnieuw voor de 27 snelste beoordelaars (i.e., groep van 50% snelste beoordelaars). Deze resultaten gaven aan dat de groep van 50% snelste beoordelaars gemiddeld 3 min. 15 s. nodig had om één vergelijking af te ronden. Hieruit kunnen we afleiden dat deze groep beoordelaars ongeveer 18 min. per bundel nodig zou hebben gehad om een betrouwbaarheid van .90 te bekomen (wat neerkomt op 11 beoordelingen per tekst)[3]. Deze tijdsinvestering is vergelijkbaar met de 15 tot 20 min. tijd die nodig was voor de beoordeling van een bundel door één beoordelaar op basis van de criterialijsten.

Pollitt (2012b) had in zijn studie niet tot doel om de tijdsinvestering en betrouwbaarheid voor het beoordelen met behulp van criterialijsten en het beoordelen via paarsgewijze vergelijking expliciet tegen elkaar af te zetten. Bijgevolg ontbreekt een precieze meting van de tijd die nodig was voor een beoordeling met criterialijsten, wat de vergelijking tussen de benodigde tijdsinvestering voor beide methoden bemoeilijkt. Daarnaast stelde Pollitt (2012b) de vergelijkingen in zijn studie op een adaptieve manier samen. De vastgestelde betrouwbaarheid (.90) is dus mogelijk een overschatting van de werkelijke betrouwbaarheid (Bramley, 2015).

McMahon en Jones (2015) gebruikten in hun studie wel willekeurig samengestelde vergelijkingen voor het beoordelen van het begrip van studenten over een chemie-experiment aan de hand van vier korte open vragen (N=154). In deze studie beoordeelde één beoordelaar elk product eerst analytisch met behulp van een beknopte criterialijst. Het kostte een beoordelaar minder dan anderhalve minuut om een product op basis van deze criterialijst te beoordelen. De totale tijdsinvestering voor de beoordeling van alle 154 producten bedroeg 3 uur. Na deze analytische beoordeling, beoordeelden vijf beoordelaars elk product gemiddeld 20 keer via paarsgewijze vergelijking. Dit leverde een betrouwbaarheid op van .87. De totale tijdsinvestering voor de beoordeling van alle producten kwam neer op ongeveer 14 uur. McMahon en Jones (2015) concludeerden dat, zelfs indien een tweede beoordeling zou worden toegevoegd bij de criterialijsten (zijnde 6 uur tijdsinvestering), paarsgewijze vergelijking meer dan dubbel zoveel tijd vroeg. Hierbij hielden zij echter geen rekening met de verschillen in snelheid tussen beoordelaars. Ook kon de betrouwbaarheid niet worden nagegaan van de beoordeling met de criterialijst.

De studies leiden niet tot eenduidige inzichten omtrent de tijdsinvestering bij een beoordeling via criterialijsten in vergelijking met die bij de methode van paarsgewijze vergelijking. Pollitt (2012b) concludeerde dat paarsgewijze vergelijking een vergelijkbare tijdsinvestering vraagt als beoordelen met behulp van een criterialijst. McMahon en Jones (2015) lieten daarentegen zien dat de tijdsinvestering bij paarsgewijze vergelijking meer dan dubbel zo groot was. In deze studie werd echter elk product slechts één keer beoordeeld met behulp van de criterialijst waardoor niet duidelijk is hoeveel beoordelingen per product nodig zijn om tot een vergelijkende betrouwbaarheid te komen. Daarnaast bieden deze studies geen inzicht in de tijd die nodig is om teksten meerdere keren te beoordelen, verschillen in tijd tussen beoordelaars, en dat beoordelaars (waarschijnlijk) sneller worden naarmate zij meer beoordelingen hebben gemaakt.

2.4 Deze studie

Om een beter inzicht te krijgen in de tijd die nodig is voor een beoordeling, is het belangrijk om voor beide beoordelingsmethoden in kaart te brengen hoe sterk de tijdsduur per beoordeling varieert. Hiervoor verwachten

we verschillen tussen beoordelaars, maar ook een afname van de tijdsduur naarmate beoordelaars meer gewend zijn aan de beoordelingstaak. De eerste onderzoeksvraag luidt:

1. Hoeveel tijd is er nodig om een tekst te beoordelen op basis van criterialijsten en op basis van paarsgewijze vergelijking?

Op basis van deze informatie kunnen we vervolgens de benodigde tijdsinvestering en de betrouwbaarheid van beide methoden nagaan.

2a. Wat is het effect van een grotere tijdsinvestering op de betrouwbaarheid wanneer er met criterialijsten wordt beoordeeld?
2b. Wat is het effect van een grotere tijdsinvestering op de betrouwbaarheid wanneer er met paarsgewijze vergelijking wordt beoordeeld?

Daarnaast willen we voor beide beoordelingsmethoden nagaan welke tijdsinvestering nodig is om tot een stabiele rangorde te komen. Geformuleerd als derde onderzoeksvraag:

3. Wat is het effect van een grotere tijdsinvestering op de mate van stabiliteit van de rangordes voor beide methoden?

3. Methoden

3.1 Materiaal

Binnen deze studie stond de competentie 'argumentatief schrijven' centraal. We definieerden deze competentie volgens de eindtermen van de derde graad van het secundair onderwijs (www.ond.vlaanderen.be): "De leerling is in staat voor een onbekend publiek op beoordelend niveau een gedocumenteerde en beargumenteerde tekst te schrijven. Meer specifiek kunnen zij 1) hun voorkennis inzetten, 2) gericht informatie ordenen en verwerken, 3) een logische tekstopbouw creëren met aandacht voor inhoudelijke en functionele relaties, 4) inhouds- en vormconventies van de taal verzorgen, 5) lay-out verzorgen en 6) correct citeren (bronvermelding)".

Om kwaliteitsvolle taken te garanderen, selecteerden we taken uit eerder wetenschappelijk onderzoek. De selectie gebeurde op grond van volgende criteria: de taak ligt in lijn met de competentiebeschrijving; de taak is relevant voor leerlingen in het vijfde leerjaar algemeen secundair onderwijs; de taak neemt niet langer dan 25 minuten in beslag; de taak resulteert in een tekst van maximaal één A4; en er is een gevalideerde criterialijst beschikbaar. Mits aanpassing aan de Vlaamse context, voldeden de taken die gebruikt zijn in de proefschriften van Van Weijen (2009) en Tillema (2012) aan deze criteria. We selecteerden drie taken, met als thema's "Orgaandonatie", "Kinderen krijgen" en "Stress bij scholieren", waarvan de laatste gebruikt is in deze studie (voor dit laatste thema, zie bijlage 1).

De teksten werden beoordeeld met behulp van een criterialijst met 20 dimensies (zie bijlage 2 voor de criterialijst). Dit is een aangepaste versie van de criterialijst zoals gebruikt door Breetvelt, van den Bergh & Rijlaarsdam (1994) en Van Weijen (2009). Zij toonden aan dat de criterialijst een hoge betrouwbaarheid opleverde bij teksten van leerlingen van 14 tot 15 jaar en leerlingen van 18 tot 19 jaar (respectievelijk $\alpha = .76$ en $\alpha = .88$). De scores op basis van de criterialijst correleerden sterk (.87) met een holistische beoordeling die hetzelfde beoogde te meten (Van Weijen, 2009). Gezien de specifieke competentiebeschrijving, voegden we drie criteria toe die betrekking hebben op taal, namelijk: grammatica/spelling, interpunctie en stijl.

