Ten Geleide

Het rapport *Peil. Schrijfvaardigheid einde (speciaal) basisonderwijs 2018-2019* toont aan dat de schrijfprestaties van leerlingen achterblijven bij de ambities uit het Referentiekader Taal. Meer aandacht voor schrijfvaardigheid is dus wenselijk, maar het meten en beoordelen van schrijfvaardigheid kan hierbij een drempel vormen, voornamelijk in high-stakes toetsing. Comparatieve beoordeling biedt hier mogelijk een uitkomst voor. In het onderzoek van Carlijn van Herpt, Kirsten van Ingen, Marleen de Jonge en Pauline Roumans is een schrijftaak afgenomen in groep 7/8 en via comparatieve beoordeling in twee studies beoordeeld door 41 beoordelaars. Dit bleek een betrouwbare en valide methode die voor beoordelaars efficiënt uit te voeren was. Deze resultaten bieden perspectieven voor het gebruik van open schrijftaken in high-stakes toetsing.

Op Nederlandse middelbare scholen waar vakspecifiek onderwijs wordt aangeboden, is een belangrijke rol weggelegd voor het vak Nederlands als het gaat om schrijfvaardigheidsonderwijs. Immers, de schrijfvaardigheden die leerlingen bij Nederlands ontwikkelen, dienen zij zelf te transfereren naar schrijftaken bij andere vakken. Maar maken leerlingen die transfer wel? Om die vraag te beantwoorden lieten Emma Hilbrink en Jochem Aben twee havo-4-klassen een betoog schrijven bij het vak Nederlands, en twee bij het vak levensbeschouwing. Hoewel alle leerlingen dezelfde opdracht kregen, waren de betogen die leerlingen schreven bij het vak Nederlands van significant hogere kwaliteit. Dit resultaat biedt aanleiding om te overwegen of schrijfvaardigheidsonderwijs anders dan vakspecifiek ingericht dient te worden.

Gijs Leenders, Rick de Graaff en Marjo van Koppen beschrijven de totstandkoming van een instrument voor het evalueren van bewuste taalvaardigheid in het grammaticaonderwijs in de L1 (Nederlands) en de L2 (Engels en Duits) in vwo 4. Items zijn ontworpen bij drie grammaticale concepten: grammaticale functie, woordvolgorde en congruentie. Deze items richten zich ofwel op het verkrijgen van het 'juiste antwoord' (taalvaardigheid), ofwel op de redenering die aan dat antwoord voorafgaat (taalbewustzijn). Beschreven wordt hoe het taalvaardigheidsniveau momenteel wordt vastgesteld in het grammaticaonderwijs, hoe de items tot stand zijn gekomen, op basis waarvan ze zijn geselecteerd, en hoe drie gelijkwaardige toetsversies zijn samengesteld die gebruikt kunnen worden om het bewustetaalvaardigheidsniveau van leerlingen te evalueren.

Roel van Steensel bespreekt het proefschrift van Suzanne Bogaerds-Hazenberg, *Text structure instruction in Dutch primary education: Building bridges between research and practice.*

Joy de Jong bespreekt het proefschrift van Jeroen Steenbakkers, *Speelse toewijding: een pedagogisch-didactisch onderzoek naar schrijfstijlen formuleeronderwijs in klas 3 en 4 havo/vwo.*

Namens de redactie
Helge Bonset

Comparatieve beoordeling van schrijfvaardigheid in het primair onderwijs

Carlijn van Herpt, Kirsten van Ingen,
Marleen de Jonge & Pauline Roumans

Het rapport Peil.Schrijfvaardigheid einde (speciaal) basisonderwijs 2018–2019 toont dat de schrijfprestaties van leerlingen achterblijven bij de ambities uit het Referentiekader. Meer aandacht voor schrijfvaardigheid is dus wenselijk, maar het meten en beoordelen van schrijfvaardigheid kan hierbij een drempel vormen, voornamelijk in high-stakes toetsing. Comparatieve beoordeling biedt hier mogelijk een uitkomst voor. In dit onderzoek is een schrijftaak afgenomen in groep 7/8 en via comparatieve beoordeling in twee studies beoordeeld door 41 beoordelaars. Dit bleek een betrouwbare en valide methode die voor beoordelaars efficiënt uit te voeren was. Deze resultaten bieden perspectieven voor het gebruik van open schrijftaken in high-stakes toetsing.

Volgens het Referentiekader taal en rekenen (Doorlopende leerlijnen Taal en Rekenen, 2009) dienen leerlingen aan het eind van het basisonderwijs korte, eenvoudige teksten te kunnen schrijven over alledaagse onderwerpen of over onderwerpen uit de leefwereld (1F). Bovendien is het streven dat zoveel mogelijk leerlingen samenhangende teksten kunnen schrijven met een eenvoudige, lineaire opbouw over uiteenlopende vertrouwde onderwerpen (2F). De prestaties van leerlingen blijven echter achter bij deze ambities, blijkt uit het peilingsonderzoek schrijfvaardigheid (Inspectie van het Onderwijs, 2021). Meer aandacht voor schrijfvaardigheid is dus wenselijk. Hiervoor zijn instrumenten om de schrijfontwikkeling van leerlingen in kaart te kunnen brengen van belang. Het meten en beoordelen van schrijfvaardigheid is echter complex en kan een drempel vormen.

