Op naar een beter Schoolexamen Schrijfvaardigheid
Renske Bouwer, 28-08-2020

Schrijfvaardigheid is een belangrijk domein binnen het vak Nederlands en daarbij hoort ook een weldoordacht examenprogramma dat het schrijfniveau van leerlingen aan het einde van het voortgezet onderwijs goed in kaart brengt. Zo'n 40 docenten Nederlands van 20 scholen verspreid door heel Nederland presenteren hier hoe ze hun schoolexamen schrijfvaardigheid in het afgelopen schooljaar hebben verbeterd. Ze hebben geëxperimenteerd met meer schrijftaken in het examen, meer beoordelaars per tekst, geïntegreerde lees- en schrijftaken, een beoordelingstraining, vergelijkend beoordelen en een praktijkexamen voor schrijven. In deze bijdrage bespreken we de noodzaak van deze veranderingen in het licht van de wetenschappelijke kennis over het valide en betrouwbaar toetsen en beoordelen van schrijfvaardigheid. Met deze wetenschappelijke kennisbasis en de rijke praktijkvoorbeelden hopen we andere scholen te inspireren om ook hun PTA aan te pakken en zo een kwaliteitsslag te maken met de schoolexamens schrijfvaardigheid.

Het belang van een goed examenprogramma
Het is essentieel dat leerlingen zich schriftelijk goed kunnen uitdrukken. Een goede schrijfvaardigheid bevordert succes op school, de arbeidsmarkt en in de maatschappij (Graham & Perin, 2007). Schrijven geeft leerlingen een stem waarmee ze zich kunnen mengen in maatschappelijke discussies en anderen kunnen informeren, overtuigen en vermaken. In vervolgstudies is schrijven een noodzakelijk middel tot leren. Dit alles maakt schrijven een belangrijk onderdeel van het schoolvak Nederlands, wat vraagt om goed onderwijs en een goed examenprogramma.

Het domein schrijfvaardigheid wordt getoetst in de schoolexamens. Dat geeft scholen de mogelijkheid om leerlingen vaker teksten te laten schrijven over een langere periode. Scholen bepalen hierbij zelf welke teksten er geschreven worden en hoe deze teksten beoordeeld worden. Dit leidt tot vaak grote verschillen in de manier van examinering. Nu zegt variatie tussen de examens niet noodzakelijkerwijs iets over de kwaliteit, maar volgens een recent adviesrapport van Levende Talen Nederlands in samenwerking met Nederlands Nu (2018) is een betere kwaliteitsborging van de schoolexamens voor schrijfvaardigheid wel degelijk hard nodig. Scholen hebben volgens hen vooral behoefte aan meer kennis over het toetsen en beoordelen van schrijfvaardigheid en het creëren van meer eenheid in (de kwaliteit van) de schoolexamens. Dit project speelt in op deze behoefte door het delen van praktijkvoorbeelden waarin een team van docenten en wetenschappers veranderingen hebben doorgebracht in de schoolexamens voor schrijfvaardigheid.

Toetsen van eindtermen
Bij de uitwerking van het examenprogramma voor schrijfvaardigheid dienen scholen te toetsen of leerlingen voldoen aan de eindtermen voor schrijfvaardigheid, zoals geformuleerd door het College voor Toetsen & Examens. Deze eindtermen zijn online te raadplegen via https://www.examenblad.nl/examen/nederlands-vwo-2/2020. Hiernaast zijn scholen sinds 2011 ook verplicht om in het examen rekening te houden met het referentiekader voor schrijven op niveau 3F (havo) en 4F (vwo), zoals vastgesteld door de commissie Meijerink (2008). In dit referentiekader is ook op taakniveau gespecificeerd wat leerlingen aan het einde van het voortgezet onderwijs moeten kunnen op het

gebied van schrijven. In het Programma van Toetsing & Afsluiting (PTA) leggen scholen vooraf vast hoe ze deze doelen toetsen, en welke opdrachten en beoordelingsprocedures ze daarvoor gebruiken.

Om scholen te ondersteunen bij het opstellen van het PTA dat voldoet aan de eindtermen en doelen uit het referentiekader adviseert het SLO (2012) om voor het schoolexamen havo/vwo te werken met een schrijfdossier dat bestaat uit meerdere schrijfopdrachten waarbij leerlingen over diverse onderwerpen schrijven en waarbij zowel de tussentijdse als gereviseerde versies van de teksten worden verzameld. Maar is dat ook waar scholen in de praktijk voor kiezen? Scholen hebben immers de ruimte om in de uitwerking van het schoolexamen hun eigen accenten aan te brengen. Het is dan ook relevant om nader te bekijken in hoeverre de schoolexamens van elkaar verschillen en wat dit betekent voor de kwaliteit ervan.

