Ten Geleide

Op basis van een docent-oordeel over één geschreven tekst valt nauwelijks te generaliseren naar de schrijfvaardigheid van een leerling, gezien de grote verschillen tussen oordelen van verschillende docenten én de grote verschillen tussen prestaties op verschillende schrijftaken van leerlingen. In het onderzoek van Renske Bouwer en Huub van den Bergh, waarmee dit nummer van *Levende Talen Tijdschrift* opent, is nagegaan hoeveel taken en beoordelaars nodig zijn voor een betrouwbaar oordeel over de individuele schrijfvaardigheid in het Nederlands. Op basis van teksten geschreven door leerlingen van groep 8 in uiteenlopende onderwerpen en genres, kan geconcludeerd worden dat leerlingen voor een betrouwbaar oordeel over hun schrijfvaardigheid tenminste drie teksten in vier verschillende genres moeten schrijven (twaalf teksten in totaal), en dat de kwaliteit van teksten door ten minste twee beoordelaars moet worden bepaald.

Kunnen docenten en opleiders met 'oude' Cambridge-examens tot oordelen over de taalvaardigheid Engels van hun leerlingen en studenten komen die vergelijkbaar zijn met de oordelen van Cambridge-examinatoren? Gezien een aantal financiële, organisatorische en inhoudelijke bezwaren die kleven aan de officiële Cambridge-examens is het antwoord op deze vraag voor scholen en opleidingen interessant. Een pilot-onderzoek van Ton Koet en Huub van den Bergh onder studenten aan de lerarenopleiding Engels van vier hogescholen geeft echter geen aanleiding te veronderstellen dat het antwoord bevestigend zou zijn. Op basis van de gepresenteerde resultaten lijkt het aanbevelenswaardig voor docenten en opleiders om grote terughoudendheid te betrachten bij pogingen de taalvaardigheid Engels aan de hand van oude Cambridge-examens te beoordelen.

Een nieuwe rubriek in dit nummer is *Work in Progress*, waarin promovendi, onderzoekers, lectoren en lerarenopleiders in kort bestek een tussenstand kunnen geven van hun onderzoek of dat van hun masterstudenten. Johan van Hattum geeft een beschrijving van zijn promotieonderzoek naar effectief Engels uitspraakonderwijs.

Rick de Graaff bespreekt het proefschrift van Gerdineke van Silfhout met de titel *Fun to read or easy to understand? Establishing effective text features for educational texts on the basis of processing and comprehension research*. Ton Koet signaleert het recent verschenen boek *Language proficiency in native and non-native speakers; Theory and research* van Jan Hulstijn, hoogleraar Tweedetaalverwerving aan de Universiteit van Amsterdam. Marjon Tammenga-Helmantel signaleert een speciale editie van het tijdschrift *Language Learning* over onderzoek naar volgorde in het verwerven van een tweede en vreemde taal, en de ontwikkelingen die daarin de laatste decennia hebben plaatsgevonden.

Namens de redactie,
Helge Bonset

Toetsen van schrijfvaardigheid: hoeveel beoordelaars, hoeveel taken?

Renske Bouwer & Huub van den Bergh

Op basis van een docent-oordeel over één geschreven tekst valt nauwelijks te generaliseren naar de schrijfvaardigheid van een leerling, gezien de grote verschillen tussen oordelen van verschillende docenten én de grote verschillen tussen prestaties op verschillende schrijftaken van leerlingen. In dit onderzoek is nagegaan hoeveel taken en beoordelaars nodig zijn voor een betrouwbaar oordeel over de individuele schrijfvaardigheid. Op basis van teksten geschreven door leerlingen van groep 8 in uiteenlopende onderwerpen en genres, kan geconcludeerd worden dat leerlingen ten minste drie teksten in vier verschillende genres moeten schrijven (12 teksten in totaal), en dat de kwaliteit van teksten door ten minste twee beoordelaars moet worden bepaald.

Om te bepalen of leerlingen begrijpelijke teksten kunnen schrijven voor uiteenlopende communicatieve doeleinden, geven docenten Nederlands schrijfopdrachten. Meestal is het de docent zelf die vervolgens beslist of de tekst die is geschreven van voldoende kwaliteit is, en dus of de leerling voldoende schrijfvaardig is. Zowel het geven van schrijfopdrachten als het beoordelen van teksten is een tijdrovende klus. Het is dan ook niet gek dat een docent vaak op basis van een enkele tekst beslist of de desbetreffende leerling zich schriftelijk voldoende kan uitdrukken. Maar in hoeverre is zo'n beslissing gerechtvaardigd? Een beslissing over rekenvaardigheid wordt toch ook niet op basis van de prestatie op een enkele som genomen? Is het bij het toetsen van schrijfvaardigheid dan niet ook nodig om een leerling meerdere teksten te laten schrijven voordat een uitspraak gedaan kan worden over diens schrijfvaardigheid?

