Ronde 5

Carlijn van Herpt, Kirsten van Ingen, Marleen de Jonge, Pauline Roumans
Cito
Contact: carlijn.vanherpt@cito.nl
kirsten.vaningen@cito.nl
marleen.dejonge@cito.nl
pauline.roumans@cito.nl

Comparatief beoordelen van schrijfvaardigheid in de bovenbouw po

1. Inleiding

Volgens het 'Referentiekader Taal en Rekenen' (Meijerink e.a. 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, zo blijkt uit het rapport 'Peiling Schrijfvaardigheid' (Inspectie van het Onderwijs 2021). Hoog tijd dus dat er meer aandacht besteed wordt aan schrijfvaardigheid.

Het beoordelen van schrijftaken kan hierbij een drempel vormen. Het is namelijk lastig om bij schrijftaken tot een betrouwbare meting te komen. Het begrip 'betrouwbaarheid' speelt in (*high-stakes*)-toetsing een grote rol: een toets moet consistente resultaten geven. Zo moet een leerling die twee keer exact dezelfde tekst schrijft hetzelfde resultaat behalen, en moeten twee beoordelaars die dezelfde tekst nakijken tot dezelfde score komen. Beoordelaarseffecten kunnen de betrouwbaarheid van een schrijftaak echter beïnvloeden (Meuffels 1994). Comparatieve of paarsgewijze beoordeling biedt wellicht een uitkomst voor deze betrouwbaarheidsproblematiek. We hebben de mogelijkheden van comparatief beoordelen van schrijftaken in het primair onderwijs, in combinatie met een niveaubepaling aan de hand van ankerteksten, verkend.

In deze workshop lichten we ons onderzoek toe en laten we de deelnemers zelf kennismaken met comparatieve beoordeling.

2. Comparatieve beoordeling

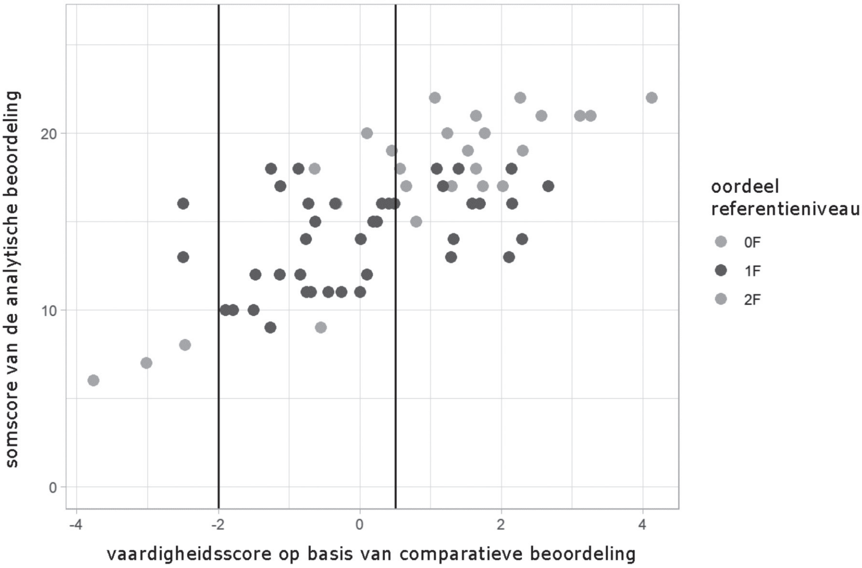
Bij comparatieve beoordeling van schrijfvaardigheid worden de teksten van leerlingen met elkaar vergeleken. Een beoordelaar krijgt twee teksten te zien en kiest welke uitwerking beter is. Deze keuze is gebaseerd op een holistisch oordeel, waarbij de beoordelaar de tekst als geheel bekijkt. Dit, in tegenstelling tot analytische beoordelingsmethoden, waarbij een beoordelaar een tekst op verschillende aspecten beoordeelt. Iedere tekst wordt meermaals vergeleken door verschillende beoordelaars en met verschillende combinaties van teksten. Uiteindelijk ontstaat een rangorde van ‘minder vaardig’ naar ‘meer vaardig’, waar een vaardigheidsscore op gebaseerd kan worden.

Het gebruik van comparatieve beoordeling kent meerdere voordelen. Zo leidt het tot betrouwbare resultaten (o.a. Bramley 2015) en biedt het de kans om schrijfvaardigheid op een valide manier te toetsen (Lesterhuis 2018), namelijk door middel van open schrijftaken. Daarnaast is het snel te verrichten en hebben beoordelaars vaak weinig instructies nodig (Heldsinger & Humphry 2010). Anderzijds kleven er ook enkele nadelen aan vast. Zo vinden sommige beoordelaars het moeilijk om tot een holistisch oordeel te komen en zijn er meerdere beoordelaars nodig om tot een betrouwbare beoordeling te komen (van Daal e.a. 2017).
Desalniettemin bieden de vele voordelen perspectieven voor het gebruik van comparatieve beoordeling in (*high-stakes*) toetsing.

