Kan schrijfvaardigheid betrouwbaar gemeten worden?

Een onderzoek naar de toepassingsmogelijkheden van schaalbeoordeling in het basisonderwijs

CARLA VAN ROOIJEN

Het beoordelen van schrijfvaardigheid in het basisonderwijs vormt een problematische kwestie: verschillende beoordelaars zijn het vaak sterk oneens over de kwaliteit van schrijfopdrachten. Een mogelijke oplossing voor dit probleem is de schaalbeoordelingsmethode. Uit voorgaand onderzoek is gebleken dat gebruik van deze methode een hogere beoordelaars-overeenstemming bewerkstelligt. De methode is echter nog niet grootschalig inzetbaar omdat het opstellen van een beoordelingsschaal te arbeidsintensief is. In dit onderzoek werd getest of een enkele schaal breder inzetbaar is. De resultaten suggereren dat dezelfde beoordelingsschaal bij de beoordeling van verschillende opdrachten te gebruiken is. Meer onderzoek is nodig om de grenzen van deze generalisering te definiëren.

Sinds een aantal jaren is er veel aandacht voor de schrijfvaardigheid van basisschoolleerlingen. Uit verschillende periodieke peilingen is gebleken dat de schrijfvaardigheid van leerlingen uit groep 8 ernstig te wensen overlaat (Inspectie van Onderwijs, 2010). Om verandering te brengen in deze negatieve trend wordt dan ook veel onderzoek gedaan ter verbetering van het schrijfonderwijs en ter ontwikkeling van nieuwe lesmethoden. Of deze maatregelen een gunstig effect hebben dient vervolgens gemeten te worden door de schrijfvaardigheid van leerlingen te blijven toetsen.

Schrijfvaardigheidsonderzoek brengt echter altijd een groot probleem met zich mee: het construct 'schrijfvaardigheid' is lastig te definiëren en om deze reden ook lastig te meten. Want hoe bepaal je eigenlijk of een schrijver schrijfvaardig is? De beste manier lijkt te zijn door de schrijver zijn vaardigheden te laten demonstreren, en dus door hem een schrijftaak te laten uitvoeren. De schrijftaak functioneert hierbij als de operationalisatie van het construct schrijfvaardigheid. Door het schrijfproduct te beoordelen hoopt men een uitspraak te kunnen doen over de vaardigheden van de schrijver. Beoordelaars kunnen het echter sterk met elkaar oneens zijn over de eigenschappen van een goede tekst, waardoor de betrouwbaarheid bij de beoordeling van schrijfvaardigheid vaak ver te zoeken is.

In dit onderzoek zullen verschillende beoordelingsmethoden van schrijfvaardigheid onder de loep genomen worden, en zal vervolgens gepoogd worden de toepasbaarheid van een van deze methoden, de schaalbeoordeling, uit te breiden. Het doel is hiermee een beoordelingsmethode te creëren die zowel betrouwbaar als toepasbaar is in het onderwijs.

Globale beoordeling

Om de schrijfvaardigheid van leerlingen te bepalen worden verschillende methoden gehanteerd. De meest gebruikte methode door leerkrachten is *globale beoordeling*: een beoordelingsmethode waarbij het cijfer wordt gebaseerd op een globale indruk van de tekstkwaliteit. Deze methode heeft enkele voordelen: het kost weinig tijd om een cijfer vast te stellen, en iedere tekst wordt als geheel beschouwd (Wesdorp, 1981; Pollmann, Prenger & De Glopper, 2012). Door een gebrek aan duidelijke beoordelingscriteria brengt deze aanpak echter veel nadelen met zich mee.

Het voornaamste gebrek van de globale beoordelingsmethode is dat de beoordeling (door het gebrek aan beoordelingscriteria) afhankelijk is van de tekstfactoren die voor de betreffende docent het belangrijkst zijn (het *signifisch effect)*. Zodoende kunnen er enorme verschillen ontstaan tussen beoordelingen van verschillende docenten en wordt de meting onbetrouwbaar (Wesdorp, 1981). Een veelgehoorde en -gebruikte oplossing voor dit probleem is de zogenoemde *jurybeoordeling*, waarbij verschillende beoordelaars dezelfde opdracht beoordelen. Het nadeel hiervan is echter dat beoordelingen vaak gereduceerd worden tot een nietszeggend gemiddelde, waarvan geen enkel specifiek criterium de basis lijkt te vormen. Dit verhoogt dus wellicht de betrouwbaarheid van de meting, maar niet de validiteit.