Schrijfvaardigheid kan gemeten worden via directe metingen, bijvoorbeeld open schrijftaken waarbij leerlingen zelf een tekst schrijven, maar ook via indirecte metingen, bijvoorbeeld door leerlingen aan de hand van meerkeuzevragen een geschreven tekst te laten reviseren. In de Centrale Eindtoets (CET)[1] werd schrijfvaardigheid via laatstgenoemde methode gemeten. Een schrijftaak met meerkeuzevragen is makkelijker te beoordelen, maar heeft een lagere validiteit (Pullens et al., 2013). Met name indruksvaliditeit, wat wil zeggen of de taak op het eerste gezicht lijkt te meten wat hij zou moeten meten (Hughes, 2002), is een belangrijke

factor in centrale toetsing, omdat niet alleen leerlingen, maar ook ouders en leerkrachten verwachten dat schrijfvaardigheid gemeten wordt door leerlingen een tekst te laten schrijven. Bovendien wijkt een indirecte meting af van de onderwijspraktijk waar schrijfvaardigheid op een actieve manier gemeten wordt aan de hand van open schrijftaken.

Directe metingen van schrijfvaardigheid kunnen een hogere (indruks)validiteit waarborgen en hebben daarom wellicht de voorkeur in een centrale of high-stakes toets[2], maar zijn minder eenvoudig te beoordelen. Het is namelijk lastig om bij open schrijftaken tot een betrouwbare meting te komen. Het begrip betrouwbaarheid speelt in centrale toetsing een grote rol: een toets moet consistente resultaten geven. Zo moet, in de hypothetische situatie dat een leerling twee keer exact dezelfde tekst schrijft, deze leerling twee keer hetzelfde resultaat behalen. Daarnaast moeten twee beoordelaars die dezelfde tekst nakijken tot dezelfde score komen. Beoordelaarseffecten, zoals een volgorde-effect of een normverschuiving, kunnen bij open schrijftaken optreden en de betrouwbaarheid van een schrijftaak beïnvloeden (Meuffels, 1994).

In dit artikel verkennen we de mogelijkheid om open schrijftaken centraal te toetsen met behulp van comparatieve beoordeling. Allereerst omschrijven we wat comparatieve beoordeling inhoudt. Daarna doen we verslag van de resultaten van ons onderzoek.

Beoordelingsmethoden

Er bestaan verschillende beoordelingsmethoden voor het beoordelen van open schrijftaken. De bekendste en waarschijnlijk meest gebruikte beoordelingsmethode is de analytische beoordeling, waarbij verschillende kenmerken van een tekst, zoals samenhang of woordgebruik, beoordeeld worden, bijvoorbeeld aan de hand van een beoordelingsformulier of een rubric (Weigle, 2002). Hiermee kan diagnostische informatie van de schrijfvaardigheid van een leerling op verschillende tekstkenmerken ingewonnen worden.

Hiertegenover staat de holistische beoordeling, waarbij een globaal oordeel over de tekst gevormd wordt. Holistische beoordeling kan leiden tot een lagere betrouwbaarheid dan analytische beoordeling, mede doordat scores tussen verschillende beoordelaars kunnen variëren als ze hun oordeel op andere (globale) waarden baseren (Hughes, 2002; Myford & Wolfe, 2003). Andere studies laten echter zien dat ook met holistische beoordelingsmethoden betrouwbare resultaten behaald kunnen worden. Daarnaast kan holistische beoordeling minder tijd kosten dan analytische beoordeling (Hughes, 2002; Weigle, 2002). Tot slot kan holistische beoordeling als meer valide beschouwd worden, omdat ze de ware reactie van een lezer het dichtst benadert (White, 1984).

Comparatieve beoordeling

Een voorbeeld van een beoordelingsmethode die gebaseerd is op een holistische aanpak, is comparatieve of paarsgewijze beoordeling. Bij comparatieve beoordeling van schrijfvaardigheid worden teksten die geschreven zijn door leerlingen met elkaar vergeleken. Een beoordelaar krijgt twee teksten te zien en kiest welke uitwerking beter is. Deze keuze is gebaseerd op een holistisch oordeel, waarbij de beoordelaar de tekst als geheel in beschouwing neemt. Iedere tekst wordt meermaals vergeleken, door verschillende beoordelaars en met verschillende combinaties van teksten. Uiteindelijk ontstaat er een rangorde van minder vaardig naar meer vaardig, waar per tekst een vaardigheidsscore op gebaseerd kan worden.

Het gebruik van comparatieve beoordeling kan bepaalde beoordelaarseffecten verminderen en tot betrouwbare resultaten leiden (o.a. Bouwer et al., 2023; Bramley, 2015). Daarmee biedt het wellicht een uitkomst voor de eerdergenoemde beoordelaarseffecten en betrouwbaarheidsproblematiek. Uit de meta-analyse van Verhavert et al. (2019) bleek dat een hoger aantal vergelijkingen per tekst kan leiden tot een hogere betrouwbaarheidscoefficiënt. Andere factoren, zoals het aantal beoordelaars of hun deskundigheidsniveau, leken geen rol te spelen.