Het ene schoolexamen is het andere niet
Uit een inventarisatie aan de start van dit project kwamen grote verschillen in de schoolexamens naar voren. Zo bleken lang niet alle 20 deelnemende scholen meerdere schrijftaken bij hun leerlingen af te nemen. Het merendeel van de 20 deelnemende scholen schrijfvaardigheid gaf aan maar een enkele schrijftaak te gebruiken, en maar 8 scholen verzamelen meerdere teksten per leerling, variërend van 2 tot 4 teksten. Als we kijken naar de tekstsoorten die in het examen aan bod komen, dan valt op dat op de meeste scholen leerlingen betogen en beschouwingen schrijven. Leerlingen mogen daarbij vaak zelf kiezen welk van de twee teksten ze willen schrijven. Er zijn ook scholen die leerlingen andere type teksten laten schrijven, zoals uiteenzettingen (3 scholen), essays (5 scholen), zakelijke teksten (4 scholen), boekrecensies (3 scholen), of (autobiografisch) verhalen (2 scholen).

Alle scholen rapporteren bij de schrijfopdrachten gebruik te maken van bronnen. In de meeste gevallen zijn dit bronnen die vooraf aan leerlingen worden gegeven (11 scholen). In de andere gevallen zoeken leerlingen hun eigen bronnen (5 scholen) of is het een combinatie van eigen en gegeven bronnen (4 scholen). Verder is er grote variatie tussen scholen in de voorbereiding van de teksten, zoals het aantal oefenmomenten die er worden geboden, en de mogelijkheid tot feedback van docent/peer en het reviseren van de tekst. Op de meeste scholen schrijven de leerlingen hun teksten op de computer, waarbij ze gebruik mogen maken van de spellingscontrole. Er zijn ook scholen waar leerlingen hun teksten met pen en papier schrijven, zonder extra hulpmiddelen.

Scholen lijken op het eerste oog minder te verschillen in de manier waarop de kwaliteit van teksten wordt beoordeeld. Op alle scholen is het de eigen docent (met soms een tweede onafhankelijke beoordelaar bij twijfelgevallen) die de teksten beoordeeld aan de hand van een analytisch beoordelingsschema. Deze beoordelingsschema's worden door elke school zelf ontwikkeld samen met de sectie Nederlands. Er worden weinig tot geen beoordelingsinstrumenten tussen scholen uitgewisseld. Aangezien het referentiekader voor elke school als uitgangspunt zou moeten dienen, kan verwacht worden dat de beoordelingsschema's dezelfde criteria bevatten. Toch blijken er in de praktijk aanzienlijke verschillen te zijn in het aantal beoordelingscriteria, het aantal punten dat er per criterium kan worden toegekend en hoe de losse onderdelen optellen tot een totaalscore voor tekstkwaliteit.

Deze verschillen in de manier van toetsing en beoordeling, maken dat het ene examen schrijfvaardigheid het andere niet is. Bovendien blijken docenten vaak onvoldoende goed op de hoogte van wetenschappelijke kennis over valide en betrouwbare toetsing en beoordeling en de mogelijkheden om deze kennis naar de onderwijspraktijk te vertalen. Zo geven de meeste docenten aan nog onvoldoende kennis en ervaring te hebben met holistisch en vergelijkend beoordelen, of met het afnemen van meerdere schrijftaken en het beoordelen van schrijfprocessen. Terwijl er juist op deze terreinen in de afgelopen jaren veel relevant praktijkgericht onderzoek is gedaan.

Aanknopingspunten voor verbetering uit de wetenschap
Onderzoek laat zien dat er ten minste twee problemen zijn met de huidige manier van examineren. Een eerste probleem is het bepalen van de kwaliteit van teksten. Het oordeel van één docent blijkt uit