Dat de validiteit van dergelijke beoordelingen al in het geding is, blijkt eveneens uit het gebrek aan beoordelingscriteria: wanneer iedere docent verschillende criteria hanteert, wordt niet bij iedere tekst hetzelfde construct gemeten (Mullis, 1984). Hier kom nog bij dat zelfs dezelfde docent niet altijd dezelfde criteria lijkt te hanteren. Er ontstaan namelijk niet alleen grote verschillen tussen beoordelaars, maar ook de consistentie binnen beoordelaars is ver te zoeken (Van den Bergh & Meuffels, 2000; Barkaoui, 2011). Beoordelaars zijn blijkbaar niet in staat om via globale beoordeling iedere keer op dezelfde wijze tot een oordeel te komen. Dit verklaart wellicht deels waarom prestaties van leerlingen bij verschillende schrijfopdrachten relatief laag met elkaar correleren (Schoonen, 2005).

Globale beoordeling vergroot bovendien de kans op ongewenste invloeden, bijvoorbeeld door het *halo-effect,* het *sequentie-effect* of door *normverschuiving*. Het *halo-effect* houdt in dat een docent zich in zijn oordeel laat beïnvloeden door irrelevante eigenschappen van de betreffende leerling of tekst (Wesdorp, 1981). Zo hebben tekstlengte en handschrift vaak een (te) grote invloed op het cijfer (Huot, 1990; Lee, Gentile, & Kantor, 2009). Ook kunnen schrijfproducten een lagere beoordeling krijgen wanneer de schrijver uit een lager sociaal milieu komt (Oudenhoven, 1983). Bij het *sequentie-effect* heeft de volgorde waarin schrijfproducten worden nagekeken invloed op de beoordeling, bijvoorbeeld wanneer een gemiddelde opdracht na een reeks goede opdrachten een te laag cijfer krijgt (Wesdorp, 1981; Hajer, Meestringa, Van der Leeuw, Prenger, De Glopper, & Van Dijk, 2012). Hieraan gerelateerd is *normverschuiving*, waarbij de beoordelaar zijn oordelen gedurende het beoordelingsproces aanpast aan het schrijfniveau (Hajer et al., 2012).

Hoewel de globale beoordelingsmethode dus gemakkelijk in gebruik is, zorgt hij niet voor een betrouwbaar en valide oordeel over de schrijfvaardigheid van leerlingen.

Analytische beoordeling

Om enkele van de vele bovenstaande problemen het hoofd te bieden wordt in veel leeromgevingen de *analytische beoordeling* toegepast. Hierbij wordt een oordeel over het schrijfproduct gevormd door een aantal deelkenmerken van de tekst te beoordelen en deze oordelen vervolgens te combineren tot een totaaloordeel.

Deze methode lijkt het grootste probleem van de globale methode – namelijk dat er geen duidelijke criteria aan de beoordeling ten grondslag liggen – op te lossen. Hierbij moet echter de vraag gesteld worden of via analytische beoordeling wel 'schrijfvaardigheid' gemeten wordt: kan een tekst wel beschouwd worden als een optelsom van zijn deel-kenmerken? Het gevaar is dat door enkel deelkenmerken te beoordelen de

tekst(kwaliteit) als geheel uit het oog wordt verloren (Mullis, 1984; Wesdorp, 1981; Barkaoui, 2011).