Uit eerder onderzoek blijkt dat het oordeel over de kwaliteit van een willekeurig geschreven tekst inderdaad niet alleen wordt bepaald door de kwaliteit van de schrijver, maar ook door de beoordelaar en de schrijftaak. In dit onderzoek wordt gekeken naar de impact van de beoordelaar en de taak op oordelen over schrijfvaardigheid, om zo vast te stellen in hoeverre oordelen op basis van een enkele geschreven tekst gegeneraliseerd kunnen worden naar een algemeen oordeel over iemands schrijfvaardigheid, en wat de implicatie hiervan is voor het toetsen van schrijfvaardigheid in de praktijk.

Effecten van de beoordelaar en schrijftaak

Meningen van beoordelaars over tekstkwaliteit lopen vaak enorm uiteen (zie o.a. Godshalk, Coffman, & Swineford, 1966; Wesdorp, 1981). Verschillen tussen beoordelaars zijn een gevolg van verschillen in strengheid, aspecten van de tekst waarop wordt gelet (op de inhoud, op zinsopbouw en woordgebruik, of een combinatie daarvan) en de mate van betrokkenheid bij de schrijver.

Naast verschillen tussen beoordelaars blijken ook taken te verschillen in moeilijkheid, wat gevolgen heeft voor de kwaliteit van de geschreven tekst (zie o.a. Huang, 2009). Zo is het bijvoorbeeld voor leerlingen lastiger te schrijven over onderwerpen die verder weg staan van hun belevingswereld. Daarnaast blijken er ook verschillen tussen genres te zijn: het schrijven van argumentatieve teksten is bijvoorbeeld moeilijker dan het schrijven van narratieve teksten. Dit blijkt niet alleen uit de kwaliteit van de teksten, maar ook uit het schrijfproces dat bij argumentatieve teksten anders verloopt (Thierry, Favart, Beauvais, & Beauvais, 2009). Over het algemeen geldt dat hoe bekender een leerling is met een onderwerp of genre, hoe makkelijker het is om een tekst te schrijven die de boodschap helder overbrengt en voldoet aan de conventies behorende bij het vereiste doel en publiek.

Door beoordelaars- en taakeffecten is het uiterst ingewikkeld om een oordeel over de kwaliteit van één geschreven tekst te generaliseren naar de schrijfvaardigheid van een leerling. Immers, wanneer de leerling een willekeurig andere schrijfopdracht had gekregen, of de tekst door iemand anders beoordeeld was, dan had de score voor tekstkwaliteit zomaar hoger of lager kunnen zijn. En dat betekent een ander oordeel over de schrijfvaardigheid van die leerling.

Taakeffecten zouden niet zo problematisch zijn als alle leerlingen dezelfde taken even moeilijk vonden. Op grond van de geschreven tekst komt men dan immers nog altijd tot dezelfde rangorde van goede en minder goede schrijvers; er hoeft alleen maar gecorrigeerd te worden voor de moeilijkheidsgraad van de schrijftaak. De praktijk leert echter dat niet elke leerling dezelfde taak even moeilijk vindt. Dit maakt het bepalen van een rangorde bijna onmogelijk, deze verschilt immers van taak tot taak. Zie bijvoorbeeld figuur 1 voor een illustratie van de verschillen tussen twee teksten, geschreven door dezelfde leerling. Op basis van deze beide teksten kan men tot verschillende oordelen over schrijfvaardigheid van deze leerling komen; de ene tekst is duidelijk minder adequaat dan de andere. Om een valide en betrouwbaar oordeel te vellen over de schrijfvaardigheid van leerlingen is daarom meer informatie nodig dan de prestatie op een enkele tekst.

Hoeveel beoordelaars en taken zijn nodig voor een betrouwbare schrijftoets?

Uit eerder onderzoek blijkt dat er ten minste 5 taken en 3 beoordelaars nodig zijn voor een goede meting van schrijfvaardigheid (o.a. Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012; Schoonen, 2005). Genoemd onderzoek verschilt echter nogal in het type taken dat is gebruikt; waar in het ene onderzoek leerlingen zeer vergelijkbare teksten moesten schrijven, zoals allemaal (korte) betogen (Van den Bergh et al, 2012), moesten ze in het andere onderzoek juist zeer diverse teksten schrijven, bijvoorbeeld betogen, maar ook routebeschrijvingen en instructie-teksten (Schoonen, 2005).

