Beoordeling van schrijfopdrachten door leerlingen;

Effecten van een korte beoordelaarstraining

Liselore van Ockenburg, Daphne van Weijen & Gert Rijlaarsdam

Doorgaans beoordelen docenten schrijfopdrachten van hun leerlingen zelf, maar wellicht kunnen leerlingen ook een rol spelen bij het (mee)beoordelen van de kwaliteit van elkaars teksten. Daartoe onderzochten we in hoeverre leerlingen door middel van een korte training kunnen leren de kwaliteit van syntheseteksten net zo te beoordelen als ervaren beoordelaars (experts) dat doen. De uitkomsten van dit onderzoek bieden steun voor het inschakelen van leerlingen als medebeoordelaars in een summatieve beoordelingssituatie. Voorafgaand aan de training kon er geen verschil worden aangetoond tussen experts en leerlingen voor de aspecten' samenhang tussen twee leerlingenoordelen' en 'samenhang tussen een leerlingoordeel en een expertoordeel'. De training had hierop geen invloed. Wel verschilden leerlingen en experts in strengheid. Na training was het leerlingoordeel nog steeds hoger, maar het verschil met de expertoordelen was kleiner geworden en verdween volledig bij het beoordelen van teksten met een ander onderwerp. Daarnaast bleek dat de betekenis van het holistisch leerlingoordeel, dat wil zeggen: de weging van verschillende aspecten van tekstkwaliteit in het holistisch eindoordeel, op alle meetmomenten niet (volledig) overeenstemde met het holistisch expertoordeel, maar na training wel meer overeenstemming ging vertonen met de expertweging. Omdat het oordeel van één beoordelaar bij summatieve beoordeling onvoldoende betrouwbaar is, lijkt het een reële optie om leerlingen, na een korte training, mee te laten beoordelen en daardoor schrijfproducten betrouwbaarder te evalueren.

In een eerder artikel in *Levende Talen Magazine* onderzochten we het gebruik van een online tool waarin leerlingen elkaars schrijfproducten eenvoudig en snel van commentaar kunnen voorzien en eventueel paarsgewijs kunnen beoordelen (https://comproved.com/; Van Ockenburg, 2019). Wanneer leerlingen commentaar geven op elkaars teksten levert dit niet alleen de docent een voordeel op, maar ook de leerlingen zelf. Er zijn zelfs aanwijzingen dat leerlingen meer leren van commentaar geven op elkaars teksten dan van ontvangen (Chanski & Ellis, 2017, p. 58). Ook SLO adviseert in de *Handreiking schoolexamens havo/vwo Nederlands* (Bonset et al., 2012, p. 48) om leerlingen elkaars conceptteksten te laten becommentariëren en zo meer te laten leren over de aspecten waarop tekstkwaliteit is gebaseerd.

Dit onderzoek wil meer licht werpen op de

vraag of een vervolgstap zou kunnen zijn leerlingen ook te betrekken bij de summatieve beoordeling van het eindproduct. Wij zien verschillende voordelen. Ten eerste ontlast dit docenten van hun rol als enige beoordelaar van schrijfproducten, wat wellicht mogelijkheden creëert om leerlingen meer teksten laten schrijven in een schooljaar zonder de werkdruk verder te verhogen. Daarnaast kunnen de uitkomsten van dit onderzoek meer licht werpen op de vraag hoe schrijfvaardigheid betrouwbaarder beoordeeld kan worden. Dat vraagstuk werd in Nederland geïntroduceerd door Wesdorp (1974, 1981), en later besproken door Schoonen (2012) en Van den Bergh e.a. (2012). Het lijkt erop dat schrijfvaardigheid alleen betrouwbaar beoordeeld kan worden als er verscheidene teksten per leerling beoordeeld worden, én als iedere beoordeling door verscheidene onafhankelijk van elkaar werkende beoordelaars wordt uitgevoerd. In de onderwijspraktijk lijkt de uitbreiding van het aantal beoordelaars per schrijfopdracht nauwelijks haalbaar: er zullen maar weinig vaksecties zijn die per rapportperiode alle teksten van alle leerlingen door drie docenten onafhankelijk laten beoordelen. Maar wat als we leerlingen kunnen inschakelen bij het beoordelen?

Er is ons in Nederland één experiment bekend waarin leerlingen betrokken werden bij summatieve beoordeling: het beoordelen van doel- en publiekgericht gedocumenteerd schrijven voor een schoolonderzoekcijfer. De gegevens uit dit experiment gebruiken we als referentiekader. In twee artikelen rapporteren Rijlaarsdam en Blok (Rijlaarsdam en Blok, 1981; Blok en Rijlaarsdam, 1981) een experiment waarin werkstukken van leerlingen voor een schoolonderzoek werden beoordeeld op basis van het oordeel van de docent en twee leerlingoordelen. De uitkomsten van het experiment lieten zien dat het oordeel van zo'n jury redelijk betrouwbaar is (betrouwbaarheid van 0,62 voor holistische tekstkwaliteit), en zeker betrouwbaarder dan de beoordeling van een enkele docent. Daarnaast analyseerden Blok en Rijlaarsdam ook de validiteit van de beoordelingen: verstaan de beoordelaars hetzelfde onder de te beoordelen aspecten? Een sluitend antwoord op deze vraag konden zij niet geven, maar zij merken op dat globale beoordelingen altijd weinig valide zijn, ongeacht of ze van een docent of een leerling komen, en dat de validiteit van analytische beoordelingen hoger is dan die van globale beoordelingen, mits het analytisch beoordelingsschema de onderwijsdoelen goed representeert.