Er is al veel onderzoek uitgevoerd naar de categorieën die onderscheiden zouden moeten worden ter beoordeling van teksten. Hierbij formuleren verschillende onderzoekers echter zeer uiteenlopende categorieverdelingen. Zo onderscheidt de ene onderzoeker vier categorieën (zoals organisatie, conventies, stijl, en inhoud) en de andere onderzoeker zes (t.w. algemene stilistische evaluatie, persoonlijke affectie, versiering, abstractheid, ernst, en typerend vs. verhalend) (Wesdorp, 1981). Daarnaast is niet duidelijk hoe de verschillende categorieën zich tot elkaar zouden moeten verhouden: dienen ze allemaal even zwaar mee te tellen bij de beoordeling, of is de ene categorie belangrijker dan de andere? Ook over deze vraag is men het niet eens (Van den Bergh & Meuffels, 2000). Bij gebrek aan een eenduidig beoordelingsschema kan men zich voorstellen dat de keuze van het schema invloed heeft op de uiteindelijke beoordeling. De vraag rijst dus wederom of via deze methode wel per definitie het construct 'schrijfvaardigheid' gemeten wordt.

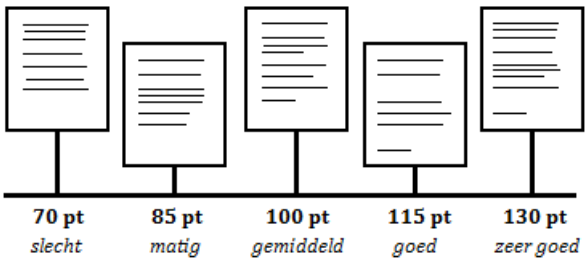
Daarnaast biedt deze methode geen garantie dat alle afzonderlijke categorieën ook daadwerkelijk onafhankelijk van elkaar worden beoordeeld. Doordat de beoordelaar dezelfde tekst meerdere keren moet bekijken met steeds een andere invalshoek, kan het oordeel over het ene tekstkenmerk invloed hebben op het oordeel over het andere tekstkenmerk. Oordelen over verschillende tekstkenmerken blijken dan ook vaak hoog met elkaar te correleren (Wesdorp, 1981; Huot, 1990; Bacha, 2001; Lee et al., 2009). Bovendien correleren de scores van analytische beoordeling (zowel afzonderlijk als gemiddeld) vaak hoog met het globale beoordelingscijfer (Huot, 1990; Lee et al., 2009). Wanneer docenten niet in staat is de categorieën onafhankelijk van elkaar te beoordelen, voegt deze methode niets toe ten opzichte van de globale methode.

Toch zou deze methode verbetering opleveren wanneer de betrouwbaarheid (in dit geval: de overeenstemming tussen verschillende beoordelaars) zou worden verhoogd ten opzichte van globale beoordeling. Dit is geen onrealistische verwachting: door de categorieën te onderscheiden en duidelijk te definiëren zou men verwachten dat verschillende beoordelaars het meer eens zijn over de afzonderlijke oordelen (Barkaoui, 2011). De onderzoeksresultaten zijn op dit gebied echter allesbehalve eenduidig. In sommige gevallen leidde de methode tot iets hogere correlaties, maar in ander onderzoek was er geen verschil tussen globale en analytische beoordeling, of was de betrouwbaarheid zelfs iets hoger bij globale beoordeling (Wesdorp, 1981; Lee et al., 2009; Barkoui, 2011).

Hoewel de analytische methode de beoordeling wel transparanter maakt, levert deze methode dus minder verbetering op het vlak van betrouwbaarheid dan men in eerste instantie zou verwachten (Wesdorp, 1981; Van den Bergh & Meuffels, 2000). Bovendien gaat bij gebruik van deze methode een groot voordeel van globale beoordeling verloren: de methode vergt meer inspanning van de beoordelaar en dus ook meer tijd (Huot, 1990; Lee et al., 2009; Barkaoui, 2011).

Schaalbeoordeling

Een andere mogelijke beoordelingsmethode is de *schaalbeoordeling*. Deze methode vergt meer voorbereiding dan de andere methoden, namelijk de samenstelling van een beoordelingsschaal. Hiertoe dienen allereerst alle schrijfproducten van één schrijfopdracht globaal te worden beoordeeld. Vervolgens worden er door meerdere beoordelaars enkele teksten uitgekozen die representatief zijn voor verschillende schrijfniveaus (de zogenoemde 'ankerteksten') en in de gewenste volgorde op



Figuur 1. Schematische weergave van een beoordelingsschaal.

een puntenschaal worden geplaatst (zie figuur 1). De verdere beoordeling van de schrijfproducten vindt plaats door iedere tekst te vergelijken met de teksten op de schaal en ze op basis van deze vergelijking een score toe te kennen.