3.2 Testpersonen

Deze studie maakte deel uit van een ruimer onderzoek waaraan tien scholen deelnamen, met leerlingen uit het vijfde jaar (16-17 jaar). In totaal schreven 135 leerlingen drie teksten binnen een tijdsbestek van twee lesuren. Alvorens zij begonnen, kregen zij informatie over het doel van het onderzoek, de competentie die beoordeeld zou worden en de drie specifieke taken. Na deze uitleg ondertekenden ze een consent formulier. Voor deze studie focusten we op de beoordelingen van de 35 teksten met als onderwerp "Stress bij scholieren". Deze teksten werden random geselecteerd uit de 135 teksten.

3.3 Beoordelaars

De werving van de beoordelaars (taalleerkrachten, lerarenopleiders en leraren in opleiding voor het vak Nederlands) gebeurde via (persoonlijke) netwerken, lerarenopleidingen en vacaturesites. In totaal beoordeelden 58 beoordelaars de taak "Stress bij scholieren". Op basis van toeval wezen we de beoordelaars toe aan een van beide beoordelingscondities (zie Tabel 1).

3.4 Procedure beoordelen

Het beoordelen nam twee middagen in beslag. Ongeacht de beoordelingsconditie, startte de eerste bijeenkomst met een uitleg over de te beoordelen competentie (argumentatief schrijven) en de schrijftaak en het ondertekenen van een consent formulier. Vervolgens beoordeelden de deelnemers twee uur lang met een pauze van een kwartier. De tweede namiddag startten zij direct met beoordelen en na twee uur kregen zij een uitgebreide uitleg over het onderzoeksproject.

In de conditie 'criterialijst' werd beoordeeld met behulp van de software Qualtrics (Provo, UT). Dit platform registreerde de tijd tussen het moment dat een beoordelaar een tekst ontving en het moment dat de beoordelaar de ingevulde criterialijst terugstuurde. Gemiddeld maakte elke beoordelaar 10 vergelijkingen (SD = 3). Voor elk van de beoordelaars kozen we de teksten op basis van toeval. Het gevolg van deze werkwijze is een verschil in de volgorde waarin een beoordelaar de teksten te zien kreeg. Dit maakte ook dat sommige teksten vaker werden beoordeeld dan andere. Na de beoordelingen genereerden we per beoordeling een totaalscore door alle 20 criteria van de criterialijst op te tellen. Elk van de 35 teksten was minstens vijf keer beoordeeld, op twee teksten na. Om de 173 beoordelingen optimaal te kunnen gebruiken, kenden we aan de twee teksten die slechts vier keer waren beoordeeld een vijfde score toe op basis van het gemiddelde van de andere vier scores[4]. Het databestand voor de conditie 'criterialijst' bestond dus uit 175 datapunten.

In de conditie 'paarsgewijze vergelijking' gebruikten de beoordelaars het Digitaal Platform voor Assessment van Competenties (D-PAC, www.d-pac.be). In D-PAC worden teksten automatisch in random paren gedeeld, op basis van het aantal keer dat een tekst al is vergeleken. Hiermee wordt gegarandeerd dat elke tekst ongeveer even vaak in een paar terugkomt. Dit was in totaal 27 of 28 keer per tekst, wat maakte dat de beoordelaars in totaal 474 paren vergeleken (M= 11, SD = 4 per beoordelaar). Dit platform logde eveneens de tijd tussen het moment dat een beoordelaar een vergelijking ontving en het moment dat de beoordelaar aangaf welke van de twee teksten beter was. Tijdens het beoordelingsproces werd de te beoordelen competentie op het whiteboard geprojecteerd. De beoordelaars bekeken eerst teksten over de andere thema's (zie 3.1. Materiaal), wat maakte dat zij al ervaring hadden opgedaan met de vergelijkende manier van beoordelen. Dit gold echter niet voor de groep leraren in opleiding, die vanwege organisatorische redenen direct met de teksten over "Stress bij scholieren" startten.

3.5 Analyse

Tijd nodig om een tekst te beoordelen

Voor de eerste onderzoeksvraag gingen we voor de beide beoordelingsmethoden na hoeveel tijd er nodig was om één tekst één keer te beoordelen. Hiervoor moesten we rekening houden met het feit dat de duur van een tekst-

Tabel 1
Verdeling beoordelaars over condities

	Criterialijsten (N = 18)	Paarsgewijze vergelijking (N = 40)
Taalleerkrachten	6 (33.3%)	13 (32.5%)
Lerarenopleiders	4 (22.2%)	8 (20%)
Leraren in opleiding	8 (44.4%)	19 (47.5%)

beoordeling afhankelijk was van de beoordelaar en de tekst die beoordeeld werd. Ook gingen we na of de duur van een beoordeling per tekst afnam naarmate de beoordelaar al meerdere beoordelingen had gemaakt. Om deze onderzoeksvraag te beantwoorden schatten we twee mixed effect modellen[5]. Het voordeel van deze modellen is dat ze rekening houden met zowel fixed als random effecten (Baayen & Milin, 2015), wat zorgt voor zuivere schattingen. De analyses werden uitgevoerd in SPSS (IBM corporation, versie 24.0).

Zowel in het model voor de conditie 'criterialijst' als het model voor de conditie 'paarsgewijze vergelijking' werd een intercept geschat voor de duur van een beoordeling van de eerste tekst. Daarnaast voegden we als fixed effect toe de hoeveelste beoordeling het was voor de beoordelaar die deze beoordeling maakte voor het thema "Stress bij scholieren". Op deze manier gingen we na of de schattingen voor de duur van een beoordeling per tekst al dan niet significant toe- of afnam naarmate een beoordelaar reeds meerdere beoordelingen afrondde.

In het model voor de conditie 'criterialijst' werden zowel beoordelaars en teksten als random effecten opgenomen. Dit laat toe om de variantiecomponent na te gaan dat toe te schrijven is aan de verschillen tussen beoordelaars in de specifieke beoordelingsmethoden. Voor de conditie 'paarsgewijze vergelijking' kon enkel de variantiecomponent voor de beoordelaars worden berekend, omdat in deze methode geen enkele combinatie van teksten in een paar vaker dan één keer voorkwam.

Het effect van tijdsinvestering op de betrouwbaarheid

Om na te gaan wat het effect was van tijdsinvestering op de betrouwbaarheid binnen de conditie 'criterialijst' berekenden we de one-way random intra-klasse correlatiecoefficiënt (ICC, Gwet, 2014; Shrout & Fleiss, 1979), waarmee we de consistentie tussen beoordelaars in kaart brachten. De ICC werd berekend als de ratio van de variantie tussen de teksten tot de totale variantie (Gwet, 2014). Een hoge ICC betekent dus dat de proportie variantie tussen teksten in de totale variantie groot is. De totale variantie is de som van de variantie tussen de teksten en de error variantie. Deze error variantie bestaat uit twee componenten: de beoordelaarsfactor en de error factor. Doordat elke beoordelaar binnen de conditie 'criterialijst' echter een toevallige set van teksten beoordeelde, konden beide componenten niet van elkaar onderscheiden worden[6]. We berekenden de ICC in SPSS op grond van vijf beoordelingen per tekst. Via de Spearman-Brown formule konden we terugrekenen wat de betrouwbaarheid was voor twee tot en met vier beoordelingen per tekst (Bouwer & Koster, 2016).

Om na te gaan wat het effect was van tijdsinvestering op de betrouwbaarheid binnen de conditie 'paarsgewijze vergelijking', berekenden we de SSR (Bramley, 2015) in R (pakket BradleyTerry2), wat ook een maat is voor de consistentie tussen beoordelaars.