onderzoek onvoldoende betrouwbaar voor het maken van zak-en-slaag beslissingen. Al sinds de eerste studies in 1960 wijzen onderzoeksresultaten op grote verschillen tussen beoordelaars (zie bijvoorbeeld Diederich, French & Carlton, 1966; McColly, 1970; Huot, 1990). Zelfs ervaren docenten blijken niet unaniem te zijn in hun oordeel. In de praktijk hangt de beslissing in een examen dus sterk af van wie de tekst heeft beoordeeld. Ook de wijze van beoordelen doet ertoe. Analytische beoordelingsschema's die nu veelvuldig in de praktijk worden gebruikt om docenten op dezelfde manier naar teksten te laten kijken, bieden geen garantie voor kwaliteitsvolle oordelen. Hoe specifiek het schema ook is, er blijft altijd ruimte over voor docenten om een schema op zijn of haar eigen manier in te vullen. Er zijn ook docenten die zelfs bij beoordelingsschema's nooit heel hoge of heel lage cijfers geven, of die bij meer dan vijf spelfouten of een onduidelijk handschrift de tekst helemaal niet verder bekijkt. Het is daarnaast de vraag in hoeverre een somscore van losse onderdelen iets zegt over de kwaliteit van een tekst als geheel (Sadler, 2009). Dat heeft ook zijn weerslag op leerlingen: ze krijgen weliswaar feedback over criteria waar ze goed of minder goed op scoren, maar leren ze ook iets over de communicatieve effectiviteit van hun tekst als geheel? Ook zijn analytische schema's vaak zeer taakafhankelijk, de oordelen zijn nauwelijks te generaliseren naar andere teksten (Bouwer & Koster, 2016; van den Bergh, De Maeyer, van Weijen, & Tillema, 2012). Om goede uitspraken te vellen over tekstkwaliteit zijn er dus meerdere beoordelaars per tekst nodig, die (ook) op andere manieren dan met analytische schema's tot een oordeel komen.

Een tweede probleem met het huidige schoolexamen is dat op de meeste scholen het schrijfexamen bestaat uit het schrijven van één tekst, terwijl uit onderzoek blijkt dat we op grond van één geschreven tekst nauwelijks uitspraken kunnen doen over de algemene schrijfvaardigheid van een leerling (van den Bergh et al., 2012). Net als bij andere complexe vaardigheidsvakken, zoals bijvoorbeeld wiskunde, zijn hiervoor meerdere items in de toets nodig. Nu is het schrijven van een tekst natuurlijk wel een grote taak waar veel uit af valt te leiden, maar het is als voorspelling van schrijfvaardigheid nog steeds veel te beperkt. De kwaliteit van een tekst hangt namelijk niet alleen af van hoe goed de leerling over het algemeen kan schrijven, maar ook hoeveel ze weten van het specifieke onderwerp of het genre. De correlatie tussen verschillende schrijftaken is dan ook vaak erg laag. Een goede (of slechte) prestatie op de ene schrijftaak zegt dus niet zoveel over hoe goed de schrijver presteert op een andere schrijftaak.

Kortom, als we met het schoolexamen kwaliteitsvolle beslissingen willen maken over de schrijfvaardigheid van individuele leerlingen, dan zijn meerdere taken en meerdere onafhankelijke beoordelaars noodzakelijk. Hoeveel taken en beoordelaars zijn er dan precies nodig? Hiervoor kunnen we gebruikmaken van generaliseerbaarheidsstudies. Dit type onderzoek maakt inzichtelijk hoe groot de verschillen zijn tussen schrijvers, beoordelaars en taken, zodat we hier in het examen rekening mee kunnen houden. Hoe groter bijvoorbeeld de verschillen in schrijfprestaties zijn van leerlingen tussen taken, hoe meer taken je moet afnemen om een goed oordeel over leerlingen te kunnen vellen. Uit dit onderzoek weten we dat voor betrouwbare oordelen ten minste 3 teksten per genre nodig zijn, die elk ten minste door 2 beoordelaars worden beoordeeld (Bouwer & van den Bergh, 2015). Met andere woorden, als docenten uitspraken willen doen over hoe goed hun leerlingen overtuigende teksten kunnen schrijven, dan moeten ze deze leerlingen 3 overtuigende teksten laten schrijven, en de kwaliteit van deze teksten beoordelen samen met een tweede onafhankelijke beoordelaar. Als docenten echter niet alleen uitspraken willen doen over hoe goed leerlingen kunnen schrijven in een specifiek genre, maar in hun uitspraken kunnen generaliseren naar de schrijfvaardigheid in het algemeen, dan zijn meerdere teksten van verschillende genres nodig.

De wetenschap biedt ook aanknopingspunten voor verbeteringen in de beoordelingen van tekstkwaliteit. Zo lijken vergelijkende beoordelingsmethoden over het algemeen tot betere oordelen te leiden dan analytische beoordelingsmodellen. Het grote verschil is dat de kwaliteit van de tekst als geheel wordt beoordeeld, in plaats van op losse criteria, en dat dit gebeurt op basis van een vergelijken. Dit vergelijken kan bijvoorbeeld met een of meer voorbeeldteksten van oplopende kwaliteit (ook wel beoordelingsschaal met ankerteksten genoemd), of in een reeks van paarsgewijze vergelijkingen. Het vergelijken van teksten gaat vaak redelijk snel (behalve als twee producten erg

veel op elkaar lijken, zie Van Daal, 2020), en sluit aan bij hoe we normaal gesproken beslissingen maken (Laming, 2004).