Wat betreft efficiëntie worden in eerdere studies verschillende resultaten gerapporteerd. Zo blijkt iedere vergelijking snel te maken, alhoewel sommige beoordelaars het moeilijk vinden om tot een holistisch oordeel te komen en een keuze tussen twee teksten te maken (Van Daal et al., 2017). Ook hebben beoordelaars vaak weinig instructies nodig (Heldsinger & Humphry, 2010; Steedle & Ferrara, 2016). Wel zijn er meerdere beoordelaars nodig. Het aantal beoordelaars is afhankelijk van onder andere het aantal teksten dat beoordeeld moet worden en de beoordelingslast die men beoordelaars wil opleggen. Daardoor kan het aantal variëren van een paar tot zelfs tientallen beoordelaars. Zeker in centrale toetsing, waarbij de wens zou zijn om schrijfvaardigheid landelijk te meten, zouden veel beoordelaars nodig zijn om de beoordelingslast per beoordelaar te beperken. Dit kan hoge kosten met zich meebrengen (Steedle & Ferrara, 2016). Daarnaast laten verschillende recente studies zien dat de totale tijdsinvestering bij comparatieve beoordeling niet per se lager, maar eerder vergelijkbaar aan of zelfs hoger is dan de tijdsinvestering bij analytische beoordelingsmethoden (Coertjens et al., 2017; McMahon & Jones, 2015). Maar een individuele beoordelaar kan het proces toch als minder arbeidsintensief ervaren, doordat de beoordelingslast verdeeld wordt over meerdere beoordelaars en iedere vergelijking snel te maken is.

Tot slot biedt comparatieve beoordeling de kans om schrijfvaardigheid op een valide manier te toetsen, namelijk middels open schrijftaken. Lesterhuis (2018) onderzocht de validiteit van comparatieve beoordeling voor de beoordeling van tekstkwaliteit. Daarvoor vroeg ze beoordelaars waar zij hun keuzes op baseerden en waarom ze een tekst beter of slechter vonden. Beoordelaars focusten op verschillende tekstkenmerken, maar gaven wel vergelijkbare scores aan de teksten. Hieruit was te concluderen dat comparatieve beoordeling verschillen in focus van beoordelaars toestaat en dat het een valide beoordelingsmethode is voor de beoordeling van schrijfvaardigheid. Daarbij moet de kanttekening gemaakt worden dat, om schrijfvaardigheid valide te toetsen, het noodzakelijk is om per leerling meerdere schrijftaken af te nemen. Zo lieten Bouwer et al. (2023) zien dat minimaal vier schrijftaken nodig zijn om generalisaties te kunnen maken over het schrijfvaardigheidsniveau van een leerling. Desalniettemin bieden de vele voordelen perspectieven voor het gebruik van comparatieve beoordeling in centrale toetsing.

Dit onderzoek

In dit onderzoek hebben we de mogelijkheden van comparatieve beoordeling van open schrijftaken aan het eind van het primair onderwijs onderzocht in twee studies. Het doel was om te verkennen of deze beoordelingsmethode geschikt zou zijn voor het toetsen van schrijfvaardigheid met behulp van open schrijftaken in centrale of high-stakes toetsing, zoals de CET[1]. De resultaten uit eerdere onderzoeken waren veelbelovend, maar er was nog weinig bekend over het gebruik van comparatieve beoordeling in het (Nederlandse) primair onderwijs. Met dit onderzoek hopen wij bij te dragen

aan de kennis hierover.

Voor dit onderzoek is de volgende hoofdvraag geformuleerd: *Is comparatieve beoordeling een betrouwbare, efficiënte en valide beoordelingsmethode voor de beoordeling van de schrijfvaardigheid van leerlingen aan het eind van het basisonderwijs?*

We realiseren ons dat onze onderzoeksopzet zijn beperkingen heeft voor het meten van schrijfvaardigheid, maar verwachten dat de resultaten wel aanknopingspunten geven voor het inzetten van open schrijftaken in high-stakes toetsing.

Methode

Participanten en procedure

Aan deze studie namen in totaal 89 leerlingen uit groep 7 en 8 van vijf Nederlandse basisscholen deel. Deze basisscholen werden geselecteerd via het docentennetwerk van Cito en bevonden zich in Gelderland en Noord-Holland. De leerlingen kregen een schrijftaak voorgelegd met als doel het schrijven van een mail naar een vakantievriend, waarin ze de spelregels van verstoppertje uitlegden. De mail moest een aantal componenten bevatten: een introductie, uitleg over wat ze in de mail gingen schrijven, met hoeveel mensen je verstoppertje kunt spelen, waar je het kunt spelen, het doel van het spel, wanneer het afgelopen is en een afsluiting. Leerlingen moesten ongeveer 150 woorden gebruiken.

Voorafgaand aan de schrijftaak gaf de leerkracht een klassikale instructie waarin de taak werd toegelicht. Daarnaast kregen de leerlingen een kort filmpje over verstoppertje te zien om voorkennis te activeren. Na deze instructie gingen ze zelfstandig met de schrijftaak aan de slag. De mail werd door de leerlingen op papier geschreven, maar is ten behoeve van de beoordelingen overgetypt. Daarin zijn alle schrijf- en spelfouten overgenomen.