3. Onderzoek

Het uitgangspunt in deze studie was een schrijftaak die is afgenomen bij leerlingen in groep 7 en 8 van vijf verschillende scholen. Leerlingen moesten een mail schrijven aan een vakantievriend, waarin ze de spelregels van hun versie van verstoppertje uitlegden. Deze schrijfproducten zijn in fase 1 van het onderzoek comparatief beoordeeld in twee beoordelingssessies middels het programma ‘*Comproved*’. Beoordelaars hadden gemiddeld drie minuten nodig voor iedere vergelijking. Naarmate ze meer vergelijkingen hadden gemaakt, ging het beoordelen sneller. In de eerste beoordelingssessie werd elke tekst ongeveer zeventien keer vergeleken door een heterogene groep van 23 leerkrachten en taalexperts. Dat leidde tot een betrouwbaarheidscoëfficiënt van .76. De tweede beoordelingssessie, waarin elke tekst ongeveer 30 keer werd vergeleken door een homogene groep van achttien leerkrachten, leidde tot een nog hogere betrouwbaarheid van .86. Volgens de COTAN-richtlijnen is dit voldoende voor (*high-stakes*) toetsing (Evers e.a. 2010). Ook is dit hoger dan in de papieren Centrale Eindtoets 2022, waarbij het onderdeel ‘schrijfvaardigheid’ een betrouwbaarheid van .71 had (CvTE 2022).

De vervolgstap in fase 2 van het onderzoek was om te bepalen of, en zo ja, hoe, de scores van comparatieve beoordeling vertaald kunnen worden naar een referentieniveau. In *high-stakes* toetsing is het immers noodzakelijk om te kunnen concluderen of een leerling een bepaalde norm – in dit geval: het referentieniveau 1F of 2F – behaald heeft. Voor dit doeleinde zijn de teksten allereerst analytisch beoordeeld middels een *rubric* die gebaseerd is op een *rubric* van Hogeschool Stenden (Berenst e.a. 2013). In de *rubric* werden de schrijftaken op enkele aspecten ingeschaald op referentieniveau. Hier volgde een score uit, op basis waarvan een referentieniveau aan een specifieke tekst toegekend kon worden. De twee beoordelingsmethoden zijn samengenomen om per vaardigheidsscore en referentieniveau een cesuur te kunnen leggen. De resultaten zijn te zien in Afbeelding 1.



Afbe eling versus de analytische beoordeling.

Afbeelding 1 toont een hoge mate van overeenstemming tussen de scores van de comparatieve beoordeling en de analytische beoordeling. De correlatie tussen beide scores is 0.72. Om te weten hoe de vaardigheidsscore van de comparatieve beoordeling vertaald kan worden naar een referentieniveau, is het noodzakelijk om cesuren te leggen bij de juiste vaardigheidsscores.

Dit is gedaan door te bepalen waar volgens ons de grens voor 1F en de grens voor 2F ligt op de vaardigheidsscore. We hebben gekeken binnen welke vaardigheidsscores de meeste uitwerkingen vallen die met de analytische beoordeling in een bepaald referentieniveau gescoord werden. Deze grenzen worden in Afbeelding 1 aangegeven door de twee zwarte verticale lijnen. Deze grenzen zijn inhoudelijk gevalideerd door te kijken of we konden verklaren waarom sommige uitwerkingen met de ene methode hoger of juist lager scoorden dan met de andere methode, en dus buiten de niveaugrenzen vielen (de zogenaamde *outliers*).

Vervolgens hebben we rond de cesuur ankerteksten (0F, 1F en 2F) geselecteerd die we aan een panel van experten hebben voorgelegd met de vraag of zij het eens waren met het toegekende referentieniveau. Uit deze valideringssessie bleek dat het panel zich in hoofdlijnen kon vinden in de toegekende referentieniveaus.

4. Conclusie

Uit ons onderzoek bleek comparatieve beoordeling, in combinatie met niveaubepaling door ankerteksten een valide, betrouwbare en praktische methode voor de beoordeling van schrijfopdrachten. Onze werkwijze zou een blauwdruk kunnen vormen voor de beoordeling van schrijftaken. We stellen ons dan een werkwijze voor, waarbij sprake is van een pre-test per schrijftaak waarin ankerteksten voor de verschillende referentieniveaus worden geselecteerd. Deze ankerteksten kunnen vervolgens in de daadwerkelijke afname in de set van te beoordelen teksten worden geplaatst, waardoor de uitwerkingen van de leerlingen kunnen worden vergeleken met de ankerteksten. Hiermee kan op een betrouwbare manier een referentieniveau aan een schrijftaak toegekend worden.

Referenties

Berenst, J., S. Faasse, R. Linthorst & M. Pulles (2013). ‘Observeren van presentaties, gesprekken en geschreven teksten in midden- en bovenbouw: 11 observatie-instrumenten om mondelinge en schriftelijke taalvaardigheid in beeld te brengen’. Leeuwarden/Groningen: NHL/RUG.

Bramley, T. (2015). ‘Investigating the reliability of adaptive comparative judgement’. Cambridge: Cambridge Assessment.

CvTE (2022). ‘Terugblik Centrale Eindtoets 2022’. Utrecht: CvTE.

Daal, T. van, M. Lesterhuis, L. Coertjens, M-T. van de Kamp, V. Donche & S. De Maeyer (2017). “The Complexity of Assessing Student Work Using Comparative Judgement: The Moderating Role of Decision Accuracy”. In: *Frontiers in Education,* 2 (44), p. 1-13.

Evers, A., W. Lucassen, R. Meijer & K. Sijtsma (2010). ‘COTAN Beoordelingssysteem voor de kwaliteit van tests’. Utrecht: NIP.

Heldsinger, S. & S. Humphry (2010). “Using the method of pairwise comparison to obtain reliable teacher assessments”. In: *Australian Educational Researcher,* 37 (2), p. 1-19.

Inspectie van het Onderwijs (2021). ‘Peil. Schrijfvaardigheid einde (s)bo 2018-2019’. Utrecht: Inspectie van het Onderwijs.

Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality.* Antwerpen: Universiteit van Antwerpen. [Proefschrift].

Meijerink, H.P. e.a. (2009). ‘Referentiekader taal en rekenen. De referentieniveaus’. Enschede: SLO.

Meuffels, B. (1994). *De verguisde beoordelaar*. Amsterdam: Thesis.