Bij het schrijven in verschillende genres werd echter steeds maar één tekst per genre geschreven. Het is daarom onduidelijk in hoeverre verschillen tussen teksten worden veroorzaakt door het onderwerp of het genre van de tekst. Te verwachten valt dat teksten binnen hetzelfde genre meer op elkaar lijken dan teksten van andere genres, met als gevolg

dat er minder taken afgenomen hoeven te worden om een goed beeld te krijgen over de onderliggende schrijfvaardigheid. Maar zegt de prestatie op soortgelijke taken dan alleen iets over schrijven binnen het betreffende genre, of zegt ze ook iets over schrijfvaardigheid in het algemeen?

Om te bepalen hoe een schrijftoets eruit moet zien om een valide en betrouwbaar oordeel te kunnen geven over schrijfvaardigheid, is het dus van belang om te weten wat het effect is van genre op de generaliseerbaarheid van scores voor tekstkwaliteit. In dit onderzoek wordt nagegaan in hoeverre verschillen in tekstkwaliteit worden bepaald door verschillen tussen leerlingen, taken (binnen en tussen genres) en beoordelaars. Daarnaast wordt gekeken hoeveel beoordelaars en taken (in verschillende genres) nodig zijn om een uitspraak te doen over de schrijfvaardigheid van leerlingen.

Beste MR supercoop

Ik ben Jip Jansen en heb een opmerking over de Smurfen actie. Er staat dat het tot 20 april 2008.
Maar het is nu 10 15 april en de actie lich g al meer
dan een week stil. Hier kan ik niet mee eens.
ik moet nog 2 smurfen maar anderen misschien nog wel
10 of 20 mij vader heeft ook al een klacht ingedient
maar het holp niet. Daarom hoop ik dat het met deze brief wel lukt.

groeten Jip Jansen

o, ps, dit is mij adres kruidenstraat 12
4567 DP halsteren Tel 0123-456789 of
ik hoop snel iets van u te horen.

Hallo smikkel

Ik ben door jullie actie als een dolle stier naar de winkel gegaan om voor die muziek cd wikkels te kopen. Maar nu ben ik een teleurgestelde en misschien wel boze stier. want zoals jullie weten is het nog lang
geen 15 juli maar zittten er geen spaarpunten meer op de wikkel. ik heb het bij elke supermarkt waar je wikkels kunt krijgen gebrobeert. Maar niks en ik moest er nog maar 2 hebben. daarom schrijf ik deze brief en zitten de 8 spaar punten en 2 wikkels zonder punten opgestuurt. In de hoop deze arme stier toch nog die cd krijgt.

Mijn adres is: kruidenstraat 12
en me naam is Jip Jansen

Smikkel spaar actie
postbus 3333
1273 AB D etten-leur

Figuur 1. Twee argumentatieve brieven geschreven door dezelfde leerling. De linker tekst heeft een duidelijk benedengemiddelde score en de rechtertekst heeft een duidelijk bovengemiddelde score

Methode van onderzoek

In dit onderzoek is gebruik gemaakt van teksten geschreven door 67 leerlingen uit groep 8 van drie random geselecteerde basisscholen in Nederland (zie Pullens, 2012). Op drie verschillende momenten schreven zij in totaal 12 teksten in vier verschillende genres. De genres verschilden van elkaar in het communicatieve doel (argumentatief of verhalend) en het beoogde publiek (specifieke lezer of abstract publiek). Binnen elk genre schreven de leerlingen drie teksten die alleen qua onderwerp

van elkaar verschilden; zie tabel 1 voor een overzicht van de gebruikte schrijftaken.

De kwaliteit van de teksten is beoordeeld door een team van 32 leraren in opleiding (pabo). Om te controleren voor mogelijke effecten van handschrift op de beoordelingen zijn alle teksten overgetypt, waarbij fouten en doorhalingen van de leerling zijn overgenomen. De beoordelaars kregen de teksten geanonimiseerd en random toegewezen, op zo'n manier dat elke tekst door drie personen werd beoordeeld. Beoordelaars gaven de leerlingteksten een holistisch oordeel voor tekstkwaliteit. Om tot dit oordeel te komen vergeleken ze elke tekst met een ankertekst van gemiddelde kwaliteit, waarbij voor de criteria Inhoud, Structuur, Conventies en Taalgebruik was aangegeven waarom de tekst de gemiddelde kwaliteit representeerde. Er waren in totaal vier ankerteksten, voor elk genre één. Deze ankerteksten zijn in een eerder stadium geselecteerd door een team van ervaren beoordelaars. De ankerteksten hadden een willekeurige score van 100 punten; betere teksten kregen daarom meer dan 100 punten en slechtere teksten minder dan 100 punten. De gemiddelde betrouwbaarheid van de oordelen van de jury's van drie beoordelaars was hoog, variërend van $\rho = 0{,}73$ tot $\rho = 0{,}86$ per taak.