In totaal voldeden 22 teksten niet aan de opdracht, waardoor ze niet meegenomen zijn in de beoordelingen. Dit waren bijvoorbeeld teksten die minder dan 100 woorden bevatten of die niet over verstoppertje gingen. Dit bracht het totale aantal te beoordelen teksten op 67.

Voor de comparatieve beoordeling is gebruikgemaakt van het digitale programma Comproved, een samenwerking van de Universiteit Antwerpen, Universiteit Gent en het Interuniversity Microelectronics Centre (IMEC). In Comproved kunnen teksten die geschreven zijn door leerlingen geüpload worden. Vervolgens selecteert Comproved twee teksten voor een beoordelaar. Zodra de beoordelaar deze twee teksten vergeleken en beoordeeld heeft, selecteert Comproved twee nieuwe teksten. Dit wordt herhaald totdat alle teksten het gewenste aantal keer beoordeeld zijn. Het gewenste aantal vergelijkingen per tekst kan in Comproved per beoordelingssessie ingevuld worden.

Studie 1

In studie 1 namen 23 beoordelaars, waaronder leerkrachten uit het basisonderwijs, pabo-studenten en taalkundigen, deel aan een beoordelingssessie. In eerste instantie was de intentie om een homogene groep van leerkrachten te laten beoordelen, omdat zij ervaring hebben met het nakijken van schrijftaken binnen deze doelgroep en om de situatie in de onderwijspraktijk zo dicht mogelijk te benaderen. Maar het bleek lastig om voldoende leerkrachten basisonderwijs te vinden die deel konden nemen aan de studie. Volgens Verhavert et al. (2019) zou de deskundigheid van de beoordelaars geen invloed hebben op de betrouwbaarheid van een beoordelingssessie. Daarom is ervoor gekozen om ook andere experts deel te laten nemen als beoordelaar.

De beoordelingssessie startte met een korte instructie. Hierin werd uitgelegd dat

beoordelaars telkens twee willekeurig geselecteerde teksten te zien zouden krijgen met de vraag welke van de twee teksten ze beter vonden. Daarnaast werd ze verzocht om in steekwoorden te omschrijven waarom ze de ene tekst beter vonden dan de andere. Alle beoordelaars ontvingen bovendien een handleiding van Comproved en de schrijftaak die de leerlingen hadden gekregen. Beoordelaars werden geïnstrueerd om maximaal 10 minuten aan één vergelijking te besteden. Ze werkten zelfstandig aan de beoordeling en rondden de beoordelingssessie in geval van tijdnood op een later moment af.

Alle teksten werden 17 keer vergeleken met een andere tekst. Dit aantal werd gebaseerd op richtlijnen van Comproved en Verhavert et al. (2019), die concludeerden dat 17 vergelijkingen leiden tot een betrouwbaarheidscoefficiënt van minimaal 0,80, wat volgens de COTAN-richtlijnen voldoende is voor high-stakes toetsing (Evers et al., 2010).

Studie 2

De tweede studie week op een aantal punten af van de eerste studie. Deze tweede studie werd uitgevoerd om te onderzoeken of een hogere betrouwbaarheidscoëfficiënt behaald kon worden met een hoger aantal vergelijkingen. Het aantal vergelijkingen per tekst werd opgehoogd tot 30. Tevens hadden we, net als in studie 1, de intentie om een homogene groep van leerkrachten als beoordelaar te laten optreden, omdat de beoordelaars bij een daadwerkelijke afname ook leerkrachten zouden zijn. Ditmaal waren er voldoende gegadigden, namelijk 18 leerkrachten uit het basisonderwijs. Tot slot kregen de beoordelaars een gerichtere instructie mee voorafgaand aan de beoordeling. Deze instructie was gebaseerd op de omschrijvingen van de referentieniveaus schrijfvaardigheid uit het Referentiekader (Doorlopende Leerlijnen Taal en Rekenen, 2009). Beoordelaars kregen als instructie dat ze op basis van de informatie in de mail een beeld moesten kunnen vormen van het spel en dat daarvoor een samenhangende tekst met een logische opbouw belangrijk is. Ook werden de volledigheid van informatie, passend taalgebruik en correcte toepassing van de schrijfconventies als relevante aandachtspunten genoemd.

Betrouwbaarheid, efficiëntie en validiteit

Het doel van dit onderzoek was om de betrouwbaarheid, efficiëntie en validiteit van comparatieve beoordeling in kaart te brengen. De eerste variabele, betrouwbaarheid, werd gegenereerd door Comproved. Comproved berekent per beoordelingssessie de *Scale Separation Reliability* (SSR). Deze maat drukt uit in hoeverre een tekst na een extra vergelijking door dezelfde (of een vergelijkbare) groep beoordelaars op dezelfde plek in de schaal terecht zou komen. Daarmee voorziet ze ons van de interbeoordelaarsbetrouwbaarheid (Verhavert et al., 2019).

Efficiëntie werd gemeten door de gemiddelde tijd te berekenen die beoordelaars besteedden aan het maken van één vergelijking. Deze tijd werd bijgehouden door Comproved, dat registreerde hoeveel seconden verstreken tussen het ontvangen van twee nieuwe teksten en het aanklikken van de beste tekst.