Leerlingen als beoordelaars van teksten

Vanaf 2016 onderzoeken we hoe we leerlingen in het voortgezet onderwijs (3-vwo) syntheseteksten kunnen leren schrijven. Dat zijn teksten die de bronnen waarop ze berusten goed en geïntegreerd representeren. Een lezer die de bronnen niet heeft gelezen, moet de informatie goed kunnen begrijpen (Klein & Boscolo, 2016). We ontwierpen een lessenreeks over syntheseteksten en testten die op drie vo-scholen in Nederland (Van Ockenburg e.a., 2021a). Omdat leerlingen in de lessenserie gevraagd werd om (delen van) elkaars syntheseteksten formatief te beoordelen, waren we nieuwsgierig of leerlingen dit voldoende betrouwbaar kunnen. We onderzochten daartoe in deze studie in hoeverre leerlingen in staat zijn om elkaars teksten adequaat te beoordelen met behulp van het beoordelingsinstrument dat we maakten voor experts, en of een korte training de beoordelingsvaardigheid van leerlingen zou kunnen verbeteren. Als blijkt dat leerlingen in staat zijn teksten van een beoordeling te voorzien die overeenstemt met de beoordeling van experts, dan kunnen leerlingbeoordelingen ook in summatieve beoordelingssituaties

worden benut om de betrouwbaarheid van de beoordelingen te verhogen.

Onderzoeksvraag

De studie die we hier rapporteren spitst zich toe op de vraag:

> *In hoeverre beïnvloedt een korte training de overeenstemming tussen leerling- en expertbeoordelingen?*

We onderzoeken vier aspecten van overeenstemming:

a. Interne consistentie: de mate waarin de kwaliteitsrangordening van teksten tussen beoordelaars onderling overeenstemt.
b. Strengheid: de mate waarin de waarde die wordt toegekend aan teksten overeenstemt.
c. Predictieve validiteit: de mate waarin een leerlingoordeel overeenstemt met een expertoordeel.
d. Constructvaliditeit: de mate waarin het holistische oordeel van leerlingen en experts op dezelfde weging van beoordelingsaspecten is gebaseerd.

Methode

Deelnemers

Aan deze studie namen 75 leerlingen deel van één school, namelijk de school waaraan de docent-onderzoeker is verbonden. Zij kwamen uit tien verschillende 3- en 4-vwo-klassen. Dit vergroot de generalisatiewaarde van de uitkomsten, omdat leerlingen binnen een klas meer op elkaar lijken dan leerlingen uit verschillende klassen. Zij hadden allen geen ervaring met het lezen, schrijven of beoordelen van syntheseteksten. In tabel 1 is meer informatie over de leerlingen en hun verdeling over de condities te vinden. Na een oproep van de docent-onderzoeker via Magister en in de klas meldden zij zich vrijwillig aan voor een voorlichtingsbijeenkomst buiten schooltijd. Bij deelname ontvingen zij een kleine financiële vergoeding in de vorm van een waardebon, mits zij alle onderzoeksactiviteiten hadden voltooid. Alle deelnemers en hun ouders verleenden actief toestemming voor deelname.

Onderzoeksontwerp

Het onderzoeksontwerp was gebaseerd op het *Solomon four group design*: een uitbreiding van het klassieke onderzoeksontwerp met de kenmerken (1) voormeting-nameting, (2) controlegroep en (3) aselecte toewijzing aan condities, waarbij de experimentele groep de training ontvangt en de controlegroep niet. In het *Solomon four group design* voegt de onderzoeker twee condities toe: een conditie die niet deelneemt aan de voormeting maar wel aan de training, en een conditie die uitsluitend deelneemt aan de nameting. Zo kan een eventueel effect van deelname aan de voormeting op de effectiviteit van de training worden bepaald. Die laatste conditie, alleen deelname aan de nameting, hebben wij niet gerealiseerd omdat het aantal beschikbare deelnemers beperkt was. Minder deelnemers per conditie zou ten koste gaan van de betrouwbaarheid van de analyses.

Met de toevoeging van conditie 3 wilden we vaststellen of er van de voormeting eventueel een leereffect uitging. Wanneer we bij

Conditie	Geslacht		leeftijd m (sd)
	jongen	meisje	
1	7	18	14,28 (0,5)
2	13	12	14,08 (0,4)
3	8	17	14,08 (0,6)
Totaal	28	47	14,15 (0,6)

Tabel 1. Kenmerken deelnemers per conditie

conditie 1 een positief effect van training zouden vaststellen, dan was dat effect verkregen na een voormeting. Een leereffect van de voormeting konden we niet bij voorbaat uitsluiten: het was mogelijk dat leerlingen die beoordelaarservaring hadden opgedaan, door deze ervaring meer baat hadden bij de training en beter beoordeelden bij de nameting dan leerlingen die de training volgden zonder deze ervaring. Door conditie 2 en 3 te vergelijken, konden we het effect van deelname aan de voormeting op de nameting isoleren van het trainingseffect. Daarnaast wilden we, door toevoeging van een transfermeting, vaststellen of het effect van de training standhield bij de beoordeling van teksten over een ander, ongetraind onderwerp (onderwerp B). Tabel 2 toont het ontwerp dat dus drie meetmomenten en drie condities omvatte: Conditie 1: wel voormeting, wel training; Conditie 2: wel voormeting, geen training; Conditie 3: geen voormeting, wel training.