Om de tijd in te schatten, gebruikten we de schattingen op basis van het twee multilevel modellen (zie de subparagraaf hiervoor: tijd nodig om een tekst te beoordelen). Gemiddeld rondde een beoordelaar in de conditie 'criterialijst' 10 beoordelingen af. In de conditie 'paarsgewijze vergelijking' rondde een beoordelaar gemiddeld 11 vergelijkingen af. Omdat de duur van de beoordeling werd beïnvloed door het aantal beoordelingen, werd voor beide condities de geschatte tijd van de vijfde beoordeling genomen. In de conditie 'criterialijst' bedroeg dit 5 min. 47 s., in de conditie 'paarsgewijze vergelijking' hadden beoordelaars per tekst 1 min. 4 s. nodig bij het maken van de vijfde beoordeling.

Vergelijking van stabiliteit rangorde tussen beide methoden bij gelijke tijdsinvestering

Gezien beide methoden vergelijkbare kwaliteiten beoordeelden, was het relevant om na te gaan of een van beide methoden met minder tijdsinvestering een stabiele rangorde bereikte. De Kendalls tau rangorde correlatiecoëfficiënt (ofwel Kendalls τ) tussen de scores na vijf beoordelingen in de conditie 'criterialijst' en de scores na 27 rondes in de conditie 'paarsgewijze vergelijking' was .66 ($p<.001$), de Pearson-product correlatie coëfficiënt (ofwel Pearsons r) .85 ($p<.001$)[7].

We opteerden om de stabiliteit van de rangorde na te gaan, omdat dit variabelen van slechts ordinaal niveau veronderstelt. Aan deze voorwaarde is voldaan: in beide condities waren de scores van ordinaal niveau. De scores ordenden immers de teksten van de minst goede naar de beste. Wanneer een extra tijdsinvestering leidde tot een verandering in de rangorde van teksten, impliceert dit dat de extra beoordeling informatie toevoegt. Een stabiele rangorde daarentegen betekent dat een extra beoordeling weinig informatie toevoegt.

De stabiliteit van de rangorde gingen we voor de conditie 'criterialijst' als volgt na. Na de eerste beoordeling werden de scores genoteerd voor elke tekst. Voor elke nieuwe beoordeling werd de gemiddelde score van de tekst tot dan toe berekend. Bijvoorbeeld, indien een tekst een score van 48 kreeg bij een eerste beoordeling en een score 43 bij een tweede beoordeling, gebruikten we voor de eerste beoordeling 48 en voor de tweede beoordeling het gemiddelde van beide beoordelingen, wat 45.5 is. Wanneer de derde beoordelaar vervolgens een score 33 toekende, is het derde datapunt het gemiddelde van 48, 43 en 33 (dat is 41.33). Om de stabiliteit van de rangordes na te gaan, berekenden we Kendalls τ op de scores van twee opeenvolgende beoordelingen (bijv. na één en na twee beoordelingen per tekst).

Om de resultaten met betrekking tot de benodigde tijd voor paarsgewijze vergelijking vergelijkbaar te maken met die uit de criterialijstenconditie, gingen we uit van de tijdsinvestering in de conditie 'criterialijst'. Net als in onderzoeksvraag 2, baseerden wij de tijd op de uitkomsten van het multilevel-model van onderzoeksvraag 1. De tijd die nodig om een tekst één keer te laten beoordelen op basis van de in deze studie gebruikte criterialijst (5 min. 47 s., zie tabel 2), kwam bijvoorbeeld ongeveer overeen met de tijd die nodig was om vijf vergelijkingen per tekst af te ronden in de conditie 'paarsgewijze vergelijking' (5 min. 18 s.). Voor twee beoordelingen op basis van een criterialijst (11 min. 34 s.), kwam dit overeen met 11 vergelijkingen per tekst in de conditie 'paarsgewijze vergelijking' (11 min. 39 s.). Ook voor de derde, vierde en vijfde beoordeling werd een vergelijkbare tijdsinvestering in de conditie 'paarsgewijze vergelijking' gezocht. Dit was respectievelijk 16, 22 en 27 beoordelingen per tekst (zie tabel 6). Vervolgens berekenden we de logit score voor de teksten voor elk van deze 5 momenten (5, 11, 16, 22 en 27 beoordelingen per tekst in de conditie 'paarsgewijze vergelijking'). Daarna bekeken we, met behulp van de Kendalls τ, de stabiliteit van de rangorde tussen twee opeenvolgende rangordes (bijvoorbeeld de rangorde na 11 vergelijkingen per tekst en na 16 vergelijkingen per tekst).

4. Resultaten

4.1 Tijd nodig om een tekst te beoordelen

Om in te schatten hoeveel tijd nodig was om een beoordeling per tekst af te ronden, werden twee multilevel modellen geschat. De resultaten in tabel 2 geven aan dat de eerste beoordeling van een tekst voor een gemiddelde beoordelaar 5.5 keer sneller ging via paarsgewijze vergelijking dan met behulp van een criterialijst. In de conditie 'criterialijst' deed een gemiddelde beoordelaar 7 min. 08 s. over de eerste beoordeling, terwijl in de conditie 'paarsgewijze vergelijking' een gemiddelde beoordelaar 1 min. 17 s. per tekst nodig had voor de eerste beoordeling.

Daarnaast laten de resultaten zien dat het voor beide beoordelingsmethoden loonde om beoordelaars meerdere beoordelingen te laten maken (zie figuur 2). Het loonde sterker in de conditie 'criterialijst', waar voor elke extra beoordeling de benodigde tijd 6.6 keer sterker daalde dan via 'paarsgewijze vergelijking'. Bij elke extra beoordeling die een gemiddelde beoordelaar maakte in de conditie 'criterialijst', nam deze benodigde tijd immers af met 20 s. In de conditie 'paarsgewijze vergelijking' nam de tijd met 3 s. af voor elke extra beoordeelde tekst.

De resultaten geven ook weer dat de verschillen tussen beoordelaars in benodigde tijd per tekst groter waren in de conditie 'criterialijst' dan bij paarsgewijze vergelijking. De variantie in tijd die werd verklaard door beoordelaars was in de conditie 'criterialijst' 21.9%, en in de conditie 'paarsgewijze vergelijking' 14.6%.

Tabel 2
Duur van een beoordeling per tekst voor de conditie 'criterialijst'

	Schatting (in sec)	Standaard error	df	t-waarde	Sig (p)
Fixed effecten					
Intercept	428.42	29.52	33.68	14.52	<.001
Hoeveelste beoordeling	-20.31	3.91	161.27	-5.20	<.001
Random effecten					
Residuen	26055.4	161.42			
Tekst	200.4	14.16			
Beoordelaar	7380.4	85.91			

Tabel 3
Duur van een beoordeling per tekst voor de conditie 'paarsgewijze vergelijking'

	Schatting (in sec)	Standaard error	df	t-waarde	Sig (p)
Fixed effecten					
Intercept	77.07	4.64	122.60	16.62	<.001
Hoeveelste beoordeling	-3.38	0.56	462.40	-6.11	<.001
Random effecten					
Residuen	1979	18.38			
Beoordelaar	338	44.49			