Validiteit werd, gebaseerd op Lesterhuis (2018), gemeten door te inventariseren op welke aspecten van schrijfvaardigheid de beoordelaars hun oordeel baseerden. In Comproved zijn per tekstpaar en per tekst open antwoordvelden met plussen of minnen zichtbaar voor beoordelaars. Beoordelaars werden verzocht om hier maximaal drie positieve of negatieve feedbackpunten per tekst in te vullen. Deze feedbackpunten, zowel positief als negatief, werden door de onderzoekers ingedeeld in de zes kenmerken van schrijfvaardigheid volgens het Referentiekader (Doorlopende Leerlijnen Taal en Rekenen, 2009):

1. Samenhang
2. Afstemming op tekstdoel
3. Afstemming op publiek
4. Woordgebruik en woordenschat
5. Spelling, interpunctie en grammatica
6. Leesbaarheid.

Feedbackpunten die niet onder een van deze categorieën vielen, werden ondergebracht in de categorie overig.

Resultaten

Betrouwbaarheid

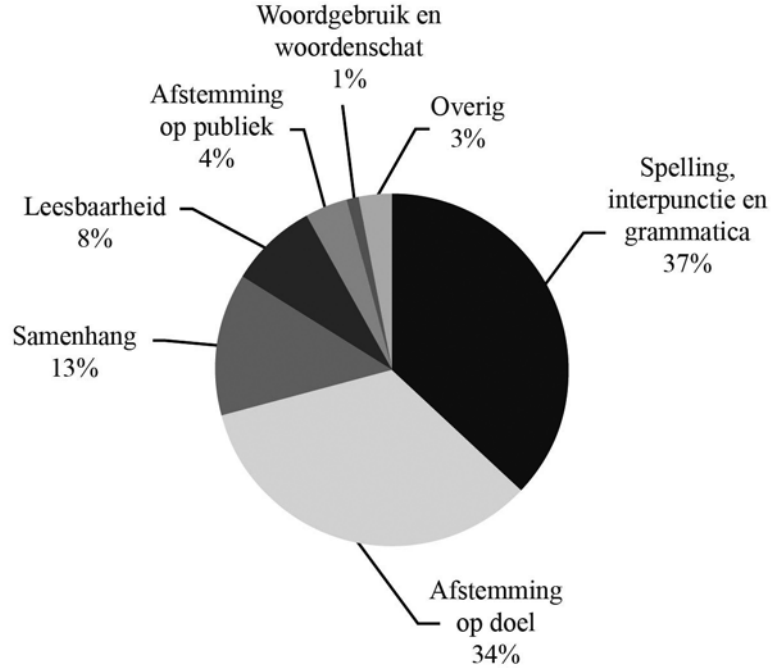
De betrouwbaarheidscoëfficiënt in studie 1 was 0,77. De betrouwbaarheidscoëfficiënt in studie 2 was 0,86. Deze laatste waarde is hoger dan 0,80 en daarmee volgens de COTAN-richtlijnen voldoende voor high-stakes toetsing (Evers et al., 2010). Ook is ze hoger dan in de papieren CET 2023, waar het onderdeel schrijfvaardigheid een betrouwbaarheid van 0,74 had (College voor Toetsen en Examens, 2023).

Efficiëntie

In studie 1 deden beoordelaars gemiddeld iets minder dan 3 minuten (M = 165,33 seconden) over het maken van één vergelijking. In die tijd gaven de beoordelaars ook feedbackpunten bij de teksten.

Studie 2 leidde tot vergelijkbare resultaten; beoordelaars deden gemiddeld iets minder dan 3 minuten (M = 169,35 seconden) over het maken van één vergelijking. In studie 2 is er ook gekeken naar de tijd die beoordelaars nodig hadden zonder het geven van feedbackpunten. Dit duurde gemiddeld anderhalve minuut (M = 96,10 seconden) per vergelijking.

In beide studies waren beoordelaars bovendien sneller in het maken van een vergelijking naarmate ze meer vergelijkingen hadden afgerond.



Figuur 1. Relatieve frequenties van feedback per categorie van schrijfvaardigheid in studie 1

