Meer beoordelaars per tekst

Startpunt

Bij het beoordelen van schrijfproducten is de betrouwbaarheid van de beoordeling een probleem. Uit onderzoek door Bouwer en Koster is duidelijk gebleken dat het vergroten van het aantal beoordelaars de betrouwbaarheid van de beoordeling verbetert.[1] De vraag is hoe dat in de praktijk vormgegeven kan worden, zonder dat de werkdruk onevenredig veel groter wordt. In de subgroep 'beoordelingen uitwisselen' van het project Schrijfvaardigheid in het Schoolexamen hebben we verschillende mogelijkheden onderzocht om met docenten van verschillende scholen hetzelfde schrijfproduct te beoordelen. Als het alleen gaat om de betrouwbaarheid van oordelen te verhogen, dan zou dit heel goed binnen één school kunnen plaatsvinden. Door uitwisseling tussen verschillende scholen, kunnen we ook veel te weten komen over de aard en het niveau van schoolexamens schrijfvaardigheid op andere scholen.

Welke acties hebben we ondernomen?

Het project is te verdelen in twee fasen: eerst hebben we verschillende opties om te beoordelen onderzocht en vergeleken (fase 1). Het doel van deze fase was om erachter te komen hoe groot de verschillen waren tussen verschillende beoordelaars bij verschillende manieren van beoordelen. In de tweede fase van het project hebben we een van die opties gekozen om verder uit te werken en komen we met een voorstel voor nader onderzoek en toepassing.

De eerste optie is analytisch beoordelen. Alle afzonderlijke beoordelingscriteria staan op papier, al dan niet in de vorm van een rubric. Voor de beoordelaar is het concreet waar hij op moet beoordelen. Voor beginnende beoordelaars is dit een duidelijke vorm van beoordelen.

Bij optie 2 gebruikten we drie teksten die een 5, een 7 en een 9 representeerden. De te beoordelen teksten werden ingeschaald door te vergelijken met die drie teksten (benchmarken).

Bij optie 3 kregen we met behulp van software genaamd *Comproved* steeds twee teksten aangeboden en moesten we aangeven welke we de beste vonden van de twee. *Comproved* is een tool die docenten helpt beoordelen door comparatieve vergelijkingen.[2] De tool structureert deze vergelijkingen.

Bij optie 2 en 3 kijken we op een holistische wijze naar tekstkwaliteit: het schrijfproduct wordt in zijn geheel beoordeeld.

---

[1] Bouwer & Koster (2016, blz. 79)

[2] www.comproved.com

Fase 1
In fase 1 zijn drie verschillende schrijfexamens gebruikt voor het vergelijken van drie beoordelingsopties.
We hebben de hierboven beschreven drie vormen van beoordeling toegepast om te kijken wat daarvan het resultaat was:

1. Analytisch met een beoordelingsmodel
2. Holistisch met behulp van ankerteksten
3. Holistisch door middel van comparatieve vergelijking met behulp van de software *Comproved*

Ons onderzoek betrof drie schrijfvaardigheidsexamens. Van elk examen hebben de vier groepsleden 10 teksten beoordeeld en vervolgens de beoordelingen onderling vergeleken.

1. Analytisch beoordelen van beschouwingen

Examen A is een schoolexamen 'gedocumenteerd schrijven'. Het examen is afgenomen in havo 5 op NSG Groenewoud. De leerlingen kregen de opdracht om een beschouwing van rond de 500 woorden te schrijven op basis van zelf verzamelde bronnen. Alle groepsleden hebben (kopieën van) tien teksten beoordeeld op basis van het analytische beoordelingsmodel dat al in gebruik was op de school waar het examen afgenomen werd. In dit beoordelingsmodel worden punten toegekend op basis van de tekststructuur en punten afgetrokken voor taalverzorging (zijn de verschillende functies van de inleiding correct uitgewerkt, wat is de kwaliteit van de argumentatie, is de tekst correct afgesloten, maakt de leerling gebruik van signaalwoorden en tot slot het taalgebruik). We hebben hier geen oordeel willen geven over de kwaliteit van het beoordelingsmodel maar zijn uitgegaan van de situatie zoals die op deze school is. De tien teksten zijn geselecteerd op basis van het onderwerp: er waren tien leerlingen die hun gedocumenteerde betoog over hetzelfde onderwerp hebben geschreven. Er heeft geen selectie vooraf plaatsgevonden om een goede verdeling te krijgen van goede en minder goede teksten.