COMMUNICATIEF DOEL	PUBLIEK	
	GESPECIFICEERD	ONGESPECIFICEERD
ARGUMENTATIEF	Formele brief aan een fictief bedrijf Onderwerpen: • Spaaractie voor smurfen • Spaaractie voor musicalzegels • Spaaractie voor muziek-CD	Betogende teksten ter voorbereiding van een discussie met de klas Onderwerpen: • Rookverbod • Snoepen • Klikken
VERHAAL	Avonturenverhaal voor lezers van een schoolkrant Onderwerpen: • Avontuur op een sportveld • Avontuur van een bosbrand • Avontuur met een vergiftiging	Verhaal over een persoonlijke ervaring Onderwerpen: • Geschrokken • Betrapt • Alleen thuis

Tabel 1. Een overzicht van de gebruikte schrijftaken, ingedeeld naar het communicatieve doel en publiek

Generaliseerbaarheidstheorie

Om te onderzoeken in hoeverre schrijfscores worden bepaald door de vaardigheid van een leerling worden verschillende bronnen van meetfouten (variantiecomponenten) onderscheiden, zoals het effect van de beoordelaar en taak (zie ook Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In ons onderzoek zijn er, conform de vraagstelling, zeven variantiecomponenten: 1. leerling, 2. genre, 3. interactie tussen leerling en genre, 4. taak binnen genre, 5. interactie tussen leerling en taak, 6. beoordelaars, 7. (onverklaarde) foutenvariantie.

In de eerste stap (de G-studie) wordt de (relatieve) bijdrage van elke variantiecomponent aan de totale variantie geschat. Hierdoor is het mogelijk om te bepalen wat nodig is om schrijfvaardigheid zo goed mogelijk te meten. Bijvoorbeeld: als de component *taak* verhoudingsgewijs veel groter is dan de component *beoordelaar*, dan is het, om tot een oordeel over de schrijfvaardigheid te komen, veel efficiënter om meer taken af te nemen dan om meer beoordelaars in te schakelen.

Hoeveel taken en beoordelaars precies nodig zijn om tot een valide en betrouwbare uitspraak te komen over de schrijfvaardigheid van een leerling wordt berekend in de tweede stap (de D-studie). Hierbij wordt voor diverse scenario's met verschillend aantal taken, beoordelaars en genres de generaliseerbaarheidscoëfficiënt berekend*, dit is de mate waarin scores de schrijfvaardigheid weerspiegelen.

Resultaten

Uit de resultaten blijkt dat de verschillen tussen scores voor tekstkwaliteit voor maar 10% worden bepaald door de schrijfvaardigheid van leerlingen. Dat betekent dat er slechts een matig verband is tussen de prestatie op twee willekeurige taken, waarbij de kwaliteit van teksten wordt beoordeeld door twee willekeurige beoordelaars ($r = 0{,}32$). Verschillen tussen teksten worden dus voor maar liefst 90% bepaald door factoren die niet (direct) gerelateerd zijn aan individuele schrijfvaardigheid, bijvoorbeeld het genre of het onderwerp van de schrijftaak, of de beoordelaar. In tabel 2 is een overzicht gegeven van het percentage verklaarde variantie voor de verschillende variantiecomponenten.

COMPONENT	VARIANTIE (IN %)
Leerling (l)	9,98
Genre (g)	11,42
Leerling x genre (lg)	4,01
Taak binnen genre (t:g)	1,71
Leerling x taak binnen genre (l(t:g))	19,13
Beoordelaar binnen taak en binnen genre (r:t:g)	18,05
Leerling x beoordelaar binnen taak, binnen genre, plus onverklaarde variantie (l(r:t:g), e)	35,71

Tabel 2. Percentage van verklaarde variantie voor elk van de zeven variantiecomponenten