Ook aan deze methode zitten haken en ogen. Het voornaamste nadeel is dat het samenstellen van de schaal een zeer arbeidsintensieve klus is, die bovendien met uiterste zorgvuldigheid dient te worden uitgevoerd: wanneer de schaal niet klopt, is de hele verdere beoordeling onbetrouwbaar (Wesdorp, 1981). Het is dus van het grootste belang om de schaal te laten samenstellen door meerdere deskundigen, die het onderling eens moeten worden over de selectie en volgorde van de ankerteksten. Hierbij rijst uiteraard de vraag wie gezien kunnen worden als 'deskundigen' en hoeveel er nodig zijn om een betrouwbare schaal op te stellen.

De schaalbeoordelingsmethode heeft echter ook een groot voordeel: de betrouwbaarheid van de oordelen wordt ermee verhoogd. Uit onderzoek van Pollmann et al. (2012) is gebleken dat de oordelen van verschillende beoordelaars relatief hoog met elkaar correleren bij gebruik van een beoordelingsschaal. Daarnaast bleek uit onderzoek van Schoonen (2005) dat oordelen op basis van een schaal beter te generaliseren zijn over andere schrijftaken dan oordelen op basis van een analytische beoordeling. Er zijn zodoende minder beoordelaars/schrijftaken nodig om een representatief beeld te krijgen van de schrijfvaardigheid van een leerling.

Ook vermindert gebruik van een schaal de kans op sequentie-effecten of normverschuiving, doordat de beoordelaar gedwongen wordt ieder schrijfproduct met de ankerteksten te vergelijken (Schoonen, 2005; Hajer et al., 2012; Pollmann et al., 2012). Hiermee combineert het de voordelen van analytische beoordeling met het voornaamste voordeel van globale beoordeling: de tekst wordt als geheel beschouwd, en niet als een som der delen (Barkaoui, 2011). Bovendien kost schaal-beoordeling relatief weinig tijd zodra de beoordelaar bekend is met de ankerteksten (Pollmann et al., 2012).

Aanleiding voor het huidige onderzoek

Beoordelingsschalen leveren dus winst op ten opzichte van de andere methoden. Toch blijft er aan deze methode een enorm nadeel kleven: het kost erg veel werk om een beoordelingsschaal op te stellen. Om dit voor iedere schrijfopdracht in het basisonderwijs te doen zou een immense klus zijn, en zodoende is deze methode nog niet geschikt voor grootschalige toepassing in de praktijk.

Een oplossing voor dit probleem zou zijn om de methode efficiënter toe te passen. Dit zou bereikt kunnen worden door dezelfde schaal vaker te gebruiken, bijvoorbeeld door hem op meerdere scholen toe te passen. Hiermee blijft het probleem van het grote aantal verschillende opdrachten echter bestaan. Een nog efficiëntere toepassing van beoordelingsschalen zou bereikt kunnen worden door dezelfde schaal te gebruiken bij de beoordeling van verschillende opdrachten. Hierbij komen de ankerteksten dus niet overeen met de betreffende schrijfopdracht, maar is dit wel noodzakelijk? Schrijfproducten zijn immers slechts operationalisaties van het construct schrijfvaardigheid: het zijn als het ware representaties van een bepaald schrijfniveau. Hierdoor kan de exacte inhoud van de tekst als onafhankelijk gezien worden van de gemeten schrijfvaardigheid. Zodoende zou het niet ondenkbaar zijn om teksten met verschillende onderwerpen met elkaar te vergelijken om zo tot een oordeel te komen.

In een ideale situatie zou iedere willekeurige opdracht met iedere willekeurige schaal beoordeeld kunnen worden. Dit klinkt echter als een te optimistische aanname: wanneer de schrijfproducten en de ankerteksten te veel verschillen, is het maar de vraag of beoordelaars in staat zijn om op gelijke wijze tot een adequate vergelijking tussen deze teksten te komen. Maar wat betekent 'te veel verschillen'? Hoe bepalen we welke teksten genoeg op elkaar lijken en welke niet?