Het idee van vergelijkend beoordelen via paarsgewijze vergelijkingen (in het Engels: *Comparative Judgment*) vindt zijn oorsprong in het werk van Thurstone (1927), die liet zien dat we beter zijn in het vergelijken van twee objecten met elkaar dan in het geven van absolute scores aan losse objecten. Door een reeks van paarsgewijze vergelijkingen is het mogelijk om objecten op dezelfde schaal in te delen. Dit principe werd eerst vooral toegepast in de psychofysica, bijvoorbeeld voor het beoordelen van lichtintensiteit of het gewicht van objecten, of voor het meten van attitudes. In 2004 werd het door Pollitt voor het eerst gebruikt voor het beoordelen van schrijfproducten in het onderwijs. Het bleek dat paarsgewijze vergelijkingen zich ook goed lenen voor dit soort complexe beoordelingen. Het is voor beoordelaars bijvoorbeeld veel gemakkelijker om aan te geven welke van twee teksten beter is, dan ook nog een cijfer aan de teksten toe te moeten kennen. Algoritmes op de computer maken het beoordelingswerk helemaal makkelijk: teksten worden automatisch in paren aan beoordelaars toegewezen en op grond van hun keuzes worden teksten gerangschikt van slecht naar goed. Dit maakt het ook mogelijk om samen als groep te beoordelen; de uiteindelijke rangorde is dan een gedragen oordeel van de gehele sectie (Van Daal et al., 2019). Recent onderzoek laat zien dat de oordelen met paarsgewijs vergelijken betrouwbaar en valide zijn (Lesterhuis, 2018; Verhavert, 2018). Zo blijken docenten zich bij het vergelijken vooral te richten op de inhoud en structuur van de teksten. Ook minder ervaren docenten vergelijken teksten op basis van relevante tekstkenmerken en laten zich niet teveel afleiden door oppervlakkige aspecten, zoals lay-out of taalfouten. Op grond van de rangorde van teksten kunnen docenten bepalen waar het omslagpunt tussen een voldoende en onvoldoende zit. Hiermee kunnen automatisch de andere cijfers worden berekend.

Op basis van de rangorde uit paarsgewijs vergelijken is het ook mogelijk om ankerteksten te selecteren die representatief zijn voor de verschillende kwaliteitsniveaus in schrijven. Met deze ankerteksten kan een beoordelingsschaal worden ontwikkeld die docenten kunnen gebruiken voor het beoordelen van nieuwe teksten of voor het geven van feedback aan leerlingen. Het is dan wel belangrijk dat er voor elk van de ankerteksten wordt toegelicht wat de ene tekst beter maakt dan de andere. Onderzoek laat zien dat het ook mogelijk is om met een beoordelingsschaal teksten over andere onderwerpen, maar binnen hetzelfde genre, te beoordelen (Bouwer & Koster, 2016).

Aan de slag met een nieuw schoolexamen: vijf praktijkvoorbeelden
Om de wetenschappelijke inzichten te vertalen naar de onderwijspraktijk, zijn 20 scholen in het schooljaar 2019-2020 in zes werkgroepen aan de slag gegaan met het uitwerken van de volgende scenario's:

1. Meer schrijftaken opnemen in het schoolexamen;
2. Integratie van lees- en schrijftoetsen;
3. Beoordelingen uitwisselen met docenten van andere scholen;
4. Holistisch beoordelen met ankerschalen en paarsgewijs vergelijken;
5. Een beoordelingstraining voor een sectie;
6. Verkennen van een praktisch examen voor schrijfvaardigheid, waarbij net als bij de kunstvakken meer aandacht komt voor het onderliggende schrijfproces.

De werkgroepen maakten vooraf gezamenlijk een plan over welke veranderingen in het schoolexamen ze (op hun eigen school) wilden gaan uitproberen en waarom. Voor het opdoen van kennis en inspiratie kregen ze toegang tot een database met wetenschappelijke literatuur en waren er workshops, bijvoorbeeld over vergelijkend beoordelen met de online tool Comproved en over het praktisch examen in de kunstvakken. Bij elke stap in het proces, van de ontwikkeling van nieuw materiaal of instrumenten tot aan de implementatie en analyse van de resultaten, kregen de groepen feedback van wetenschappers en collega-docenten.