2. Holistisch beoordelen van beschouwingen met ankerteksten

Examen B is afgenomen in vwo 6 op het Aventus Lyceum (vavo). De leerlingen kregen de opdracht om een beschouwing van 800-900 woorden te schrijven op basis van zelf verzamelde bronnen. Alle groepsleden hebben (kopieën van) tien teksten beoordeeld met behulp van ankerteksten. De ankerteksten en de te beoordelen teksten zijn geselecteerd door de examinator, waarbij zij globaal heeft gekeken naar mindere en betere teksten.

3. Comparatief beoordelen van betogen

Examen C is een betoog over een onderwerp naar keuze van 700 woorden gebaseerd op bronnen uit vwo 6 van gymnasium Bernrode. Alle groepsleden hebben tien willekeurig geselecteerde teksten beoordeeld in Comproved door deze teksten met elkaar te vergelijken. Bij deze software krijg je als beoordelaar telkens twee teksten aangeboden. Je geeft bij elke vergelijking aan welke tekst je het beste vindt. Alle groepsleden hebben elk 25 keer twee teksten met elkaar vergeleken. De tien te beoordelen teksten zijn willekeurig geselecteerd door iemand van de ICT-afdeling van gymnasium Bernrode.

De uitvoering

De uitkomsten van dit kleine onderzoek vormden een bevestiging van wat al bekend was: bij alle drie de methoden werd duidelijk dat de verschillen tussen de beoordelaars groot zijn.[3][4] Bovendien hadden de gebruikte methoden ook ieder hun eigen nadelen.

Bij examen A bleek dat het zonder meer inzetten van een door een ander ontwikkeld beoordelingsmodel lastig is, omdat het schema zelf geen duidelijkheid geeft over de interpretatie ervan. Iedere docent vertaalt de prestatie-indicatoren op zijn eigen manier. Bij gebruik zonder voorafgaand overleg zorgt dat voor heel verschillende beoordelingen. Als er in het model bijvoorbeeld staat “Tekstsoort aankondigen d.m.v. stelling: 5 punten”, geeft de beoordelaar dan 5 of 0 punten of kun je ook een kwaliteitsverschil maken in de aankondiging en daarmee ook 1, 2, 3 of 4 punten toekennen? Een ander voorbeeld is “Een sterk argument: 5 punten”: wat versta je nu precies onder een sterk argument?
Uit deze vergelijking blijkt wel dat de beoordelaars met het analytische beoordelingsmodel het goed eens zijn welke tekst het beste en minst goed is (betrouwbaarheid 0.83). Er is minder overeenstemming over welk cijfer daar dan bij hoort. Deze manier van beoordelen is erg arbeidsintensief: de beoordelaars investeren 20 tot 30 minuten per leerling.

	Beoordelaar 1	Beoordelaar 2	Beoordelaar 3	Beoordelaar 4 (examinator)
Laagste cijfer	5,7	3,1	4,7	4,0
Hoogste cijfer	8,8	7,2	7,8	8,2
Range	3,1	4,1	3,1	4,2
Gemiddelde	7,1	4,9	6,2	6,0
Betrouwbaarheid	10 teksten, 4 beoordelaars Betrouwbaarheid van de scores: 0.83			

Tabel 1: Beoordeling met behulp van een analytisch beoordelingsmodel

Bij examen B was het werken met ankerteksten (in dit geval) lastig omdat de schrijfopdracht niet duidelijk genoeg geformuleerd was. Dat maakt het ook lastig om te beoordelen wat de waarde is van de schrijfproducten. Ook daardoor waren er grote verschillen in de beoordeling. Deze manier van beoordelen is, hoewel minder dan analytisch beoordelen, ook nog arbeidsintensief: de beoordelaars investeren 15 tot 20 minuten per leerling.