Van deze 90% wordt 11% van de variantie verklaard door het genre waarin wordt geschreven. Sommige genres blijken moeilijker te zijn dan andere genres. Dit heeft gevolgen voor beslissingen die worden gemaakt op basis van schrijfprestaties. In het geval van de absolute beslissing of een leerling voldoet aan een bepaald niveau, maakt het nogal uit in welk genre er wordt geschreven. Als een leerling bijvoorbeeld toevallig een tekst in een moeilijk genre heeft geschreven, dan kan ten onrechte de beslissing worden genomen dat de leerling niet voldoende kan schrijven. Was het een makkelijker genre geweest, dan had de tekst immers wel voldoende kunnen zijn. Om een goed beeld te krijgen van het niveau van schrijfvaardigheid moeten leerlingen dus meerdere teksten in meerdere genres schrijven, of moet er gecorrigeerd worden voor de moeilijkheid van de taak.

Naast een hoofdeffect van genre blijkt 4% van de verschillen tussen scores verklaard te worden door de interactie tussen leerlingen en genre. Dit betekent dat de rangorde van leerlingen enigszins verschilt tussen genres: niet alle leerlingen vinden dezelfde genres even lastig of gemakkelijk. De rangorde van leerlingen blijkt ook duidelijk te verschillen binnen genres, aangezien maar liefst 20% van de variantie wordt verklaard door de interactie tussen leerling en taak (binnen eenzelfde genre). Waar de ene leerling naar aanleiding van een schrijfopdracht redelijk gemakkelijk een goede tekst schrijft, heeft een andere leerling hier aanzienlijk meer moeite mee, maar bij een andere taak is dit juist weer andersom. De resterende variantie tussen tekstscores wordt verklaard door verschillen tussen beoordelaars (18%) en door de interactie tussen leerlingen en beoordelaars (op taakniveau) plus nog onverklaarde variantie (36%).

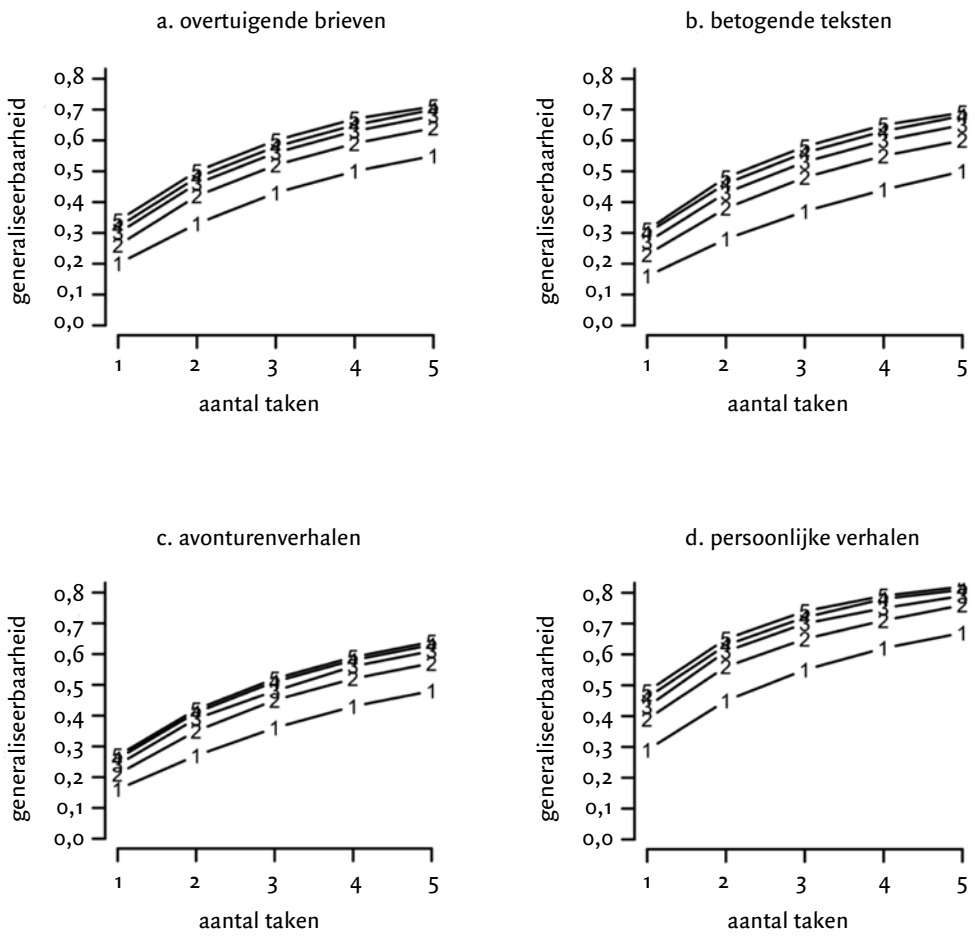
De prestatie van leerlingen is dus afhankelijk van de schrijftaak (genre, onderwerp) en de beoordelaar; hoe meer taken en beoordelaars betrokken zijn bij de toets, hoe betrouwbaarder het oordeel is over iemands schrijfvaardigheid. Om een generaliseerbaarheidscoëfficiënt van 0,70 te bereiken, waarbij het verschil tussen leerlingen voor 70% wordt verklaard door verschil in schrijfvaardigheid, moeten ten minste drie verschillende schrijftaken in vier verschillende genres worden afgenomen (12 teksten in totaal) waarbij de teksten worden beoordeeld door ten minste twee beoordelaars. Een hogere generaliseerbaarheidscoëfficiënt is natuurlijk wenselijk, maar vereist een nog hoger aantal taken en beoordelaars.