Procedure

Voorafgaand aan hun deelname aan het onderzoek, kregen alle geïnteresseerde leerlingen een korte voorlichting tijdens een fysieke bijeenkomst op school in kleine groepen. Hierbij nam de eerste auteur, docent Nederlands op de betreffende school, de onderzoeksprocedure met hen door, dus welke taken zij zouden moeten uitvoeren binnen welk tijdsbestek. Na deze bijeenkomst konden leerlingen definitief besluiten of zij wilden deelnemen.

Vervolgens vonden vier onderzoeksactiviteiten plaats in vier aaneengesloten weken: drie beoordelingssessies en één trainingssessie. Iedere activiteit duurde ongeveer 60 minuten en werd flexibel ingepland. De drie beoordelingssessies (voormeting, nameting en transfermeting) werden door de leerling-beoordelaars thuis uitgevoerd.

Het materiaal dat we gebruikten voor de beoordelingssessies kwam uit twee tekstsets uit eerder onderzoek (Van Ockenburg e.a., 2021a). Iedere set bestond uit 124 syntheseteksten van leerlingen uit tien 3-vwo-klassen van drie verschillende middelbare scholen in Nederland. Deze teksten moesten ongeveer 200 woorden lang zijn en gebaseerd zijn op drie korte bronnen die elkaar aanvulden (aantal woorden per bron $M = 188{,}9$, $SD = 55$). De eerste set (onderwerp A) bestond uit teksten over bedreigde wilde dieren in Afrika en de tweede (onderwerp B) uit teksten over kunstmatige kleurstoffen in voedsel (E-nummers). De opdrachten waren gebaseerd op synthesetaken die waren ontwikkeld en getest als onderdeel van een nationaal peilingsonderzoek naar synthesetaken in de bovenbouw van het voortgezet onderwijs (Vandermeulen e.a., 2020). Iedere tekst was onafhankelijk beoordeeld door een groep van drie beoordelaars uit een panel van 25 experts (docenten Nederlands en schrijfvaardigheidsonderzoe-

Conditie	*n*	Voormeting (onderwerp A)	Training	Nameting (onderwerp A)	Transfermeting (onderwerp B)
1	25	o	x	o	o
2	25	o		o	o
3	25		x	o	o

x = deelname aan training; o = tekstbeoordeling

Tabel 2. Onderzoeksontwerp

kers). Meer informatie over het beoordelingsinstrument dat zij gebruikten is te vinden in Van Ockenburg e.a. (2021b).

Aan het begin van de week ontvingen de leerlingen 15 teksten over onderwerp A in papieren vorm. Zodra zij klaar waren met beoordelen konden zij hun beoordelingen digitaal inleveren. De leerlingen uit conditie 3, de conditie zonder voormeting, ontvingen voor de voormeting ook een pakketje, maar dat bevatte een taak die niets met beoordelen te maken had. Het waren literaire teksten met vragen die zij moesten beantwoorden. De taak vroeg een vergelijkbare tijdsinvestering als de taak die de andere leerlingen moesten uitvoeren, maar was inhoudelijk niet gerelateerd aan het onderzoek. Na de voormeting volgden nog twee beoordelingsrondes: bij de nameting beoordeelden de leerlingen ieder een ander pakket teksten over onderwerp A uit het totale bestand waaruit we putten voor de voormeting; bij de transfermeting een pakket syntheseteksten over een ander onderwerp.

De trainingssessies voor experimentele condities 1 en 3 vonden plaats op school. De training werd op zes roostermomenten aangeboden en leerlingen schreven zich in voor een moment dat goed in hun rooster paste. Twee leerlingen die vanwege Corona in quarantaine zaten in de betreffende week, volgden de training digitaal via Teams. De leerlingen uit controleconditie 2 (wel voormeting, geen training) schreven zich ook in, maar zij beantwoordden vragen bij literaire teksten in een andere ruimte.

Inhoud beoordelingstraining

De inhoud van de beoordelingstraining baseerden we op een beoordelingstraining voor docenten (Echten et al., 2020) die we ontwikkelden in het kader van het project Schrijven in het Schoolexamen, in samenwerking met de WODN en SLO (https://didactieknederlands.nl/publicaties/wie-schrijft-die-blijft-afstemmen-methode-voor-een-betere-beoordelaarsafstemming-in-een-lesuur/). Het doel van de oorspronkelijke training was docenten in een sectie met elkaar in gesprek te brengen over hun visie op tekstkwaliteit door teksten van verschillende kwaliteit te vergelijken en te ordenen, en zo een (meer) gedeelde visie op tekstkwaliteit te creëren. Dit zou betrouwbaarder oordelen over tekstkwaliteit opleveren.

De betrouwbaarheid en validiteit van de beoordelingen kan vergroot worden door het gebruik van beoordelingscriteria en schalen met ankerteksten. Dat zijn de twee pijlers van de training. Ten eerste wilden we beoordelaars trainen om criteria te hanteren die handvatten bieden om ieder schrijfproduct op dezelfde manier te beoordelen. Een analytisch beoordelingsschema, bijvoorbeeld, legt criteria vast die beoordelaars meenemen in de beoordeling (Coertjens et al., 2017). Ten tweede kan ook een schaal met ankerteksten helpen bij de beoordeling (Koster et al., 2018; Wesdorp, 1981). Zo'n schaal bestaat uit een reeks in kwaliteit oplopende voorbeeldteksten die zijn voorzien van een onderbouwd oordeel. Een tekstschaal kan bepaalde beoordelingsproblemen voorkomen, zoals het sequentie-effect (wat inhoudt dat een goede tekst die volgt na een aantal zwakkere teksten hoger wordt beoordeeld dan wanneer hij zou volgen na een andere goede tekst), normverschuiving (wat inhoudt dat een beoordelaar zich (onbewust) aanpast aan het niveau van de teksten) of een signifisch effect (verschillende beoordelaars hechten waarde aan andere tekstaspecten en daardoor verschilt hun beoordeling van dezelfde tekst) (Koster et al. 2018; Pollmann et al., 2012; Wesdorp, 1981).