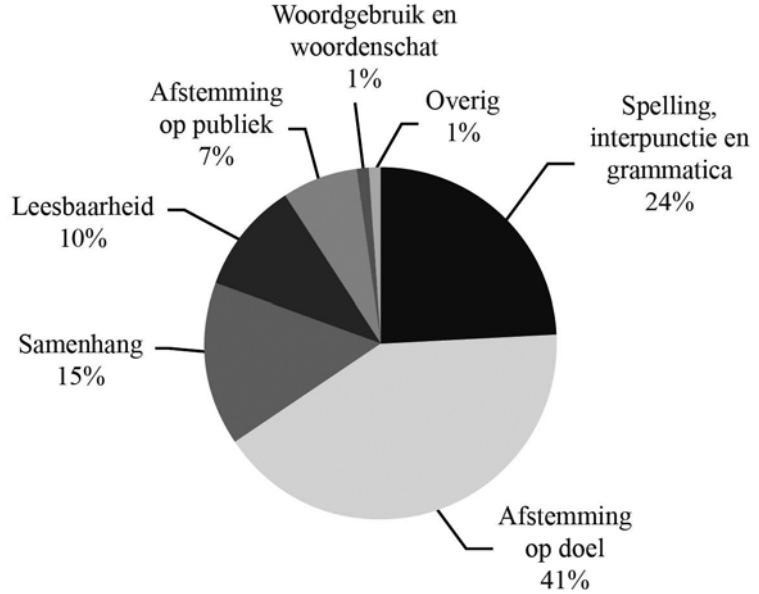
Validiteit

De relatieve frequenties van de feedback per categorie van schrijfvaardigheid van studie 1 zijn weergegeven in figuur 1. 37% van de feedbackpunten van beoordelaars viel in de categorie 'Spelling, interpunctie en grammatica'. 34% viel onder 'Afstemming op tekstdoel', 13% onder 'Samenhang', 8% onder 'Leesbaarheid', 4% onder 'Afstemming op publiek' en 1% onder 'Woordgebruik en woordenschat'. 3% van alle feedback viel in de categorie 'Overig'.

De relatieve frequenties van de feedback per categorie van schrijfvaardigheid van studie 2 zijn weergegeven in figuur 2. De meest frequent voorkomende feedbackcategorieën in studie 2 betroffen 'Afstemming op tekstdoel' (41%) en 'Spelling, interpunctie en grammatica' (24%). 'Samenhang' werd in 15% en 'Leesbaarheid' in 10% van de gevallen als feedback gegeven. 7% viel onder 'Afstemming op publiek' en 1% onder 'Woordgebruik en woordenschat'. 1% van de feedbackpunten viel in de categorie 'Overig'.

Discussie

In dit onderzoek is in twee studies onderzocht of comparatieve beoordeling een betrouwbare, efficiënte en valide beoordelingsmethode is voor de beoordeling van open schrijftaken van leerlingen aan het eind van het basisonderwijs. Hiervoor zijn 67 schrijftaken in twee beoordelingssessies beoordeeld in Comproved. Aan de eerste beoordelingssessie namen 23 beoordelaars deel en aan de tweede 18. Het uiteindelijke doel was om te verkennen of comparatieve



Figuur 2. Relatieve frequenties van feedback per categorie van schrijfvaardigheid in studie 2

beoordeling mogelijkheden biedt voor het gebruik van open schrijftaken in centrale of high-stakes toetsing.

De resultaten lieten zien dat de comparatieve beoordeling leidde tot een hoge betrouwbaarheid. In studie 1 werd een betrouwbaarheidscoëfficiënt van 0,77 behaald. Dit is volgens de COTAN-richtlijnen voldoende voor low-stakes toetsing, maar onvoldoende voor high-stakes toetsing (Evers et al., 2010). In studie 2 werd een hogere betrouwbaarheidscoëfficiënt van 0,86 gemeten, wat wel voldoende is voor high-stakes toetsing. Beide resultaten zijn hoger dan de betrouwbaarheid van het onderdeel schrijfvaardigheid in de papieren CET 2023, waarbij schrijfvaardigheid indirect met meerkeuzevragen werd getoetst en een betrouwbaarheid van 0,74 werd gemeten (College voor Toetsen en Examens, 2023).

De hogere betrouwbaarheidscoëfficiënt in studie 2 is mogelijkerwijs te verklaren doordat het aantal vergelijkingen per tekst verhoogd was van 17 naar 30. Dit komt overeen met Verhavert et al. (2019), die concludeerden dat de betrouwbaarheid steeg naarmate het aantal vergelijkingen omhoogging. Daarnaast was er in studie 2 sprake van een homogene groep beoordelaars, dit in tegenstelling tot de heterogene groep beoordelaars in studie 1. De groep beoordelaars in studie 2 bestond enkel uit bevoegde leerkrachten basisonderwijs, waardoor zij wellicht beter bekend waren met de doelgroep en onderwijsdoelen dan de heterogene groep beoordelaars uit studie 1, die naast leerkrachten ook bestond uit pabo-studenten en taalkundigen. Ondanks dat Verhavert et al. concludeerden dat het deskundigheidsniveau van de beoordelaars geen invloed had op de betrouwbaarheid van de metingen in hun studie, kunnen we niet uitsluiten dat dit in ons onderzoek wel een rol gespeeld heeft. De beoordelaars in studie 2 kregen ook gerichtere instructies mee. Vervolgonderzoek is nodig om te bepalen welke invloed deze factoren hadden op de betrouwbaarheidscoëfficiënt.

Mocht vervolgonderzoek aantonen dat de hogere betrouwbaarheidscoëfficiënt te danken is aan het hogere aantal vergelijkingen, zal dit consequenties hebben voor de efficiëntie van comparatieve beoordeling in high-stakes toetsing, aangezien een hoger aantal vergelijkingen meer tijd in beslag neemt. Desalniettemin was de beoordelingslast per beoordelaar laag. Beoordelaars hadden in beide studies gemiddeld minder dan 3 minuten nodig per vergelijking. Naarmate ze meer teksten vergeleken hadden, hadden ze minder tijd nodig. Ook behoefden de beoordelaars in beide studies weinig instructies voorafgaand aan de beoordeling. In studie 2 kregen beoordelaars vooraf meer en gerichtere instructies, maar er was geen uitgebreide trainingssessie nodig. De instructie duurde slechts enkele minuten. Dit komt overeen met eerdere studies (Heldsinger & Humphry, 2010; Steedle & Ferrara, 2016). Zoals eerder genoemd, kan onder efficiëntie ook de totale tijdsinvestering van de comparatieve beoordeling worden overwogen. Maar op basis van deze resultaten kunnen geen uitspraken gedaan worden over de totale tijdsinvestering bij comparatieve beoordeling ten opzichte van analytische beoordelingsmethoden.