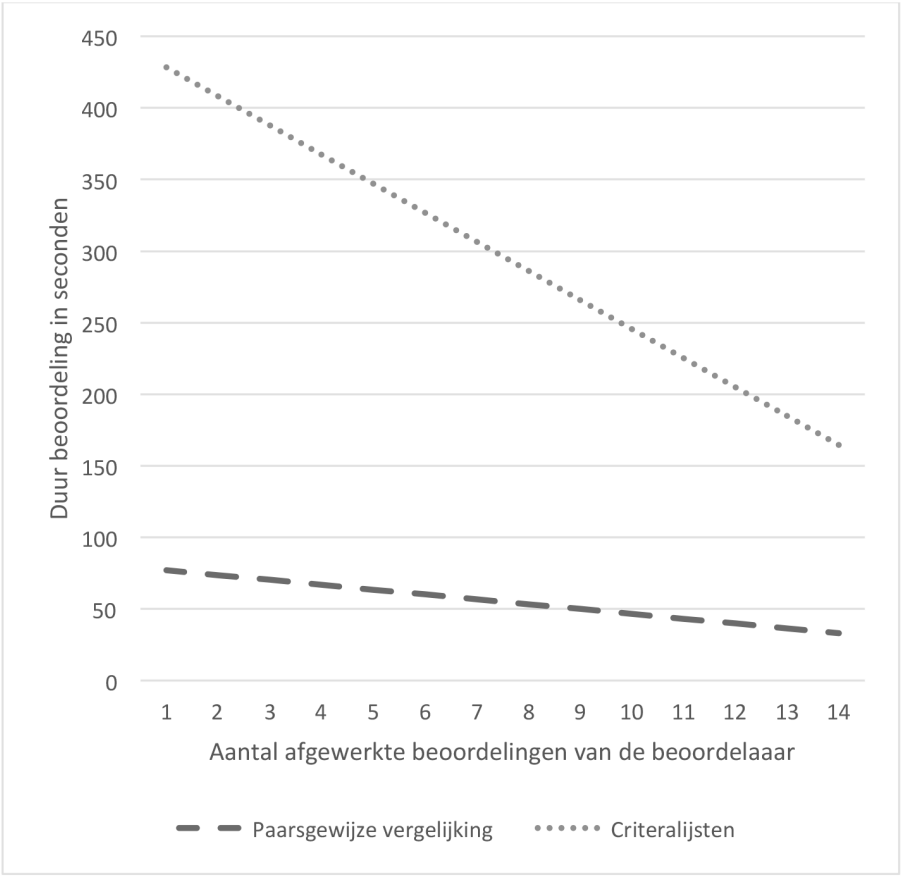
4.2 Het effect van tijdsinvestering op de betrouwbaarheid

Tabel 4 en figuur 3 laten zien dat voor de conditie 'criterialijst' de betrouwbaarheid toeneemt bij een grotere tijdsinvestering. Bij twee beoordelingen per tekst (een tijdsinvestering van 11 min. 34 s.) was de betrouwbaarheid .67, terwijl bij vijf beoordelingen per tekst (een tijdsinvestering van 28 min. 56 s.) de betrouwbaarheid .85 was. Een extra tijdsinvestering van 17 min. 22 s. leverde dus een stijging in de ICC van 0.18 op.

De resultaten voor de conditie 'paarsgewijze vergelijking' in tabel 5 en figuur 4 tonen eveneens een stijgende betrouwbaarheid naarmate er meer tijd geïnvesteerd werd. Bij een tijdsinvestering van 5 min. 18 s. (vijf beoordelingen per tekst), bedroeg de SSR .29. Een extra tijdsinvestering van 21 min. 11 s. (20 extra beoordelingen per tekst) ging gepaard met een toename in de SSR van .58. Bij 25 beoordelingen per tekst, een tijdsinvestering van 26 min. 29 s., bedroeg de SSR 0.87. De grenzen van .70 en .80 werden bereikt na respectievelijk twaalf en zeventien beoordelingen per tekst.

4.3 Vergelijking van stabiliteit rangorde tussen beide methoden bij gelijke tijdsinvestering

Tabel 6 geeft voor beide beoordelingsmethoden de Kendalls τ weer. De resultaten geven aan dat de rangorde in de conditie 'criterialijst' snel stabiel is. De correlatie tussen de scores op basis van de eerste beoordeling en de scores op basis van het gemiddelde punt op basis van de eerste en tweede beoordeling was reeds hoog (.69). De stabiliteit (en dus de betrouwbaarheid) neemt nog verder toe naarmate meerdere beoordelingen worden toegevoegd. De correlatie tussen de scores na vier en na vijf beoordelingen bedroeg .90.



Figuur 2

Duur van een beoordeling afhankelijk van het aantal beoordelingen per beoordelaar

Een vergelijkbare tijdsinvestering in de conditie 'paarsgewijze vergelijking' leidt ook tot een stabiele rangorde. Zo is de correlatie tussen de scores na respectievelijk vijf en elf beoordelingen per tekst .72. Daarnaast is de rangorde stabieler naarmate beoordelaars meer tijd investeerden. De stabiliteit nam toe van .72 tussen de rangorde op moment 1 en op moment 2, tot .92 tussen rangorde op moment 4 en op moment 5.

De resultaten van de stabiliteit van de rangordes in beide condities zijn erg vergelijkbaar, alhoewel paarsgewijze vergelijking net iets sneller lijkt. Dit impliceert dat met beide methoden met nagenoeg eenzelfde tijdsinvestering tot even stabiele rangordes leiden.

5. Discussie

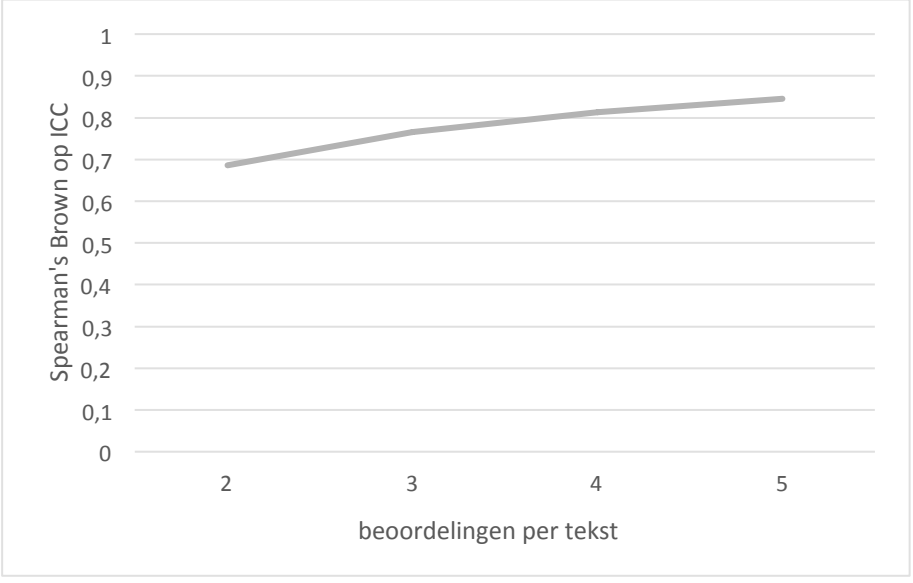
Om teksten betrouwbaar te beoordelen zijn verschillende methoden voorhanden. Meestal worden er criterialijsten gebruikt om teksten te beoordelen. Tegenwoordig wordt de alternatieve methode van paarsgewijze vergelijking meer en meer gebruikt. Er is echter weinig bekend over mogelijke verschillen in betrouwbaarheid en tijdsinvestering tussen beide methoden. Dit maakt het moeilijk een afweging te maken tussen één van beide methoden. In deze studie vergeleken we beide methoden wat de betrouwbaarheid en de benodigde tijdsinvestering betreft.