	Beoordelaar 1	Beoordelaar 2 (examinator)	Beoordelaar 3	Beoordelaar 4
Laagste cijfer	4,5	4,1	4,5	5,5
Hoogste cijfer	8,0	7,5	7,5	9,0

[3] Bouwer e.a., 2020

[4] Bouwer & Koster, 2016

Range	3,5	3,4	3,0	3,5
Gemiddelde	**5,9**	**6,3**	**6,2**	**7,3**
Betrouwbaarheid	7 teksten, 4 beoordelaars Betrouwbaarheid van de scores: 0.63			

Tabel 2: Beoordeling met behulp van ankerteksten

Ook bij examen C waren nog verschillen zichtbaar. In onderstaande tabel is de ranking te zien op basis van Comproved en in kolom 2 de ranking gebaseerd op het oordeel van de examinator. De beoordelingen komen overwegend overeen, met uitzondering van twee opvallende verschillen bij de tekst die op basis van Comproved als beste wordt beoordeeld en de tekst die door de examinator als beste wordt beoordeeld.
Een nadeel dat wordt genoemd, is dat je na het werken met Comproved nog geen schoolexamencijfer vastgesteld hebt. We vroegen ons af hoe je deze ranking nu omzet in een cijfer, waarbij je wel rekening houdt met de continuïteit van het oordeel over de jaren heen. Je moet voorkomen dat je lagere cijfers gaat geven als de gemiddelde tekst beter is, of andersom. Tot slot de tijdsinvestering: het kost minder tijd om deze 25 vergelijkingen te maken (gemiddeld 40 minuten per beoordelaar) dan de tijd die per persoon is geïnvesteerd in examen A en B. Je kunt bij deze wijze van beoordelen de nakijkdruk dus beter verdelen over verschillende beoordelaars.
Helaas werd het werken met Comproved door sommige beoordelaars als saai ervaren. Als je inderdaad Comproved als sectie op je school gaat inzetten om te beoordelen, dan moet je meer vergelijkingen gaan maken dan er nu zijn gedaan. Voor betrouwbare resultaten moet elke tekst namelijk 10 keer worden vergeleken met een willekeurige andere tekst.

Ranking uit Comproved	Ranking door examinator
1	*10*
2	4
3	3
4	2
5	5
6	6
7	*1*
8	7
9	8
10	12
11	9
12	11
Betrouwbaarheid	10 teksten, 4 beoordelaars, in totaal 100 vergelijkingen Reliability: 0.62

Tabel 3: Beoordeling met behulp van *Comproved*

De algemene conclusies van de eerste stap van het onderzoek zijn dat optie A meer bijdraagt aan de betrouwbaarheid van de beoordeling dan optie B, terwijl beide opties wel een verhoging van de werkdruk opleveren: je moet als je teksten van een ander gaat beoordelen, immers méér teksten beoordelen. Daarnaast is de vraag, wat je met de opbrengst van de verschillende beoordelingen doet: ga je de cijfers middelen? Moet de examinator zijn oordeel herzien als de andere beoordelingen sterk afwijken? Examen C, de comparatieve vergelijking, levert ook een verhoging van de betrouwbaarheid op, maar zorgt daarnaast voor andere problemen zoals mogelijke problemen met het omzetten van een ranking naar een cijfer en daarnaast ook een verhoging van de werkdruk.

Fase 2

Welke acties je verder kan ondernemen

Op basis van de ervaringen uit fase 1 hebben we besloten om voor fase 2 van de uitvoering in te zetten op het beter voorbereiden van de beoordeling door onderling overleg (kalibreren). Wij kwamen op deze gedachte omdat afstemmen op elkaar steeds weer een terugkerend thema is. Want hoe interpreteer je de beoordelingscriteria en hoe zwaar weeg je de verschillende criteria? Dat kwam bij alle drie de methodes terug. Door de coronacrisis zijn we niet meer bij elkaar gekomen om dit uit te voeren.