Als laatste is gekeken naar de validiteit van comparatieve beoordeling. Vergelijkbaar met het proefschrift van Lesterhuis (2018), focusten beoordelaars op verschillende aspecten van schrijfvaardigheid. In studie 1 viel slechts 3% van de feedback onder de categorie overig. In studie 2 was dit 1%. Dit betekent dat respectievelijk 97% en 99% van de feedback betrekking had op componenten van schrijfvaardigheid uit het Referentiekader (Doorlopende leerlijnen Taal en Rekenen, 2009). Beoordelaars leken hun oordeel dus te baseren op de omschrijvingen van de referentieniveaus schrijfvaardigheid uit het Referentiekader, ondanks dat de beoordelaars in sessie 1 niet expliciet de instructie kregen om hierop te letten. Hieruit blijkt dat comparatieve beoordeling met het bijhorende holistische karakter een valide beoordelingsmethode is, aangezien beoordelaars de belangrijkste aspecten van schrijfvaardigheid in beschouwing namen.

In beide studies werden de categorieën 'Afstemming op tekstdoel' en 'Spelling, interpunctie en grammatica' het vaakst genoemd door beoordelaars. In studie 1 was 'Spelling, interpunctie en grammatica' de meest frequent voorkomende categorie. Hierbij moet echter de kanttekening gemaakt worden dat sommige teksten in studie 1 gemanipuleerd waren door spelfouten aan de teksten toe te voegen. Daardoor viel het percentage beoordelaars dat dit aspect noemde mogelijkerwijs hoger uit dan bij gebruik van de originele teksten. Desalniettemin leken 'Spelling, interpunctie en grammatica' en 'Afstemming op tekstdoel' in beide studies het zwaarst te wegen voor beoordelaars.

De aspecten waar beoordelaars op letten, zijn echter niet de enige factor die een rol spelen bij validiteit. Ook het aantal schrijftaken heeft invloed op het valide meten van schrijfvaardigheid (Bouwer et al., 2023). Vanwege het verkennende karakter van de huidige studie, hebben leerlingen slechts één schrijftaak gemaakt. Voordat schrijfvaardigheid opgenomen kan worden in centrale toetsing, zou verder onderzocht moeten worden hoe dit met meerdere schrijftaken gerealiseerd kan worden. Ook is het onderzoek uitgevoerd op relatief kleine schaal. Er bleven slechts 67 teksten over die meegenomen konden worden in de analyse. Vervolgonderzoek op grotere schaal zou een beter beeld kunnen geven van de betrouwbaarheid, efficiëntie en validiteit van comparatieve beoordeling.

Het doel van het huidige onderzoek was om de mogelijkheden voor comparatieve beoordeling van open schrijftaken te exploreren voor een mogelijke toepassing van deze beoordelingsmethode bij de centrale toetsing van schrijfvaardigheid, zoals in de CET. Inmiddels bestaat de eindtoets niet meer[I] en maken leerlingen in groep 8 in plaats daarvan een doorstroomtoets. Schrijfvaardigheid is een optioneel onderdeel in de doorstroomtoets. Op het moment van schrijven wordt schrijfvaardigheid in geen van de doorstroomtoetsen opgenomen. De resultaten van dit onderzoek laten echter zien dat, in het geval schrijfvaardigheid in een doorstroomtoets wordt opgenomen, dit niet per se gedaan hoeft te worden middels meerkeuzevragen, maar mogelijkerwijs ook kan met open schrijftaken. Zo kan een hogere indruksvaliditeit en een betere aansluiting bij de onderwijspraktijk gewaarborgd worden. Daarnaast kunnen de resultaten wellicht gebruikt worden voor onderzoek naar beoordeling van schrijfvaardigheid bij andere gestandaardiseerde centrale toetsen, bijvoorbeeld in het voortgezet onderwijs, waar schrijfvaardigheid op dit moment deel uitmaakt van de centrale examens Nederlands vmbo. Ook zijn de resultaten inzetbaar buiten centrale toetsing, bijvoorbeeld binnen een school of klas.

Conclusie

Dit onderzoek heeft aangetoond dat comparatieve beoordeling een veelbelovende methode is voor de beoordeling van open schrijftaken aan het eind van het basisonderwijs. Comparatieve beoordeling bleek betrouwbaar, valide en voor beoordelaars efficiënt uit te voeren. Deze beoordelingsmethode kan in de toekomst mogelijk bij gestandaardiseerde centrale toetsen of op schoolniveau ingezet worden.