Om teksten onderling te kunnen vergelijken is het belangrijk om vast te stellen of het om dezelfde soort tekst gaat, en dus om ze te kunnen classificeren (Renkema, 1987). Het verdelen van teksten in categorieën is echter een problematische bezigheid. De term

tekstgenre wordt vaak gehanteerd, maar welke tekstkenmerken precies ten grondslag liggen aan de verschillende onderverdelingen is vaak onduidelijk. Zo kan men het communicatieve doel, de vorm/het medium, of het onderwerp van de tekst in acht nemen, of een combinatie van deze factoren (Pander Maat, 2002). Wanneer we echter kijken naar de 'klassieke basisverdeling' zien we dat er drie teksttypen worden onderscheiden: informerend, argumenterend/betogend en verhalend (Renkema, 1987). Binnen deze onderverdeling ligt de nadruk op het communicatieve doel van de tekst. Inmiddels zijn er vele varianten van deze basisverdeling geconstrueerd, waaronder de verdeling van Pander Maat (2002), die stelt dat een tekst gedefinieerd kan worden "als een reeks taaluitingen die gezamenlijk een communicatief doel dienen. Het communicatieve doel is dus uiteindelijk verantwoordelijk voor de samenhang tussen de uitingen in de tekst." (Pander Maat, 2002, 266)

Pander Maat (2002) onderscheidt vijf teksttypen op basis van de verschillende communicatieve doelen: informatief, instructief, persuasief, directief en expressief. De keuze van het teksttype heeft invloed op de inhoudselementen die de tekst moet bevatten, en hiermee ook op de structuur van de tekst (Pander Maat, 2002). Hierdoor is het mogelijk dat teksten met verschillende communicatieve doelen lastiger te vergelijken zijn. Dit zou kunnen impliceren dat wanneer ankerteksten en schrijfproducten verschillende tekstdoelen hebben, de ankerteksten op de schaal te weinig ondersteuning bieden om de betreffende schrijfproducten adequaat te kunnen beoordelen.

Vraagstelling

Schaalbeoordeling zou toegepast kunnen worden in het basisonderwijs door dezelfde schaal te gebruiken bij verschillende opdrachten, maar het is nog onbekend in hoeverre dit mogelijk is. Wanneer de ankerteksten en schrijfproducten te sterk van elkaar afwijken en de schaal zodoende weinig ondersteuning biedt bij de beoordeling, zou dit mogelijk gevolgen kunnen hebben voor de beoordelaarsovereenstemming. De centrale vraag in dit onderzoek is dus:

> *Wat is de invloed van de congruentie tussen de beoordelingsschaal en de schrijfopdrachten op de beoordelaarsovereenstemming?*

Om antwoord te kunnen geven op deze vraag zullen oordelen van verschillende beoordelaars over drie verschillende schrijfopdrachten uit het basisonderwijs met elkaar vergeleken worden bij gebruik van dezelfde beoordelingsschaal. Om te bepalen in hoeverre iedere opdracht congrueert met de schaal zal gekeken worden naar de opdracht en het teksttype (tekstdoel). Hierbij congrueren ankerteksten en schrijfproducten hoger wanneer ze naar aanleiding van dezelfde opdracht geschreven zijn, en lager wanneer ze bij verschillende opdrachten horen. Daarnaast congrueren ze hoger wanneer ze tot hetzelfde teksttype behoren, en lager wanneer ze onder verschillende teksttypen vallen. Op basis van deze aannames kunnen drie congruentieniveaus opgesteld worden:

<u>Hoge congruentie</u>
: Ankerteksten en schrijfproducten horen bij dezelfde opdracht (en hebben dus ook hetzelfde teksttype)

<u>Middelmatige congruentie</u>
: Ankerteksten en schrijfproducten horen bij een andere opdracht, maar hebben hetzelfde teksttype

<u>Lage congruentie</u>
: Ankerteksten en schrijfproducten horen bij een andere opdracht en hebben een ander teksttype

Het doel van dit onderzoek is om te testen of middelmatige en lage congruentie tussen ankerteksten en schrijfproducten eveneens leidt tot een hoge interbeoordelaarsbetrouwbaarheid. De deelvragen van dit onderzoek luiden dus als volgt:

RQ1: Is de beoordelaarsovereenstemming bij een middelmatige congruentie tussen ankerteksten en schrijfproducten even hoog als

bij een hoge congruentie?