Het doel van de interventie was leerlingoordelen meer laten overeenstemmen met expertoordelen. Daarom pasten we de oorspronkelijke training aan opdat die meer richtinggevend zou zijn. Tabel 3 geeft een overzicht van de inhoud van de leerlingtraining. In verschillende beoordelingsrondes vergelijken de leerlingen hun eigen beoordelingen van voorbeeldteksten met die van experts en worden deze expertoordelen toegelicht. Aan het eind van de training zou duidelijk moeten zijn hoe experts tekstkwaliteit beoordelen en waarom zij dat zo doen, zodat leerlingen hun eigen beoordelingen hierop af kunnen stemmen.

Resultaten

De vraag die we hier beantwoorden, is in hoeverre een korte training de overeenstemming tussen expert- en leerlingbeoordelingen beïnvloedt wat betreft interne consistentie, strengheid, predictieve validiteit en constructvaliditeit. Ook wilden we weten of de overeenstemming verandert wanneer

Ronde	Min.	Hoe?*	Inhoud
1	10	I	Leerlingen leggen individueel drie informatieve voorbeeldteksten op volgorde van kwaliteit en geven argumenten voor deze rangschikking.
2	10	G	Leerlingen vergelijken de eigen rangschikking met de rangschikking door experts en bespreken hoe de aspecten informatie, integratie, structuur en stijl meewegen in het beoordelen van tekstkwaliteit.
3	5	I	Leerlingen voegen individueel twee nieuwe voorbeeldteksten toe aan de expertrangschikking uit ronde 2.
4	5	G	Leerlingen vergelijken de eigen en de expertrangschikking, bediscussiëren eventuele verschillen en bepalen of de kwaliteitsaspecten uit ronde 2 helder genoeg zijn of nader moeten worden toegelicht.
5	10	I	Leerlingen bepalen voor de vijf voorbeeldteksten uit ronde 1-4 voor elk aspect de score op een schaal van 1 t/m 5 (analytisch) en de holistische score (1-100).
6	10	G	Leerlingen vergelijken hun eigen analytische en holistische oordelen met de expertoordelen en bediscussiëren eventuele verschillen.

* I = Individueel; G = Groepsgesprek geleid door trainer

Tabel 3. Structuur en globale inhoud beoordelingstraining voor leerlingen

			MEETMOMENT		
			Voormeting	Nameting	Transfermeting
ONDERWERP			A	A	B
Conditie	Voormeting	Training			
Experts			0,33	0,33	0,33
Leerlingen conditie 1	ja	ja	0,33	0,32	0,23
Leerlingen conditie 2	ja	nee	0,26	0,35	0,27
Leerlingen conditie 3	nee	ja	–	0,37	0,30

Tabel 4. Interne consistentie: Generalisatiecoëfficiënt van de correlatie tussen twee beoordelaars uit één conditie voor holistische tekstkwaliteit

			MEETMOMENT		
			Voormeting	Nameting	Transfermeting
ONDERWERP			A	A	B
Conditie	Voormeting	Training			
Experts (baseline)			52,9 (13,2)	52,9 (13,2)	66,4 (10,9)
Leerlingen conditie 1	ja	ja	60,3 (12,4)	57,8 (11,7)	67,9 (8,9)
Leerlingen conditie 2	ja	nee	60,5 (12,7)	63,1 (12,0)	69,8 (9,9)
Leerlingen conditie 3	nee	ja	-	57,5 (13,4)	66,8 (10,5)

Tabel 5. Strengheid: Gemiddelde (en standaardafwijking) van holistische oordelen (schaal 0-100)

synthese-teksten over een ander onderwerp beoordeeld worden.

a. Interne consistentie

Tabel 4 toont de betrouwbaarheid van de holistische beoordelingen binnen de groepen. De correlaties tussen twee beoordelaars uit één conditie varieerde van 0,26 (leerlingen conditie 2, voormeting) tot 0,37 (leerlingen conditie 3, nameting). Er waren op geen enkel meetmoment significante verschillen in betrouwbaarheid (test statistiek $z = 0{,}41$, $p = 0{,}34$; Lenhard & Lenhard, 2014): de correlatie tussen twee experts onderling en twee leerlingen onderling verschilde niet. Zowel training als onderwerp hadden geen effect op de interne consistentie van de beoordelingen: de correlatie tussen twee leerlingen nam niet toe, en bleef gelijk aan die van experts.

b. Strengheid

Tabel 5 toont de gemiddelde beoordeling van tekstkwaliteit per conditie per meetmoment. We voerden een multilevelanalyse uit met conditie als factor en de meetmomenten genest in deelnemers. Op de voormeting vergeleken we drie condities: experts en twee leerlingcondities. De derde leerlingconditie nam bewust niet deel aan de voormeting vanwege het onderzoeksontwerp. De analyse liet zien dat er op de voormeting een significant verschil bestond tussen de hoogte van de holistische beoordelingen door experts en leerlingen ($F(2{,}991{,}083) = 30{,}72$, $p < 0{,}001$). Paarsgewijze vergelijkingen lieten zien dat leerlingen significant hoger oordeelden dan experts (beide vergelijkingen $p < 0{,}001$; $d = 0{,}58$, wat wijst op een middelgroot effect van beoordelaarsconditie). De twee leerlingcondities verschilden onderling niet ($p = 0{,}89$).