Verschillen tussen genres

In het geval dat we alleen een oordeel willen geven over de schrijfvaardigheid binnen een specifiek genre zijn er minder teksten nodig. Het aantal taken en beoordelaars dat dan nodig is blijkt te verschillen per genre. In figuur 2 is voor elk van de vier genres te zien hoe de generaliseerbaarheidscoëfficiënt (y-as) toeneemt naargelang er meer taken (x-as) en meer beoordelaars (de lijnen in het figuur) bij de toets betrokken zijn. Bijvoorbeeld voor overtuigende brieven is de generaliseerbaarheid bij één taak en één beoordelaar slechts 0,20. Op basis van eén tekst valt dus nauwelijks iets te zeggen over de schrijfvaardigheid binnen dit genre. De generaliseerbaarheid stijgt naar 0,33 bij vijf beoordelaars die één tekst van een leerling beoordelen. Dit is nog steeds niet voldoende om tot een betrouwbaar oordeel over de schrijfvaardigheid te komen. Een minimale generaliseerbaarheid van 0,70 wordt echter wel bereikt wanneer de leerling meer teksten schrijft. In het geval van vijf teksten, beoordeeld door 5 beoordelaars, is de generaliseerbaarheid 0,71.

Uit figuur 2 blijkt dat leerlingen consistenter zijn in het schrijven van verhalen over persoonlijke ervaringen dan wanneer het

andere genres betreft. Ook blijken beoordelaars de kwaliteit van de persoonlijke verhalen consistenter te beoordelen; beoordelaars zijn het meer met elkaar eens over wat een goed persoonlijk verhaal is dan dat ze het bijvoorbeeld eens zijn over een overtuigende brief. De implicatie hiervan is dat er minder informatie nodig is om te kunnen generaliseren naar het schrijven van persoonlijke verhalen dan naar het schrijven in één van de andere drie genres. Om precies te zijn: voor persoonlijke verhalen zijn maar drie taken en drie beoordelaars nodig om tot een oordeel te komen waarmee uitspraken gedaan kunnen worden over verschillen tussen leerlingen in schrijfvaardigheid, terwijl voor het schrijven in de andere genres maar liefst vijf taken en vijf beoordelaars nodig zijn.



Figuur 2. De generaliseerbaarheid van schrijfscores voor relatieve beslissingen in elk van de vier genres, bij verschillend aantal taken (x-as) en beoordelaars (de getallen bij de lijnen representeren het aantal beoordelaars)

Conclusie en discussie

In deze studie is onderzocht hoeveel schrijftaken en beoordelaars nodig zijn om scores voor tekstkwaliteit te kunnen generaliseren naar individuele schrijfvaardigheid. Ook is gekeken wat de rol van genre hierbij is; is het mogelijk om op basis van zeer homogene schrijfopdrachten (met hetzelfde communicatieve doel en publiek) te kunnen generaliseren over genres, of moet het oordeel in dit geval beperkt blijven tot schrijfvaardigheid binnen een specifiek genre?

Uit de verschillen tussen scores voor tekstkwaliteit van teksten in vier verschillende genres, geschreven door leerlingen uit groep 8, blijkt dat maar 10% van de variantie verklaard kan worden door individuele schrijfvaardigheid en blijkt een effect van genre op de generaliseerbaarheid van tekstscores. Om te kunnen generaliseren over genres moet een schrijftoets maar liefst twaalf taken bevatten, waarbij leerlingen drie teksten in ten minste vier genres schrijven, en waarbij de kwaliteit van elke tekst door ten minste twee beoordelaars wordt bepaald. Als het uit praktisch oogpunt niet haalbaar is om zoveel taken tegelijkertijd af te nemen, dan is het wenselijk om minder, maar wel meer gelijksoortige, taken af te nemen. In dit geval kunnen de uitkomsten niet gegeneraliseerd worden over genres, maar alleen naar schrijven binnen een specifiek genre. Hoeveel taken en beoordelaars dan nodig zijn voor een betrouwbare toets hangt af van het desbetreffende genre waarnaar een docent wil generaliseren.