Tabel 4
Betrouwbaarheid en tijdsinvestering bij criterialijsten

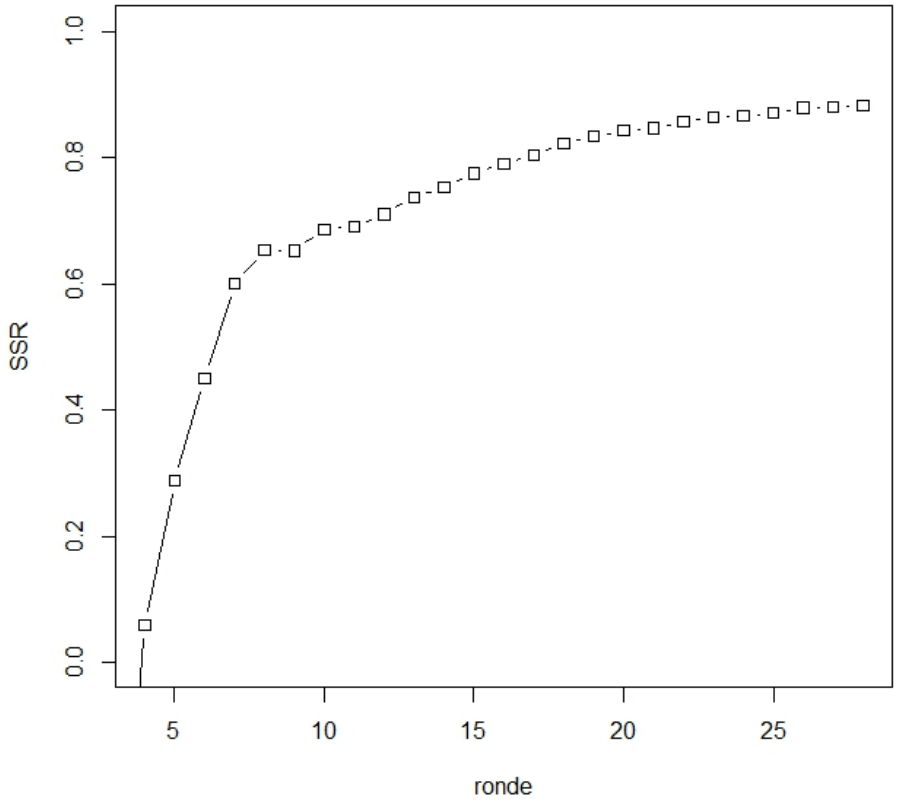
Aantal beoordelingen per tekst	ICC teruggerekend op basis van Spearmans' Brown	Tijd per tekst (in min.)
2	0.67	11 min. 34 s.
3	0.77	17 min. 22 s.
4	0.81	23 min. 09 s.
5	0.85	28 min. 56 s.

Allereerst liet deze studie zien dat bij beide methoden beoordelaars verschilden in hoe lang zij over een beoordeling per tekst deden. De variantie in de benodigde tijd per beoordeling was voor 21.9% te wijten aan beoordelaars in de conditie 'criterialijst' en voor 14.6% in de conditie 'paarsgewijze vergelijking'. In beide methoden nam de benodigde tijd af naarmate een beoordelaar al (meerdere) beoordelingen had gemaakt, al was deze afname sterker in de conditie 'criterialijsten'. Gezien Pollitt (2012b) concludeert dat snelheid niet gerelateerd is aan de kwaliteit van beoordelingen bij paarsgewijze vergelijking, lijkt het op basis van de resultaten in deze studie dat het loont om beoordelaars meerdere beoordelingen te laten maken. Er is meer onderzoek nodig om Pollitt's studie (2012b) te bevestigen en om meer zicht te krijgen in de kwaliteit en snelheid van beoordelingen wanneer beoordelaars via criterialijsten beoordelen.

Ten tweede liet deze studie zien dat er met criterialijsten minimaal twee beoordelingen per tekst nodig zijn om tot een betrouwbaarheid (in termen van ICC) te komen van .67. De betrouwbaarheid neemt toe naarmate een tekst vaker werd beoordeeld, tot .85 bij 5 beoordelingen per tekst. Bij paarsgewijze vergelijking waren er respectievelijk twaalf en zeventien beoordelingen per tekst nodig voor een betrouwbaarheid (in termen van SSR) van



Figuur 3
Betrouwbaarheid per ronde bij criterialijsten



Figuur 4

Evolutie in SSR per beoordelingsronde voor paarsgewijze vergelijking per beoordeling per tekst

.70 en .80. Deze resultaten zijn in lijn met de bevindingen van Verhavert et al. (2017). Ten derde toonde deze studie dat paarsgewijze vergelijking net iets sneller tot een stabiele rangorde komen, maar dit is minimaal. Deze resultaten spreken eerder onderzoek tegen waarin werd geconcludeerd dat paarsgewijze vergelijking veel minder (Pollitt, 2012b) dan wel veel meer (McMahon en Jones, 2015) tijd vergt dan werken met criterialijsten.

We willen opmerken dat de uitkomsten in deze studie sterk afhankelijk zijn van het design. Allereerst is de tijd die een beoordeling kost sterk bepaald door de lengte van de gebruikte criterialijst en de lengte van de teksten (Breland, 1983). Bijgevolg kunnen de bevindingen niet gegeneraliseerd worden naar andere schrijfproducten (bijvoorbeeld portfolio's) en andere criterialijsten (bijvoorbeeld meer beknopte criterialijsten). De bevindingen kunnen evenmin gegeneraliseerd worden naar andere inhoudsdomeinen, zoals wetenschappen, wiskunde of kunst. Vervolgonderzoek is nodig om de betrouwbaarheid en de benodigde tijdsinvestering bij criterialijsten en paarsgewijze vergelijking verder uit te diepen.

Ook kozen wij ervoor de beoordelaars niet te trainen. Eerder onderzoek toonde echter aan dat training van beoordelaars de betrouwbaarheid van beoordelingen beïnvloedt (Bacha, 2001; Bouwer & Koster, 2016; Weigle, 1999). Daartegenover stelt Pollitt (2012a) dat beoordelaars bij paarsgewijze vergelijking minder behoefte hebben aan training, gegeven het intuïtief vergelijken en beslissen. Er is echter een duidelijk gebrek aan onderzoek naar de rol van training. Gezien training ook een tijdsinvestering met zich meebrengt, moet nieuw onderzoek uitwijzen of training

Tabel 5
De evolutie in betrouwbaarheid en tijd in de conditie 'paarsgewijze vergelijking'

Aantal beoordelingen per tekst	SSR	Tijd per tekst (in min.)
5	0.29	5 min. 18 s.
10	0.69	10 min. 36 s.
12	0.71	12 min. 43 s.
15	0.78	15 min. 53 s.
17	0.80	18 min. 0 s.
20	0.84	21 min. 11 s.
25	0.87	26 min. 29 s.

Tabel 6
Kendalls τ correlatie bij criterialijsten en paarsgewijze vergelijking in functie van tijd per tekst

Criterialijsten			Paarsgewijze vergelijking		
Tijd per tekst (in min.)	Beoordelingen per tekst	Kendalls τ	Tijd per tekst (in min.)	Vergelijkingen per tekst	Kendalls τ
5 min. 47 s.	1		5 min. 18 s.	5	
11 min. 34 s.	2	.69	11 min. 39 s.	11	.72
17 min. 22 s.	3	.81	16 min. 57 s.	16	.82
23 min. 09 s.	4	.87	23 min. 18 s.	22	.89
28 min. 56 s.	5	.90	28 min. 36 s.	27	.92

voor snellere beoordelingen en/of een snellere stijging in betrouwbaarheid zorgt voor zowel de methode van criterialijsten als voor de methode van paarsgewijze vergelijking.