Een multilevelanalyse – de data waren immers genest in beoordelaars – van de holistische beoordelingen (tabel 5) toonde aan dat ook bij de nameting de leerlingen nog minder streng oordeelden dan de experts ($F(3{,}1363{,}028) = 32{,}49$, $p < 0{,}001$). Het verschil tussen het oordeel van experts en getrainde leerlingen uit conditie 1 bleek echter significant kleiner ten opzichte van de voormeting ($B = 5{,}17$, $SE = 2{,}25$, $p = 0{,}02$). Voor conditie 3 konden we dit niet vaststellen aangezien deze conditie niet deelnam aan de voormeting. Bij de nameting verschilden de scores van conditie 1 ($m = 57{,}8$) en conditie 3 ($m = 57{,}5$) niet ($p = 0{,}93$). Conditie 2, zonder training, beoordeelde op de nameting significant hoger dan de condities met training ($B = 5{,}61$, $SE = 1{,}59$, $p < 0{,}001$). Na training stemde de hoogte van het holistisch leerlingoordeel dus meer overeen met het expertoordeel. Met andere woorden, de leerlingen waren wat strenger gaan beoordelen.

Bij de transfermeting was er geen verschil meer tussen de oordelen van de twee trainingscondities onderling (conditie 1 en 3; $B = 1{,}16$, $SE = 1{,}28$, $p = 0{,}36$), noch tussen de expertoordelen en trainingscondities (conditie 1: + 1,4 punten ten opzichte van het holistisch expertoordeel, $SE = 1{,}3$, $p = 0{,}25$; conditie 3: + 0,32 punten ten opzichte van het holistisch expertoordeel, $SE = 1{,}3$, $p = 0{,}80$). De conditie zonder training beoordeelde ook op de transfermeting nog altijd significant minder streng dan de experts (+ 3,3 punten, SE: 1,3, $p = 0{,}009$).

Het effect van de training op de overeenstemming tussen leerling- en expertbeoordelingen lijkt dus door te werken in de beoordeling van teksten met een ander onderwerp.

c. Predictieve validiteit

Hoewel leerlingen onderling het net zo met elkaar eens zijn als experts onderling als zij teksten moeten rangordenen op kwaliteit (tabel 4), is het mogelijk dat de rangordes van leerlingen en experts verschillen. De tussengroepcorrelatie kan uitsluitsel bieden. Tabel 6 geeft de correlatie tussen leerlingen en experts, en biedt inzicht in de mate waarin de twee groepen overeenkomen in hun oordeel

			MEETMOMENT		
			Voormeting	Nameting	Transfermeting
ONDERWERP			A	A	B
Conditie	Voormeting	Training			
Leerlingen conditie 1	ja	ja	0,29	0,30	0,26
Leerlingen conditie 2	ja	nee	0,26	0,28	0,24
Leerlingen conditie 3	nee	ja	-	0,28	0,27

Tabel 6. Predictieve validiteit: Generalisatiecoëfficiënt tussen een willekeurige expert en een willekeurige leerling

		EXPERTS	CONDITIE 1	CONDITIE 2	CONDITIE 3[1]
VOORMETING	Informatie	0,77	0,62	0,67	
	Integratie	0,75	0,60	0,62	
	Structuur	0,74	0,63	0,65	
	Stijl	0,61	0,63	0,62	
NAMETING	Informatie	0,77	0,64	0,72	0,71
	Integratie	0,75	0,67	0,60	0,66
	Structuur	0,74	0,61	0,63	0,70
	Stijl	0,61	0,58	0,63	0,63
TRANSFERMETING	Informatie	0,69	0,66	0,63	0,72
	Integratie	0,77	0,66	0,67	0,68
	Structuur	0,80	0,69	0,68	0,68
	Stijl	0,70	0,58	0,63	0,61

Conditie 1: Voormeting en Training;
Conditie 2: Voormeting;
Conditie 3: Training
1. Leerlingen in conditie 3 namen niet deel aan de voormeting.

Tabel 7. Constructvaliditeit: Correlaties van aspectscores met holistische tekstkwaliteit

over welke teksten sterker en welke zwakker zijn, zonder te kijken naar de normering.

De correlatie tussen de rangordening van de teksten door de twee groepen leerlingen die deelnamen aan de voormeting (tabel 6: voormeting 0,26 vs 0,29, test statistiek $z = 0{,}11$, $p = 0{,}46$) verschilde niet significant, en ook niet van de correlatie tussen twee experts (tabel 4 0,33 vs tabel 6 0,26, test statistiek $z = 0{,}25$, $p = 0{,}40$): leerlingenoordelen over de rangschikking van tekstkwaliteit hingen dus evenzeer met expertoordelen samen als expertoordelen onderling. Zowel training als onderwerp beïnvloedden die samenhang niet: de samenhang bleef even sterk als die tussen leerlingen onderling en tussen experts onderling.