RQ2: Is de beoordelaarsovereenstemming bij een lage congruentie tussen ankerteksten en schrijfproducten even hoog als bij een hoge congruentie?

Hypothesen

De bovenstaande vragen zullen getoetst gaan worden aan de hand van twee soorten teksten: *directieve (*specifiek: *verzoekende)* en *instructieve* teksten. Door hun uiteenlopende tekstdoelen hebben deze teksttypen verschillende structuren: bij *verzoeken* ligt de nadruk op de onderbouwing van/argumentatie bij de gewenste handeling, terwijl *instrueren* vooral draait om het uitleggen/beschrijven van de gewenste handeling (Pander Maat, 2002). Op basis van deze verschillen wordt verwacht dat teksten met verschillende teksttypen te veel van elkaar verschillen om tot een adequate vergelijking te komen. Op basis van de overeenkomsten binnen teksttypen wordt verwacht dat teksten met gelijke teksttypen wel adequaat te vergelijken zijn. Dit resulteert in de volgende hypothesen:

H1: De beoordelaarsovereenstemming bij een middelmatige congruentie tussen ankerteksten en schrijfproducten is even hoog als bij een hoge congruentie.

H2: De beoordelaarsovereenstemming bij een lage congruentie tussen ankerteksten en schrijfproducten is lager dan bij een hoge congruentie.

Het onderzoek wordt uitgevoerd met twee verschillende beoordelingsschalen (beide met ankerteksten van het type *verzoeken*). Dankzij deze scheiding zijn de resultaten niet afhankelijk van slechts één beoordelingsschaal, en kunnen bovendien eventuele effecten van de schaalkeuze op de beoordeling gemeten worden. Gezien de overeenkomsten tussen de schalen (gelijke puntenverdeling, gelijk aantal ankerteksten, gelijke teksttypen van de ankerteksten) wordt verwacht dat de schaalkeuze geen invloed heeft op de beoordeling.

Methode

Proefpersonen

In het onderzoek fungeerden tien tweede- en derdejaars bachelorstudenten als beoordelaars: 1 man en 9 vrouwen tussen de 20 en de 26 jaar (M = 21.90, SD = 1.73). Wegens praktische overwegingen is ervoor gekozen studenten in te zetten als beoordelaars, en niet (basisschool)-docenten. Dit vormt geen grote beperking: uit onderzoek blijkt dat de overeenstemming tussen oordelen van leken (studenten) even hoog is als die tussen (taal)docenten (Schoonen, Vergeer & Eiting, 1997). Wel is gebleken dat studenten het lastiger vinden om de teksteigenschap 'taalgebruik' te beoordelen (Schoonen, 1991; Schoonen et al., 1997). Zodoende zijn voor het huidige onderzoek enkel studenten geselecteerd die een taalgerelateerde opleiding volgen.[1] De proefpersonen hadden verder geen ervaring met het beoordelen van teksten van basisschool-leerlingen.

Materiaal

Het materiaal is geselecteerd uit drie schrijfopdrachten die zijn uitgevoerd door leerlingen uit groep 8 van verschillende basisscholen. Bij alle opdrachten moesten de leerlingen een brief schrijven. Hier volgt een beknopte beschrijving van de opdrachten:

<u>Opdracht 'Smikkel'</u>

Het schrijven van een *verzoekende* brief aan het bedrijf Smikkel. Het doel was een gratis CD te bemachtigen die deel uitmaakte van een spaaractie, ondanks een tekort aan spaarpunten.

<u>Opdracht 'Smurfen'</u>

Het schrijven van een *verzoekende* brief aan het bedrijf SuperCoop. Het doel was om alsnog smurfenpoppetjes te ontvangen als onderdeel van een verzamelactie, ondanks het feit dat de poppetjes al op waren in de supermarkten.

<u>Opdracht 'Like'</u>

Het schrijven van een *instruerende* brief aan het meisje Like dat een jaar in Engeland had gewoond. Het doel was Like tips te geven

over het schrijven van een goede Nederlandse brief.