Noten

1. Dit onderzoek is gestart toen de Centrale Eindtoets (CET) nog bestond. Inmiddels is het onderwijsstelsel gewijzigd. Vanaf schooljaar 2023-2024 is de eindtoets, en daarmee ook de CET, afgeschaft. In plaats daarvan maken leerlingen in groep 8 een doorstroomtoets. Zie voor verdere informatie de kamerbrief over deze wetswijziging (Ministerie van Onderwijs, Cultuur en Wetenschap, 2022).
2. Een high-stakes toets is een toets waarop een belangrijk besluit wordt genomen en die daarmee grote consequenties heeft voor leerlingen. Denk bijvoorbeeld aan de doorstroomtoets primair onderwijs en de centrale examens in het voortgezet onderwijs. Een low-stakes toets is een toets waarmee het leerproces geëvalueerd wordt en waar weinig consequenties aan verbonden zijn voor leerlingen. Denk bijvoorbeeld aan een tussentijdse toets in het voortgezet onderwijs.

Literatuur

Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., & De Maeyer, S. (2023). Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement. *Journal of Writing Research* 15(3), 498–518.

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment.* Cambridge Assessment. https://pdfs.semanticscholar.org/2afc/9ebe5c55c5f349e8d49f5906b4714da17483.pdf

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

College voor Toetsen en Examens. (2023). *Terugblik Centrale Eindtoets 2023.* College voor Toetsen en Examens.

Doorlopende leerlijnen Taal en Rekenen. (2009). *Referentiekader taal en rekenen. De referentieniveaus.* SLO.

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*(2), 1–20.

Hughes, A. (2002). *Testing for Language Teachers.* Cambridge University.

Inspectie van het Onderwijs. (2021). *Peil. Schrijfvaardigheid einde (s)bo 2018–2019*. Inspectie van het Onderwijs.

Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality.* Dissertatie Universiteit Antwerpen. Proefschriften UA-FSW.

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

Meuffels, B. (1994). *De verguisde beoordelaar: opstellen over opstelbeoordeling.* Thesis Publishers.

Ministerie van Onderwijs, Cultuur en Wetenschap. (2022). Wet van 9 februari 2022 tot wijziging van een aantal onderwijswetten in verband met aanpassingen op het gebied van de doorstroom van het basisonderwijs naar het voortgezet onderwijs en wijziging van de stelselinrichting van doorstroomtoetsen en toetsen verbonden aan leerling- en onderwijsvolgsystemen in het basisonderwijs [Kamerbrief]. *Staatsblad van het Koninkrijk der Nederlanden,* 135. Geraadpleegd via https://zoek.officielebekendmakingen.nl/stb-2022-135.html

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement,* 4(4), 386–422.

Pullens, T., Ouden, J. N. den , Herrlitz, W., & Bergh, H. H. van den (2013). Kan een meerkeuzetoets bijdragen aan het meten van schriftelijke taalvaardigheid? *Levende Talen Tijdschrift,* 14(2), 31–41.

Steedle, J. T., & Ferrara, S. (2016). Evaluating Comparative Judgement as an Approach to Essay Scoring. *Applied Measurement in Education,* 29(3), 211–223.

Van Daal, T., Lesterhuis, M., Coertjens, L, van de Kamp, M-T., Donche, V, & De Maeyer, S. (2017). The complexity of assessing student work using comparative judgement: The moderating role of decision accuracy. *Frontiers in Education,* 2(44), 1–11

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice,* 26(5), 541–562.

Weigle, S. C. (2002). Scoring procedures for writing assessment. In S. C. Weigle (Red.), *Assessing writing* (pp. 108–139). Cambridge University.

White, E. M. (1984). Holisticism. *College Composition and Communication,* 35(4), 400–409.

Carlijn van Herpt is toetsdeskundige taal bij Stichting Cito, waar zij een bijdrage heeft geleverd aan verschillende (taal) onderdelen van de Centrale Eindtoets en de doorstroomtoets van de overheid (DOE). Zij heeft een master General Linguistics en Onderwijswetenschappen behaald aan de Radboud Universiteit. Dit artikel is gebaseerd op haar masterscriptie, waarvoor zij onderzoek deed naar comparatieve beoordeling van schrijfvaardigheid in het primair onderwijs.
E-mail: <carlijn.vanherpt@cito.nl>

Kirsten van Ingen is toetsdeskundige Nederlands bij Stichting Cito. Zij heeft een bijdrage geleverd aan het onderdeel Schrijven van de Centrale Eindtoets en werkt op dit moment onder andere voor de examens vmbo voor de basisberoepsgerichte leerweg en de mbo-examens voor Nederlands 2F.
E-mail: <kirsten.vaningen@cito.nl>

Marleen de Jonge behaalde (cum laude) de master Applied Linguistics aan de Rijksuniversiteit Groningen en is momenteel werkzaam als toetsdeskundige taal bij Stichting Cito. Zij heeft als toetsdeskundige onder andere een bijdrage geleverd aan verschillende taalonderdelen van de Centrale Eindtoets, waaronder het onderdeel Schrijven, en de doorstroomtoets van de overheid (DOE).
E-mail: <marleen.dejonge@cito.nl>

Pauline Roumans is als toetsdeskundige taal en adviseur primair onderwijs werkzaam bij Stichting Cito. Zij heeft als toetsdeskundige een bijdrage geleverd aan verschillende taalonderdelen van de Centrale Eindtoets. Haar werk omvat verder het adviseren van het College van Toetsen en Examens (CvTE) met betrekking tot het beoordelen van de kwaliteit en de erkenning van doorstroomtoetsen en leerlingvolgsystemen voor het primair onderwijs.
E-mail:<pauline.roumans@cito.nl>