d. Constructvaliditeit

Tabel 7 toont verschillen in de betekenis die experts en leerlingen toekennen aan holistische oordelen. De verschillen tussen de twee leerlingcondities waren niet significant bij de voormeting: alle vier de aspecten wogen in dezelfde mate mee (Conditie 1: $r = 0{,}60$ vs $r = 0{,}63$, $p = 0{,}26$; conditie 2: $r = 0{,}62$ vs $r = 0{,}67$, $p = 0{,}12$). Maar de verschillen tussen leerlingen en experts waren tweeërlei. Ten eerste wogen de aspectscores voor informatie, integratie en structuur bij experts zwaarder mee dan het aspectoordeel over stijl ($r = 0{,}61$ vs $r = 0{,}74$, $p = 0{,}001$). Leerlingen daarentegen betrokken alle vier de aspecten gelijkmatig in hun oordeel, maar hun holistisch oordeel correleerde zwakker met de aspectscores informatie, integratie en structuur dan bij de experts het geval was (het kleinste verschil $r = 0{,}65$ vs $r = 0{,}74$ is significant: $p = 0{,}009$). De correlaties voor die drie aspectscores met het globale oordeel waren voorafgaand aan de training van de leerlingen dus significant sterker bij experts.

Bij de nameting was er een flink verschil waar te nemen tussen de twee trainingscondities (conditie 1 en 3) in de bijdrage van aspectscores aan het holistische oordeel, wanneer we de bijdrage in het expertpanel als criterium namen. Voor de aspecten informatie, integratie en structuur is de bijdrage in conditie 1 significant kleiner dan het criterium (het kleinste verschil, $r = 0{,}75$ vs $r = 0{,}67$ is statistisch significant: $p = 0{,}014$). Voor conditie 3 geldt dit alleen voor het aspect integratie ($p = 0{,}007$). Kennelijk is de voormeting, waaraan conditie 1 wel deelnam en conditie 3 niet, van invloed op de mate waarin de training in dit opzicht een effect heeft. Conditie 2, die alleen meedeed met de voormeting maar geen training kreeg, scoort lager op de aspecten integratie en structuur dan het criterium (het verschil voor informatie is niet significant, $p = 0{,}06$). Voor het aspect stijl geldt dat de bijdrage aan het holistische oordeel in alle drie de condities niet verschilt van het criterium.

Bij de transfermeting was de bijdrage van de aspectscores integratie en structuur aan de holistische score in de expertoordelen steevast groter dan in de leerlingoordelen: het kleinste verschil voor integratie ($r = 0{,}68$ en $r = 0{,}77$, $p = 0{,}005$) was significant, evenals het kleinste verschil voor structuur ($r = 0{,}80$ en $r = 0{,}69$, $p < 0{,}001$). Voor het aspect informatie verschilden de bijdragen niet en voor stijl week alleen conditie 3 af met een kleinere bijdrage vergeleken met experts ($p = 0{,}02$). Hiermee leek hun oordeel meer dan het oordeel van de andere condities op dat van experts, die stijl ook minder zwaar lieten meewegen in hun holistisch oordeel.

Samenvatting

Op alle meetmomenten waren de oordelen van leerlingen onderling even consistent als de oordelen van experts onderling, en hingen leerlingoordelen over de kwaliteitsrangschikking van teksten samen met expertoordelen. Op deze twee aspecten had een beoordelaarstraining geen invloed.

Leerlingen verschilden eerst wel van experts in strengheid: bij de voormeting waardeerden ze teksten gemiddeld hoger. Training leidde ertoe dat leerlingen strenger oordeelden dan voor de training, maar nog niet even streng als experts. Conditie 2, zonder training, beoordeelde de teksten significant hoger dan condities 1 en 3, met training. Op de transfermeting kon er op dit aspect geen verschil meer worden aangetoond tussen conditie 1 en 3 enerzijds en de experts anderzijds. De beoordeling van conditie 2, zonder training, was nog altijd significant hoger. Training doet er dus toe in dit opzicht.

Voorafgaand aan de training was het holistische leerlingoordeel minder sterk verbonden met de aspectscores *informatie*, *integratie* en *structuur*, dan het expertoordeel. Bij de nameting zagen we effecten van training bij condities 1 en 3, die beide gedifferentieerder de vier aspecten meewogen in het holistische oordeel. In het bijzonder de bijdrage van *stijl* werd kleiner ten opzichte van andere aspecten, zoals experts ook *stijl* iets minder meewogen. Desondanks stemde ook na training de betekenis van het holistisch leerlingoordeel niet volledig overeen met het holistisch expertoordeel.

Discussie en implicaties voor de praktijk

In deze studie onderzochten we of een beoordelingstraining de beoordelingsvaardigheid van leerlingen beïnvloedt. We vergeleken de leerlingoordelen steeds op vier aspecten met een criterium dat werd gevormd door expertbeoordelingen van dezelfde teksten: a) interne consistentie, b) strengheid, c) predictieve validiteit en d) constructvaliditeit.

De resultaten lieten een trainingseffect zien: zowel conditie 1 als 3 stemmen sterker overeen met experts na de training. Dit is goed nieuws voor de implementatie van de beoordelaarstraining in het onderwijs, omdat een vooroefening waarin leerlingen vertrouwd raken met het te beoordelen tekstgenre niet nodig lijkt voor effectieve training. Er kon namelijk geen verschil worden aangetoond tussen conditie 3, zonder voormeting, en conditie 1, met voormeting.