Het resultaat is een schat aan nieuwe praktijkvoorbeelden die op deze website door de werkgroepen zelf worden gepresenteerd. De bijdragen bevatten ook links naar nieuwe examenopdrachten en

beoordelingsinstrumenten die door andere scholen ook gebruikt kunnen worden. Elke werkgroep bespreekt ook welke lessen ze hebben geleerd en wat ze mee nemen naar hun volgende schoolexamen schrijfvaardigheid.

Hoe nu verder? Op naar een meer prominente rol voor schrijven in het examen

Met dit project en de praktijkvoorbeelden hopen we ook andere scholen inspiratie te geven om het schoolexamen te verbeteren. Op 10 februari 2021 volgt nog een inspiratiemiddag waarin de werkgroepen hun vernieuwde schoolexamen presenteren en ervaringen uitwisselen. Tijdens deze bijeenkomst is er ook de mogelijkheid om met je sectie je eigen PTA aan te pakken en aanpassingen aan collega-docenten en wetenschappers voor te leggen. De middag duurt van 15.00 tot 20.30 en is in Utrecht.

Met een beter schoolexamen schrijfvaardigheid alleen zijn we er echter nog niet. Draagvlakonderzoek onder leden van Levende Talen Nederlands wijst uit dat docenten een meer prominente rol voor schrijven willen in het examen. Op dit moment maakt schrijfvaardigheid maar een klein deel uit van het schoolexamen, terwijl iets meer dan de helft van de havo/vwo-bovenbouwddocenten schrijfvaardigheid wensen op te nemen in het centraal examen. Dit is ook de wens van experts op het gebied van schrijfvaardigheid, zie bijvoorbeeld de recente blogs van Helge Bonset, Huub van den Bergh en Gerdineke van Silfhout over het ideale examen Nederlands op Neerlandistiek.nl. Volgens hen krijgt leesvaardigheid nu een onevenredig groot aandeel in het cijfer voor Nederlands. Door schrijven meer gewicht te geven in het eindcijfer geeft dat een betere weerspiegeling van wat het vak Nederlands inhoudt. Met de praktijkvoorbeelden uit dit project hebben we een eerste verkenning gedaan voor de mogelijkheden daartoe, zoals van een praktisch examen voor schrijven zoals bij de kunstvakken, een tweede zitting met meerdere (korte) schrijfopdrachten of een gecombineerde zitting voor lees- en schrijfvaardigheid. De werkgroepen laten ook zien dat we de beoordelingsproblematiek kunnen we aanpakken door te werken met comparatieve beoordelingen en het uitwisselen van beoordelingen tussen scholen.

Referenties

Bouwer, R., & Bergh, V. D. H. (2015). Toetsen van schrijfvaardigheid: hoeveel beoordelaars, hoeveel taken? *Levende Talen Tijdschrift*, *16*(3), 3–12.

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

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability*. Princeton, New Jersey: Educational Testing Service.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. http://doi.org/10.1037/0022-0663.99.3.445

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237–263.

Laming, D. R. J. (2004). *Human Judgment: The Eye of the Beholder.* London: Thomson Learning.

Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality: An assessor's perspective* (ongepubliceerd proefschrift). Antwerpen: Universiteit Antwerpen.

Levende Talen Nederlands (2019). *Advies curriculum Nederlands in de bovenbouw van het voortgezet onderwijs.* Geraadpleegd via https://levendetalen.nl/wp-content/uploads/2019/09/LT-advies-curriculum-Nederlands-2019.pdf

McColly, W. (1970). What Does Educational Research Say About the Judging of Writing Ability? *Journal of Educational Research*, *64*(4), 147–156.

Meijerink, H. (2008). *Over de drempels met taal en rekenen* (pp. 1–56). Enschede: Expertgroep Doorlopende Leerlijnen Taal en Rekenen.

Pollitt, A. (2004). *Let's stop marking exams* (pp. 1–21). Philadelphia.

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

SLO (2012). *Handreiking voor het schoolexamen Nederlands havo/vwo*. Enschede: SLO.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Van Daal, T. (2020). *Making a choice is not easy?! Unravelling the task difficulty of comparative judgement to assess student work* (ongepubliceerd proefschrift). Antwerpen: Universiteit Antwerpen

Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education*, *26*(1), 59–74. http://doi.org/10.1080/0969594X.2016.1253542

Van den Bergh, H., De Maeyer, S., van Weijen, D., & Tillema, M. (2012). Generalizability of Text Quality Scores. In *Measuring Writing: Recent Insights into Theory, Methodology and Practice* (pp. 1–10). http://doi.org/10.1108/S1572-6304(2012)0000027005

Verhavert, S. (2018). *Beyond a mere rank order: The method, the reliability and the efficiency of comparative judgment* (ongepubliceerd proefschrift). Antwerpen: Universiteit Antwerpen.