De resultaten van deze studie sluiten aan bij eerder onderzoek waarin al bleek dat scores voor tekstkwaliteit van individuele studenten enorm variëren tussen beoordelaars (o.a. Godshalk, Coffman, & Swineford, 1966) en taken (o.a. Huang, 2009). Waar in eerder onderzoek effecten van genre en taak altijd zijn gecontamineerd, bijvoorbeeld doordat meerdere taken zijn afgenomen in maar één genre, of in meerdere genres maar dan wel met slechts één taak per genre, werden de effecten van taak en genre in deze studie uit elkaar gehaald door leerlingen meerdere taken in meerdere genres te laten schrijven. Uit deze studie blijkt dat genre een effect heeft op de stabiliteit van individuele schrijfprestaties, nog bovenop de effecten van het onderwerp. Het gaat hierbij voornamelijk om een hoofdeffect van genre. Over het algemeen ervaren leerlingen dus dezelfde problemen met dezelfde genres en is maar een klein deel van de variantie gerelateerd aan de interactie tussen leerling en genre.

Individuele verschillen in moeilijkheid met genres kunnen verklaard worden door genre-specifieke kennis (Bhatia, 1993). Hoe bekender leerlingen zijn met een specifiek genre, hoe beter zij zich bewust zijn van genre-specifieke conventies en deze kunnen toepassen op de tekst, wat uiteindelijk resulteert in teksten van enigszins vergelijkbare kwaliteit. Het is daarom niet geheel onverwacht dat de leerlingen in deze studie meer consistent presteren bij het schrijven van persoonlijke verhalen dan bij andere genres, aangezien op de basisschool geregeld wordt gecommuniceerd over persoonlijke ervaringen.

Op basis van de resultaten uit deze studie kan men zich afvragen of er wel zoiets bestaat als een *algemene* schrijfvaardigheid. Als de prestatie op taken met verschillende communicatieve doelen en voor verschillend publiek immers zo van elkaar verschilt, moeten we schrijven dan niet eerder beschouwen als een verzameling aparte deelvaardigheden? Voor de onderwijspraktijk zou dat impliceren dat instructies zich richten op genre-specifiek schrijven en dat toetsen ook altijd genre-specifiek zijn. Uit deze studie blijkt echter dat er ook een enorme variantie is binnen genres; leerlingen lijken over het

algemeen elke schrijftaak aan te pakken alsof het een geheel nieuwe opdracht is. Om de transfer van kennis van de ene taak naar de andere taak te bevorderen, lijkt het juist van belang om aandacht te besteden aan het aanleren van algemene strategieën om schrijftaken effectief aan te pakken, en leerlingen bewust te maken hoe een tekst aankomt bij een lezer (bijvoorbeeld door feedback). Recent onderzoek laat zien dat zo'n algemene procesaanpak voor schrijven de taakverschillen binnen leerlingen kan verkleinen (Bouwer, Koster, & Van den Bergh, 2015).

Een kanttekening bij dit onderzoek is dat het niet duidelijk is of verschillen binnen genres wel echt worden veroorzaakt door het onderwerp. Om dat te onderzoeken hadden de onderwerpen voor elk genre hetzelfde moeten zijn. Dit is echter niet mogelijk omdat niet elk onderwerp zich leent voor elk genre.

Op basis van de resultaten uit dit onderzoek kunnen we concluderen dat het niet wenselijk is om schrijfvaardigheid met eén enkele schrijftaak te toetsen. De geleverde prestatie op deze ene taak blijkt nauwelijks een voorspeller te zijn voor de prestatie op een willekeurige andere taak. Net als bij rekenen is er dus meer informatie nodig (meer taken in verschillende genres) om een valide en betrouwbaar oordeel te vellen over de schrijfvaardigheid van leerlingen. Daarnaast blijkt dat ook meer beoordelaars nodig zijn. Een alternatief is om te investeren in meer betrouwbare, objectieve beoordelingsprocedures, waarbij verschillen tussen beoordelaars worden geminimaliseerd.