Door te kalibreren zou het makkelijker moeten worden om de verschillen tussen beoordelaars te verkleinen, doordat vooraf is besproken wat van belang is en hoe de prestatie-indicatoren geïnterpreteerd moeten worden. Andriessen schrijft hierover: “Een kalibreersessie is een gezamenlijke bijeenkomst van examinatoren waarin één of meerdere werkstukken van studenten worden besproken. De sessie wordt geleid door een moderator en de resultaten worden vastgelegd door een notulist. Een sessie duurt meestal 2,5 uur wanneer er één werkstuk wordt besproken, en langer wanneer er meerdere werkstukken aan de orde zijn. Een optimale groepsgrootte is tussen de 6 en 12 deelnemers.”[5] Deze wijze van afstemmen vergt vooraf dus een (tijds)investering van de betrokken docenten, maar we verwachten dat we daarna sneller kunnen nakijken, omdat duidelijk is wat wel en wat niet goed is. Een daadwerkelijke uitwisseling van de beoordelingen is dit natuurlijk niet; het “doel van de sessie is niet om te komen tot een gezamenlijk oordeel over het werkstuk dat wordt besproken” (aldus Andriessen). Het kan wel een goed hulpmiddel zijn om met collega’s, zowel uit de sectie als van andere scholen, tot een werkbare verbetering te komen in de beoordeling van schrijfproducten.

Een andere mogelijke optie om verder te gaan is door Comproved in te zetten als middel voor peer feedback. Niet de docenten doen dan het werk, maar de leerlingen zelf. Je kunt dan assessments opzetten per klas in plaats van per leerjaar, dat maakt het minder werk. Op deze manier internaliseren leerlingen de prestatie-indicatoren en krijgen ze voorbeelden.

Onze vraag was hoe we de betrouwbaarheid van de beoordeling kunnen verbeteren door het aantal beoordelaars te vergroten, zonder dat de werkdruk onevenredig veel groter wordt. De verschillende mogelijkheden die wij hebben onderzocht (analytische

---

[5] Andriessen (2015)

beoordeling, beoordeling met behulp van ankerteksten en beoordeling door comparatieve vergelijking) blijken helaas wel te zorgen voor een (flinke) verhoging van de werkdruk. In plaats van werken met meer beoordelaars, zouden kalibreersessies (op de eigen school of met verschillende scholen) en/of het inzetten van peerfeedback wel kunnen zorgen voor een verbetering in de betrouwbaarheid van beoordeling van leerlingteksten. Wat ons betreft is dit de moeite waard om alsnog te onderzoeken.

Anna Hermans
Margreeth Krommendijk
Sarah Mulders
Annemieke de Roo

Bibliografie

Andriessen, Daan (2015). *Handreiking Kalibreersessies.* Hogeschool Utrecht, 16 januari 2015

Bouwer, R., Koster, M. (2016). *Bringing writing research into the classroom*. Utrecht University Repository

Bouwer, R., De Smedt., F. Lesterhuis De Mayer, S., Van Keer, H. (2020). *A comparative approach to the assesment of writing: the reliability and validity of comparative judgment and benchmark ratings*

Expertgroep protocol (2014). *Beoordelen is mensenwerk. Bevindingen over de wenselijkheid en mogelijkheid van een gezamenlijk protocol voor het beoordelen van (kern)werkstukken.* Vereniging Hogescholen

Heijden, C. van der (2018). Roeping versus regels. Professionals aan de lopende band. *De groene Amsterdammer.* Jaargang 142, 44-45.

Schoot, Menno van der (2020). *Een scriptiebeoordeling past niet in een schema*. Science Guide, 11 februari 2020.

Bos-Horstink, M., Soeting, J., Sugito, A., Uil-Hoogerwaard, W., Wouters, M. (2017). *Hoe maak ik goede opdrachten. Toetskwaliteit in de praktijk.* Wilp: Teelen

Wheadon, C., Patrick Barmby, Daisy Christodoulou & Brian Henderson (2019). *A comparative judgement approach to the large-scale assessment of primary writing in England, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2019.1700212*