In deze studie kozen wij bovendien voor het gebruik van een criterialijst die gevalideerd en succesvol gebruikt is in eerder wetenschappelijk onderzoek. Deze studie nam bijgevolg de tijd om een dergelijke criterialijst te ontwikkelen niet mee. Voor onderwijsonderzoekers en de praktijk kan het interessant zijn juist de ontwikkel- en opzettijd mee in rekening brengen in de afweging tussen criterialijsten en paarsgewijze vergelijking.

Ook liepen wij in deze studie tegen een aantal moeilijkheden op. Zo is het nog niet duidelijk hoe vergelijkbaar de ICC en de SSR zijn. Wij kozen er daarom voor om ook de stabiliteit van de rangorde in kaart te brengen. Toekomstig onderzoek zal moeten uitwijzen hoe gelijkaardig de ICC en SSR zijn. Daarnaast zouden ook andere aspecten van betrouwbaarheid in overweging genomen moeten worden. Bijvoorbeeld, Jones en Inglis (2015) bekeken de betrouwbaarheid van paarsgewijze vergelijking van wiskundige probleemoplossingstaken door twee groepen beoordelaars onafhankelijk van elkaar te laten beoordelen. De Pearsons *r* tussen de rangordes van beide groepen, als maat voor de interbeoordelaarsbetrouwbaarheid, was hoog (.84). Een voordeel van deze benadering van betrouwbaarheid is dat het zou toelaten om de interbeoordelaarsbetrouwbaarheid in de conditie 'criterialijst' te vergelijken met deze in de conditie 'paarsgewijze vergelijking'.

Deze studie focuste op de efficiëntie van de methode van paarsgewijze vergelijking waarbij paren willekeurig werden samengesteld. De resultaten kunnen dus niet gegeneraliseerd worden naar adaptieve paarsgewijze vergelijking. Bij adaptieve paarsgewijze ver-

gelijking zijn er minder beoordelingen nodig voor dezelfde betrouwbaarheid (Pollitt, 2012b). Bramley (2015) toonde echter aan dat producten selecteren op basis van een voorlopige logit schatting mogelijk voor artificieel verhoogde betrouwbaarheden zorgt. Verder onderzoek is daarom nodig naar de efficiëntie van andere vormen van adaptieve paarsgewijze vergelijking, zoals werken met een gekalibreerde rangorde (Bramley, 2015; Heldsinger & Humphry, 2010).

De huidige studie kent enkele beperkingen. Allereerst liet de opzet van de studie bij criterialijsten (beoordelaars random verdeeld over teksten) niet toe om variantie gerelateerd aan beoordelaars te onderscheiden van andere error variantie (One-way Random ICC, Gwet, 2014). Om deze varianties in toekomstig onderzoek wel te kunnen onderscheiden, is het wenselijk een groep van beoordelaars alle teksten te laten beoordelen (Gwet, 2014; Shrout & Fleiss, 1979). Ten tweede is de benodigde tijd om een beoordeling af te ronden mogelijk beïnvloed door de procedure in beide beoordelingscondities. Beoordelaars in de conditie 'criterialijst' kregen random teksten met verschillende thema's ("Orgaandonatie", "Stress bij scholieren" en "Kinderen krijgen"). Verscheidene beoordelaars hadden dus al enkele beoordelingen afgerond (en dus ervaring opgebouwd met de criterialijst) voordat zij een tekst over "Stress bij scholieren" beoordeelden. We namen in de analyses van de benodigde tijd echter enkel de beoordelingen van de teksten "Stress bij scholieren" mee. Ook in de conditie 'paarsgewijze vergelijking' hadden de meeste beoordelaars reeds beoordelingen achter de rug van de teksten over "Orgaandonatie" en "Kinderen krijgen". Met andere woorden, de eerste vergelijkingen met teksten over "Stress bij scholieren" zijn dus niet de eerste beoordelingen die de beoordelaars maakten. We kunnen de tijdsinvesteringen die nodig zijn dus niet generaliseren. Verder onderzoek is nodig om de daling in de benodigde tijd voor een beoordeling in beide methoden accurater te schatten.

Een laatste punt is dat deze studie één belangrijk aspect van beoordelen niet mee nam, namelijk de validiteit. Beide beoordelingsmethoden zijn gebaseerd op verschillende assumpties van validiteit. Bij criterialijsten wordt validiteit nagestreefd door alle beoordelaars verschillende aspecten van schrijven te laten beoordelen. Bij paarsgewijze vergelijking vertrouwt men er op dat beoordelaars kwalitatieve teksten herkennen. Het feit dat meerdere beoordelaars betrokken zijn in het beoordelingsproces en elke tekst meerdere keren wordt vergeleken, zorgt ervoor dat het uiteindelijke kwaliteitsoordeel verschillende visies op tekstkwaliteit reflecteert (van Daal et al., 2016). Onderzoek naar de validiteit van paarsgewijze vergelijking zal in de toekomst meer aandacht moeten krijgen om ook dit te laten meewegen in de keuze voor één van beide methoden.

Ondanks deze tekortkomingen, is de huidige studie de eerste die de evolutie van betrouwbaarheid in functie van tijdsinvestering bekeek, zowel bij beoordeling via criterialijsten als bij paarsgewijze vergelijking. Daarnaast nam de studie de stabiliteit van de rangordes onder de loep. We concluderen dat paarsgewijze vergelijking een vergelijkbare tijdsinvestering van beoordelaars vraagt als de methode van criterialijsten.

Noten

1 De inbreng van de twee eerstgenoemde auteurs in dit artikel is gelijkwaardig

2 Verder rekenend met de 25.5 min. nodig per tekst voor 11 vergelijkingen, wordt dit als volgt berekend: ((25.5/11)*2)= 4 min. 40s. (We vermenigvuldigen met 2 omdat in elke vergelijking twee teksten bevat).

3 Dit wordt als volgt berekend: De totale tijdsinvestering voor alle schrijfproducten is (aantal schrijfproducten/2)*aantal vergelijkingen per tekst*seconden per vergelijking. (We delen het aantal schrijfproducten door twee omdat een vergelijking telkens informatie oplevert voor 2 schrijfproducten. Met andere woorden, om voor alle schrijfproducten 1 vergelijking te hebben zijn 500 (1000/2) vergelijkingen nodig in dit geval.) De tijdsinvestering per schrijfproduct is dan: de totale tijdsinvestering/aantal schrijfproducten. In het gegeven voorbeeld, geeft dit: 1000/2*11 vergelijkingen*195 s. (195 s.= 3 min. 15 s.). Wanneer dit totaal wordt gedeeld door 1000 en omgerekend naar minuten, geeft dit 17

min. 50 s. tijdsinvestering per schrijfproduct.

4 In het algemeen, wordt deze traditionele manier van omgaan met missing data (mean imputation) ontraden (Enders, 2010). Crameri, von Wyl, Koemeda, Schulthess & Tschuschke (2015) geven echter aan dat de impact van de manier waarop met missing data wordt omgegaan verwaarloosbaar is in het geval het aandeel missing data kleiner is dan 10%.

5 Om de power te verhogen hebben we beide methoden ook in een twee-intercepten model geschat. Hieruit bleek dat de parameterschattingen betreft de fixed effecten nagenoeg gelijk waren. De variantiecomponenten verschilden echter sterk tussen het twee-intercepten model en de twee aparte modellen. We rapporteren enkel de resultaten van de twee afzonderlijke modellen, omdat in deze modellen de residuen voor de beide methoden niet gepoold zijn.