* Formule: $$\rho = \frac{S^2_{leerling}}{S^2_{leerling} + \frac{S^2_{genre}}{N_{genre}} + \frac{S^2_{leerling*genre}}{N_{leerling}*N_{genre}} + \frac{S^2_{taak}}{N_{taak}} + \frac{S^2_{leerling*taak}}{N_{leerling}*N_{taak}} + \frac{S^2_{beo}}{N_{beo}} + \frac{S^2_{leerling*taak*beo}}{N_{leerling}*N_{taak}*N_{beo}}}$$

Verantwoording

Dit artikel is een Nederlandse bewerking van het artikel 'Effect of genre on the generalizability of writing scores' dat in 2015 verscheen in het tijdschrift *Language Testing*. We danken Theo Pullens voor het beschikbaar stellen van zijn data waarop de resultaten van deze studie zijn gebaseerd.

Literatuur

Bergh, H. van den, Maeyer, S. de, Weijen, D. van, & Tillema, M. (2012). Generalizability of text quality scores. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (vol. 27, pp. 23–32). Leiden: Brill.

Bouwer, R., Koster, M., & Van den Bergh, H. (2015). *Improving the writing skills of students in the upper elementary grades: An intervention study.* Aangeboden ter publicatie.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements.* New York: John Wiley.

Godshalk, F. I., Coffman, W. E., & Swineford, F. (1966). *The measurement of writing ability.* New York: College Entrance Examination Board.

Huang, C. (2009). Magnitude of task-sampling variability in performance assessment: A meta-analysis. *Educational and Psychological Measurement, 69*(6), 887–912.

Olive, T., Favart, M., Beauvais, C., & Beauvais, L. (2009). Children's cognitive effort and fluency in writing: Effects of genre and of handwriting automatisation. *Learning and Instruction, 19*(4), 299–308.

Pullens, T. J. M. (2012). *Bij wijze van schrijven* (dissertatie). Universiteit Utrecht.
Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22(1), 1–30.
Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs.* Den Haag: Stichting voor Onderzoek van het Onderwijs.

Renske Bouwer (1984) studeerde aan de Radboud Universiteit, waar zij in 2008 de research master Behavioural Science afrondde. In 2012 is zij gestart met een promotieproject (aan de Universiteit Utrecht) naar verbetering van de schrijfvaardigheid van leerlingen in de bovenbouw van het basisonderwijs. Hierbij concentreert zij zich onder meer op de vraag hoe schrijfvaardigheid van leerlingen valide en betrouwbaar geëvalueerd kan worden, en hoe aan leerlingen adequate feedback op hun teksten gegeven kan worden.
E-mail: <i.r.bouwer@uu>

Huub van den Bergh is hoogleraar toetsing en didactiek van taalvaardigheid aan de Universiteit Utrecht. Hij was betrokken bij vele groot- en kleinschalige projecten naar de kwaliteit van ons (taal)onderwijs.
E-mail: <h.vandenbergh@uu.nl>

Beoordeling van Engelse taalvaardigheid door docenten en Cambridge-examinatoren

Ton Koet & Huub van den Bergh

Kunnen docenten en opleiders met 'oude' Cambridge-examens tot oordelen over de taalvaardigheid van hun leerlingen en studenten komen die vergelijkbaar zijn met de oordelen van Cambridge-examinatoren? Gezien de nadelen van de officiële examens is het antwoord op deze vraag voor scholen en opleidingen interessant. Een pilot-onderzoek gaf echter geen aanleiding te veronderstellen dat het antwoord bevestigend zou zijn.

In de kennisbases voor tweedegraadsopleidingen voor leraar moderne vreemde talen (Duits, Engels, Frans en Spaans) zijn aan het Europees Referentiekader (ERK) gerelateerde eisen aan de taalvaardigheid geformuleerd (HBO-raad, 2009). De vier vreemde talen verschillen in de vereiste niveaus, waarbij alleen Frans een onderscheid maakt tussen de vaardigheden (zie tabel 1). Voor Engels staat daarnaast vermeld dat een voldoende voor het Certificate of Proficiency in English (CPE) is vereist.

TAAL	VAARDIGHEID	ERK-NIVEAU
Duits		C1
Engels		C; niet nader gespecificeerd
Frans	Leesvaardigheid	C1
	Kijk- en luistervaardigheid	B2 – C1
	Schrijfvaardigheid	niet gespecificeerd
	Spreekvaardigheid	B2 – C1
	Gespreksvaardigheid	B2 – C1
Spaans		hoog B2

Tabel 1. De vereiste taalvaardigheidsniveaus voor de vier moderne vreemde talen