Bij de transfermeting bleken getrainde leerlingen net zo streng te beoordelen als experts. Toch vonden we eerder, bij de nameting, nog wel een verschil in strengheid tussen leerlingen en experts, ondanks de training. We moeten bij de interpretatie van deze gegevens echter in de beschouwing betrekken dat het hier gaat om betrekkelijk willekeurige leerlingen, die evenwel vrijwillig opteerden voor deelname aan dit onderzoek en geconfronteerd werden met een tekstgenre waarmee zij *niet* bekend waren. Wij vergeleken hun beoordelingen met de beoordelingen van een groep die bestond uit willekeurige experts. In een onderwijssituatie leren leerlingen de beoordelingswijze van hun docent kennen, en mag verondersteld worden dat er binnen een klas meer duidelijkheid bestaat over wat goede en minder goede teksten zijn dan in deze studie het geval was.

De enige, maar niet onbelangrijke dimensie waarop op geen enkel meetmoment volledige overeenstemming werd bereikt tussen leerlingen en experts, was constructvaliditeit. Maar ook voor deze uitkomst geldt dat in een reguliere onderwijssituatie leerlingen waarschijnlijk beter kunnen inschatten wat hun docent belangrijke kwaliteitsaspecten vindt dan de leerlingen in deze studie. In het ideale geval stellen docenten in samenspraak met hun leerlingen heldere criteria voor tekstkwaliteit vast, waarna de leerlingen oefenen met het toepassen van deze criteria door het geven en ontvangen van feedback op conceptversies van teksten. Bij summatief beoordelen zijn leerlingen dan beter in staat deze criteria toe te passen in overeenstemming met de werkwijze van hun eigen docent.

Een bekend probleem bij het beoordelen van teksten is de lage overeenstemming over tekstkwaliteit. Ook bij de teksten die we selecteerden voor deze studie was de gemiddelde overeenstemming onder twee willekeurige expertbeoordelaars slechts 0,33 (zie tabel 4). Gezien deze lage interne consistentie, is het dus een goed idee om een tekst te laten beoordelen door meer onafhankelijke beoordelaars om de betrouwbaarheid te vergroten. In de praktijk is het lastig om iedere leerlingtekst door minimaal drie docenten onafhankelijk te laten beoordelen. De uitkomsten van dit onderzoek bieden steun voor het inschakelen van leerlingen als medebeoordelaars, na een korte training, in een summatieve beoordelingssituatie. Leerlingen hoeven voorafgaand aan de training geen ervaring te hebben met het schrijven of beoordelen van de betreffende teksten, en het effect werkt door in hun beoordeling van teksten met een ander onderwerp. Door leerlingen op deze eenvoudige manier te betrekken bij het beoordelingsproces kunnen schrijfproducten betrouwbaarder geëvalueerd worden. Een beoordeling door één docent is niet betrouwbaar, een beoordeling door drie docenten is niet haalbaar, een beoordeling door drie leerlingen of een docent en drie leerlingen is betrouwbaarder.

Literatuur

Bergh, H., van den, De Maeyer, S., Weijen, D. van, & Tillema, M. (2012). Generalizability of text quality scores. *Measuring writing: Recent insights into theory, methodology and practices*, *27*, 23–32.

Bergh, H. van den, & Eiting, M. H. (1989). A method of estimating rater reliability. *Journal of Educational Measurement*, *26*(1), 29–40.

Blok, H., & Rijlaarsdam, G. (1981). Beoordeling van schrijfprodukten door leerlingen: evaluatie van de kwaliteit. *Levende Talen*, *367*, 958–972.

Bonset, H., Meestringa, T., & Ravesloot, C. (2012). *Handreiking schoolexamens Nederlands Havo/vwo*.SLO.

Chanski, S., & Ellis, L. (2017). Which helps writers more, receiving peer feedback or giving it? *English Journal 106* (6), 54–60.

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.

Echten, K., Linnert, A., & Ockenburg, L. van (2020). *Wie schrijft die blijft… Afstemmen. Methode voor een betere beoordelaarsafstemming in één lesuur*. Projectpublicatie 5 van het LTN-project Schrijfvaardigheid in de schoolexamens. Levende Talen Nederlands. https://didactieknederlands.nl/publicaties/wie-schrijft-die-blijft-afstemmen-methode-voor-een-betere-beoordelaarsafstemming-in-een-lesuur/

Klein, P. D., & Boscolo, P. (2016). Trends in research on writing as a learning activity. *Journal of Writing Research*, *7*(3), 311–350.

Koster, M., Besselink, E., & Seinhorst, E. (2018). "Goed gedaan, maar kan nog beter…"; Het ontwikkelingsgericht en betrouwbaar beoordelen van leerlingteksten. *Tijdschrift voor Lerarenopleiders*, *39*(3), 83–94.

Lenhard, W., & Lenhard, A. (2014). *Hypothesis Tests for Comparing Correlations*. available: https://www.psychometrica.de/correlation.html. Bibergau (Germany): Psychometrica. DOI:10.13140/RG.2.1.2954.1367

Ockenburg, L. van. (2019). Comparatief beoordelen in D-PAC. Leerlingen vergelijken teksten en geven feedback in handige online tool. *Levende Talen Magazine*, *106*(1).