6 Voor de eigenlijke formule van de one-way random ICC verwijzen we naar Gwet (2014, p. 197 e.v.).

7 De geattenueerde correlatie is .99, maar deze is wellicht vertekend doordat de ICC en de SSR niet goed vergelijkbaar zijn.

Literatuur

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95–104.

Baayen, R. H., & Milin, P. (2015). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, *29*(3), 371–383.

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, *41*(3), 466–481.

Bouwer, R., & Koster, M. (2016). *Bringing writing research into the classroom. The effectiveness of Tekster, a newly developed writing program for elementary students*. University of Utrecht, Utrecht.

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment*. Cambridge: Cambridge Assessment.

Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction*, *12*(2), 103–123.

Breland, H. M. (1983). The direct assessment of writing skill: A measurement review. *ETS Research Report Series*, *1983*(2). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1983.tb00032.x/full

Crameri, A., von Wyl, A., Koemeda, M., Schulthess, P., & Tschuschke, V. (2015). Sensitivity analysis in multiple imputation in effectiveness studies of psychotherapy. *Frontiers in Psychology*, *6*.

Crisp, V. (2013). Criteria, comparison and past experiences: how do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice*, *20*(1), 127–144. https://doi.org/10.1080/0969594X.2012.741059

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. *ETS Research Bulletin Series*, *1961*(2), i-93. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185. doi:10.1177/0265532207086780

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Greatorex, J. (2007). *Contemporary GCSE and A-level Awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work* (pp. 5–8). Cambridge: Cambridge Assessment. Retrieved from http://beta.cambridgeassessment.org.uk/Images/109755-contemporary-gcse-and-a-level-awarding-a-psychological-perspective-on-the-decision-making-process-used-to-judge-the-quality-of-candidates-work.pdf

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg: Advanced Analytics, LLC.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, *8*(1), 5–16.

Heldsinger, S. A., & Humphry, S. M. (2010). Using the method of pairwise comparison to obtain

reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–19. https://doi.org/10.1007/BF03216919

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

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics*, *89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*(1), 151-177.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 4th edition). Westport, CT: American Council on Education/Praeger.

Lesterhuis, M., Donche, V., De Maeyer, S., van Daal, T., Van Gasse, R., Coertjens, L., Verhavert, S., Mortier, A., Coenen, T., Vlerick, P., Vanhoof, J., & Van Petegem, P. (2015). Competenties kwaliteitsvol beoordelen: brengt een comparatieve aanpak soelaas? *Tijdschrift voor Hoger Onderwijs*, *33*(2), 55–67.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Comparative judgement as a promising alternative to score competences. In G. Ion & E. Cano (Eds.), *Innovative Practices for Higher Education Assessment and Measurement.* Hershey, PA: IGI Global.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, *19*(3), 246–276.

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research*, *64*(4), 147–156.

McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 368–389.

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

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1-11.

Tillema, M. (2012). *Writing in first and second language: Empirical studies on text quality and writing processes*. Netherlands Graduate School of Linguistics, Utrecht.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, 1–16.

Van Weijen, D. (2009). *Writing processes, text quality, and task effects: Empirical studies in first and second language writing*. Netherlands Graduate School of Linguistics, Utrecht.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2017, June). *Meta-analyse naar betrouwbaarheid van paarsgewijs beoordelen: Hoeveel vergelijkingen voor een betrouwbare rangorde?* Antwerpen: Onderwijs Research Dagen.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (ter perse). *Scale Separation Reliability: What does it mean in the context of comparative judgement?* Applied Psychological Measurement.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145–178.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Yeates, P., O'Neill, P., Mann, K., & W Eva, K. (2013). "You're certainly relatively competent': assessor bias due to recent experiences. *Medical Education*, *47*(9), 910–922.

Auteurs

Liesje Coertjens werkt als docent aan het Psychological Sciences Research Institute, Université Catholique de Louvain. **Marije Lesterhuis** werkt als doctoraatsonderzoeker bij het departement Opleidings- en Onderwijswetenschappen, Universiteit Antwerpen. **San Verhavert** werkt als doctoraatsonderzoeker bij het departement Opleidings- en Onderwijswetenschappen, Universiteit Antwerpen. **Roos Van Gasse** werkt als doctoraatsonderzoeker bij het departement Opleidings- en Onderwijswetenschappen, Universiteit Antwerpen. **Sven De Maeyer** werkt als hoogleraar aan de faculteit Sociale Wetenschappen, Universiteit Antwerpen.

Correspondentieadres: Liesje Coertjens, Place de l'université 1, 1348 Louvain-la-Neuve. Email: Liesje.Coertjens@uclouvain.be

Abstract

Judging texts with rubrics and comparative judgement: Taking into account reliability and time investment.

Writing researchers and practitioners both aim for reliable judgements with a minimum investment of time. This study focuses on two judgement methods, rubrics and comparative judgement. For each method, we studied how long it took to complete a judgement per text. Moreover, we examined how the reliability evolves in relation to the time spent judging. Judges were randomly attributed to the rubrics condition or the comparative judgement condition. In each condition, the same 35 texts were judged and time was tracked during this process. Results show that, when reliability is operationalized as the stability of the rank order, both methods require a comparable time investment to reach a stable rank order. Future research on the reliability and time investment should take into account the time needed for developing the rubric and to set up a comparative judgement assessment. Further research should also clarify whether the findings can be generalized to other texts and rubrics.

Key words: judging texts, comparative judgement, rubrics, reliability, time investment

Bijlage 1 - Taak leerlingen.

Overgenomen uit Van Weijen (2009) met kleine aanpassingen

Het leven van een scholier: Een zwaar bestaan vol stress? Of valt het wel mee? De Vlaamse Scholieren Koepel, organiseert een schrijfwedstrijd, speciaal voor scholieren uit 5-ASO. Jij doet ook mee. Je wilt absoluut winnen. Het winnende opstel wordt geplaatst in het maandblad Yeti, dat gelezen wordt door leerlingen van jouw leeftijd in heel Vlaanderen.

Doelstelling:
Schrijf een opstel waarin je je mening geeft en anderen overtuigt. De vraag luidt: "Hebben scholieren een zwaar bestaan vol stress? Of valt het wel mee?"

De jury stelt de volgende eisen:
1. Je artikel moet (ongeveer) een halve pagina lang zijn.
2. Je moet in je artikel je best doen om de lezers te overtuigen.
3. Je moet jouw standpunt goed onderbouwen.
4. Je artikel moet op een goede/logische manier zijn opgebouwd.
5. Je artikel moet er goed verzorgd uitzien (denk aan taalgebruik en spelling).
6. Je moet in je artikel ten minste twee fragmenten gebruiken uit de 'Bronnen' (zie volgende pagina). Die fragmenten moet je op een zinvolle manier verwerken in je artikel.

Het onderwerp van het opstel staat vast en werd als volgt omschreven in Yeti:
Heb jij weleens last van stress? Bezwijk je soms onder de druk van deadlines, bergen huiswerk of examens? Of vind je het onzin dat er scholieren zouden zijn die aan stress lijden? In de media wordt steeds meer aandacht besteed aan fenomenen als Burn-out, 'midlifecrisis' en andere stressgerelateerde klachten. Daarom wil de Vlaamse Scholieren Koepel in een speciale editie van Yeti uitgebreid aandacht besteden aan dit onderwerp. We willen graag van scholieren zelf horen wat ze vinden. Bepaal je mening en stuur ons je reactie!