Ockenburg, L. van, Weijen, D. van, & Rijlaarsdam, G. (2021a). Choosing how to plan informative synthesis texts: Effects of strategy-based interventions on overall text quality. *Reading and Writing*. https://doi.org/10.1007/s11145-021-10226-6

Ockenburg, L. van, Weijen, D. van, & Rijlaarsdam, G. (2021b). Learning how to synthesize: The design and evaluation of a reading-writing learning unit for high-school students. *L1-Educational Studies in Language and Literature*, Volume 21, 1–33. https://doi.org/10.17239/LIESLL-2021.21.01.06

Pollmann, E., Prenger, J., & Glopper, K. de (2012). Het beoordelen van leerlingteksten met behulp van een schaalmodel. *Levende Talen Tijdschrift*, *13*(3), 15–24.

Rijlaarsdam, G., & Blok, H. (1981). Beoordeling van schrijfprodukten door leerlingen: theorie en praktijk. *Levende Talen*, *365*, 753–766.

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In *Measuring writing: Recent insights into theory, methodology and practice* (pp. 1–22). Studies in Writing vol. 27. Koninklijke Brill NV.

Vandermeulen, N., De Maeyer, S., Van Steendam, E., Lesterhuis, M., Bergh, H. van den, & Rijlaarsdam, G. (2020). Mapping synthesis writing in various levels of Dutch upper-secondary education A national baseline study on text quality, writing process and students' perspectives on writing. *Pedagogische Studiën*, *97*(3), 187–236.

Wesdorp, H. (1974). Het meten van de produktief-schriftelijke taalvaardigheid. Directe en indirecte methoden: 'opstelbeoordeling' versus 'schrijfvaardigheidstoetsen'. Universiteit van Amsterdam (dissertatie).

Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs*. Staatsuitgeverij.

Liselore van Ockenburg is docent Nederlands aan het Stedelijk Gymnasium Den Bosch. Dankzij een promotiebeurs voor leraren van het NWO doet zij sinds 2016 onderzoek aan de Universiteit van Amsterdam naar het leren schrijven van syntheseteksten in het voortgezet onderwijs.
E-mail: L.vanOckenburg@uva.nl

Daphne van Weijen is universitair docent en onderzoeker aan de Interfacultaire Lerarenopleidingen (ILO) van de Universiteit van Amsterdam. Zij voert samen met anderen domeinspecifiek onderzoek uit naar syntheseteksten schrijven en schrijfonderwijs in moedertaal en moderne vreemde talen, en begeleidt praktijkonderzoek van docenten binnen de Werkplaats Onderwijsonderzoek Amsterdam-VO/MBO.
E-mail: D.vanWeijen@uva.nl

Gert Rijlaarsdam was jarenlang leraar Nederlands in Dordrecht en is hoogleraar innovatie van het taalonderwijs aan de Universiteit van Amsterdam.
E-mail: G.C.W.Rijlaarsdam@uva.nl

Naar betekenisvol formuleeronderwijs.

Een kwestie van afstemming

Jeroen Steenbakkers, Jimmy van Rijt, Veerle Baaijen,
Ninke Stukker & Kees de Glopper

Leerlingen krijgen doorgaans vanaf klas 3 of 4 havo/vwo jaarlijks modules correct formuleren aangeboden. Ze leren daarbij formuleerfouten (zoals foutieve samentrekkingen en contaminaties) herkennen en verbeteren. In hoeverre sluiten deze modules aan bij de formuleerproblemen van de leerlingen? Tot voor kort kon deze vraag niet goed worden beantwoord, omdat er onvoldoende empirische gegevens waren over formuleerfouten in leerlingenteksten. Twee recente studies hebben nieuwe data aangeleverd. Het eerste deel van dit artikel bevat een overzicht van wat er inmiddels bekend is over formuleerfouten in leerlingenteksten. Het tweede deel van dit artikel presenteert ontwerpprincipes voor nieuw formuleeronderwijs. Nieuwe modules (correct) formuleren zouden zich niet moeten richten op vermeende formuleerproblemen maar op reële formuleerproblemen van leerlingen. Bovendien kunnen ze bijdragen aan de ontwikkeling van bewuste taalvaardigheid.

Er bestaat voor het voortgezet onderwijs geen periodiek peilingsonderzoek met betrekking tot taalverzorging, en meer in het bijzonder formuleervaardigheid. We kunnen daardoor nu niet bepalen of het niveau van formuleren van leerlingen de afgelopen decennia is gedaald, gelijk is gebleven of is toegenomen. We weten ook onvoldoende hoe de formuleervaardigheid van leerlingen zich gedurende hun middelbareschooltijd ontwikkelt. De afgelopen decennia is er wel incidenteel onderzoek gedaan naar formuleervaardigheid van vo-leerlingen, en dan met name naar hun formuleerfouten. Zo onderzocht Schuurs (1990) 448 verhalende en betogende opstellen van elf- tot veertienjarigen op veelvoorkomende formuleerfouten. Pander Maat, Raaijmakers, Vermeulen en De Glopper (2019) onderzochten 371 teksten uit de klassen 1–3 van de onderbouw op taalverzorgingsfouten. Van de Gein (2012) onderzocht 109 betogende en beschouwende teksten van eindexamenkandidaten havo-5 en vwo-6. Deze drie onderzoeken laten zich moeilijk vergelijken, allereerst omdat de onderzochte jaarlagen verschillend zijn en daarnaast omdat de fouten op verschillende wijzen zijn gecategoriseerd. Wel kunnen op basis van deze onderzoeken twee conclusies worden getrokken. (1) Formuleerfouten in leerlingenteksten komen veel voor, van de brugklas tot en met eindexamenklassen. Schuurs (1990) constateert dat één op de 5 à 6 zinnen van elf- tot veertienjarigen een evidente grammaticale fout bevat. Van de Gein (2012) con-