Je hebt voor deze opdracht 25 minuten de tijd.

Succes!

Bronnen
"Stress is meer dan alleen een populaire actuele term, het is iets wat iedereen, jong en oud op verschillende gebieden kan ondervinden, […] Onderwijs is in deze tijd gericht op resultaten en presteren maar niet op lekker in je vel zitten en jezelf mogen zijn. Het leven zelf is de grootste leerschool! Om leerlingen nog beter voor te bereiden op het leven is het belangrijk om aandacht te besteden aan het ontstaan van stress, angst en black-outs. Met deze basiskennis zijn scholieren beter voorbereid op hun toekomst."
Bron: drs B.M.G.L. Kruitz. www.heyokah.com, 2005.

"Het leven van anderen lijkt altijd mooier. Natuurlijk ben je moe na een dag op school. Het vreemde is dat veel mensen vinden dat ze van werken niet moe zouden moeten worden. Je kunt ook teveel willen!"
Bron: Ank van der Campen, "Weekblad van Leraren", 2 oktober 1975.

Uit psychologisch onderzoek van de Universiteit van Antwerpen blijkt dat eerstejaars studenten het niet makkelijk hebben. Onderzoeker Jelte Wicherts: ..Eerstejaars zijn depressiever dan hun leefdtijdgenoten. Het is ook een periode van grote levensveranderingen, waardoor ze vatbaar zijn voor problemen.''
Bron: Marloes Zevenhuizen, www.standaard.be, 22 april 2004.

"Eén op tien leerlingen lijdt aan een ernstige vorm van faalangst. Ze halen slechte cijfers omdat ze bang zijn. Bang om te mislukken, bang om niet aan de verwachtingen te voldoen die ouders, leerkrachten of zijzelf vooropstellen. Ze hebben hoofdpijn, maagkrampen of hartkloppingen. Ze hyperventileren of zijn overgevoelig.

Onze sterk prestatiegerichte maatschappij werkt faalangst in de hand. Scholen staan onder druk om succesvolle leerlingen af te leveren. Dat veroorzaakt veel nutteloze stress zoals faalangst."
Bron: Klasse voor Leerkrachten 88, oktober 1998.

Drs. S. Beijne [een studentenpsycholoog verbonden aan de Universiteit Gent] schat dat zo'n vijf procent van de studenten serieuze stressproblemen heeft. Dat valt eigenlijk nog mee. "Maar," zo zegt Beijne, "ik wil benadrukken dat stress een natuurlijk verschijnsel is. Met stress moet je leren leven. Een beetje stress is nodig voor het leveren van goede prestaties."
Bron: Mieneke Scheele, www.panoplia.org, 2004.

Onderzoekers van de VUB vergeleken studenten die een deadline moesten halen met studenten die bloederige medische documentaires te zien kregen. De documentairekijkers hadden veel minder immunoglobuline – een stof die beschermt tegen ziekteverwekkers – in hun speeksel dan de deadlinewerkers. Goede stress is dus gezond, denken de onderzoekers.
Bron: Cicero; Universitair Medisch Centrum, maart 2002.

Bijlage 2- Criterialijst

Criterialijst dataverzameling november en december 2014 (grotendeels overgenomen uit Van Weijen, 2009)

"Argumentatief schrijven" aangepast aan de eindtermen derde graad ASO
1= Onvoldoende
2= Voldoende met leemtes
3= Voldoende
4= Goed
5= Schitterend

Onderdeel	Score				
1. Structuur					
1.1 Titel De tekst heeft een titel die duidelijk past bij de inhoud van de tekst.	1	2	3	4	5
1.2 Opbouw De tekst bevat een duidelijke indeling in: inleiding, argumentatie en conclusie.	1	2	3	4	5
1.3 Lay-out De tekst is overzichtelijk. Er is een duidelijke indeling in alinea's. Alinea's zijn gescheiden d.m.v.: witregels, inspringen of beginnen op een nieuwe regel.	1	2	3	4	5
1.4 Deelonderwerp Elke alinea heeft één eigen (deel)onderwerp.	1	2	3	4	5
1.5 Relaties tussen Alinea's Er is een heldere 'gedachtegang' tussen alinea's: op basis van de tekst zijn er duidelijk (gemakkelijk) coherentierelaties tussen alinea's te identificeren.	1	2	3	4	5

Onderdeel	Score				
1.6 Continuïteit Informatie die bij elkaar hoort, staat ook bij elkaar in de tekst.	1	2	3	4	5
2. Inhoud					
2.1 Inleiding In de inleiding wordt de stelling gepresenteerd én wordt eventueel ook duidelijk wat de mening van de schrijver is over de stelling.	1	2	3	4	5
2.2 Overtuigen Het is duidelijk waar de schrijver de lezer van wil overtuigen: een keuze vóór of tegen de gepresenteerde stelling.	1	2	3	4	5
2.3 Referenties De tekst bevat minimaal twee (delen van) referenties, die op een zinvolle manier verwerkt zijn in de tekst. Ze ondersteunen bijvoorbeeld de argumentatie of worden gebruikt als voorbeeld in de inleiding.	1	2	3	4	5
2.4 Verwijzingen (citeren uit referenties) De citaten uit de referenties zijn correct gemarkeerd in de tekst. Letterlijke citaten (tussen aanhalingstekens) en parafrases hebben allebei een bronvermelding.	1	2	3	4	5
2.5 Lezergerichtheid De tekst is goed te begrijpen voor een lezer die de opdracht niet kent. Er wordt bijvoorbeeld niet verwezen naar de opdracht voor de schrijftaak of naar de omgeving van de schrijver.	1	2	3	4	5
2.6 Lezerbetrokkenheid De lezer wordt duidelijk betrokken bij de tekst door voorbeelden die verwijzen naar het dagelijks leven of ervaringen die iedereen heeft.	1	2	3	4	5
2.7 Conclusie De tekst bevat een duidelijke conclusie, die aansluit bij de rest van de tekst, én waaruit de mening van de schrijver blijkt. Het is duidelijk dat de tekst hiermee wordt afgesloten.	1	2	3	4	5
3. Argumentatie					
3.1 Ondersteuning De argumentatie bestaat uit meerdere argumenten, die de mening van de schrijver ondersteunen.	1	2	3	4	5
3.2 Relevantie De argumentatie bevat niet teveel overbodige informatie, d.w.z. informatie die niet bijdraagt aan het ondersteunen van de mening van de schrijver.	1	2	3	4	5
3.3 Aanduiding Argumentatie De argumenten zijn duidelijk herkenbaar als argument; o.a. door het gebruik van constructies als "*daarom vind ik (niet) dat*…", "*ik vind/denk*…", "*ik ben het hier (niet) mee eens*" etc.	1	2	3	4	5
3.4 Referentiële en Coherentie relaties De referentiële en coherentie relaties zijn duidelijk als ze impliciet zijn, of expliciet gemarkeerd. Voorbeelden van markeringen zijn: *daarom, daardoor, dus, want, omdat, eerste, tweede, derde, daarna* etc.	1	2	3	4	5
4. Taal					
4.1 Grammatica en spelling De tekst bevat geen grammaticale en/ of spellingsfouten.	1	2	3	4	5
4.2 Interpunctie De leestekens zijn goed toegepast.	1	2	3	4	5
4.3 Stijl De toon en het woordgebruik zijn aangepast aan het doel en lezerspubliek van de tekst.	1	2	3	4	5

5. Welk cijfer geef je de tekst op een schaal van 0 tot 10?