Voorafgaand aan het huidige onderzoek zijn bij de Smikkel- en de Smurfenopdracht beoordelingsschalen opgesteld (zie bijlage 1). Het opstellen van een dergelijke schaal gebeurde door eerst alle brieven globaal te laten beoordelen door drie onafhankelijke deskundigen. Hierbij kregen de teksten een puntenscore (100 punten voor een gemiddelde tekst). Vervolgens werden er vijf teksten geselecteerd met een hoge beoordelaars-overeenstemming, die representatief waren voor vijf schrijfniveaus (gemiddeld, -2 SD, -1 SD, +1 SD en +2 SD). Deze vijf representatieve teksten zijn de zogenoemde 'ankerteksten'. De ankerteksten werden op een beoordelingsschaal geplaatst, met een ankertekst op 70, 85, 100, 115 en 130 punten. Om de gebruiker van de schaal te helpen bij de interpretatie van de ankerteksten, is bij iedere schaal een toelichting opgesteld met daarin de belangrijkste plus- en minpunten per ankertekst (zie bijlage 2). In het huidige onderzoek werden zowel de schaal van de Smikkel- als van de Smurfenopdracht gebruikt, met hun bijbehorende toelichtingen.

Uit de brieven van de leerlingen zijn er in totaal 120 geselecteerd: 40 brieven per opdracht. Bij de selectie is gelet op de globale kwaliteit van de brieven, waarbij een realistische spreiding (volgens de normaalcurve) is bewerkstelligd over de verschillende niveaus op de schaal (veertien gemiddelde teksten, vijf teksten rond -1 SD, vijf teksten rond +1 SD, één tekst rond -2 SD en één tekst rond +2 SD). Dit om te voorkomen dat er door willekeurige selectie alleen hele slechte of hele goede teksten geselecteerd zouden worden, daar dit een ongewenste invloed zou kunnen hebben op de resultaten. De globale kwaliteit werd bepaald aan de hand van de gemiddelde beoordeling van de drie onafhankelijke deskundigen.

Procedure

De proefpersonen werden in twee groepen van vijf personen gesplitst en in twee aparte ruimtes geplaatst. Vervolgens kreeg de ene groep de schaal van de Smikkelopdracht (de 'Smikkelschaal') en de andere groep de schaal van de Smurfenopdracht (de 'Smurfenschaal'). Beide groepen kregen vervolgens alle 120 teksten, die ze dienden te beoordelen door ze te vergelijken met de betreffende schaal. In de groep met de Smikkelschaal werden eerst alle Smikkelteksten beoordeeld, vervolgens de Smurfenteksten en ten slotte de Like-teksten. In de groep met de Smurfenschaal werden eerst de Smurfenteksten, daarna de Smikkelteksten en tot slot de Like-teksten beoordeeld. In beide condities nam hierdoor de congruentie van de opdrachten met de schaal stapsgewijs af. Binnen de verzameling teksten van één opdracht (dus binnen één groep van 40 teksten) was de nakijkvolgorde voor iedere beoordelaar hetzelfde.

Voorafgaand aan de beoordeling kregen alle proefpersonen een schriftelijk instructie en een korte training. Deze waren voor beide condities gelijk. Tijdens de instructie werd hen de schaalbeoordelingsmethode uitgelegd en kregen ze enkele richtlijnen mee. Deze hielden o.a. in dat ze niet mochten overleggen, gegeven scores later niet meer mochten aanpassen er bij het beoordelen vooral op moesten letten of het tekstdoel (verzoeken of instrueren) bereikt werd.[2] Hierna kregen ze een training, waarbij enkele voorbeeldteksten beoordeeld moesten worden en er gelegenheid was om vragen te stellen.

De instructie en training duurden ongeveer 15 minuten. Hierna kregen de proefpersonen 90-120 minuten om na te kijken. Tot slot hebben alle proefpersonen een korte evaluatie ingevuld. Hierbij hebben ze o.a. per congruentieniveau (hoog/middelmatig/laag) aangegeven hoe gemakkelijk/moeilijk ze het vonden om de opdrachten te beoordelen met de betreffende schaal. Het antwoord werd genoteerd op een 5-puntsschaal die liep van 'zeer moeilijk' tot 'zeer gemakkelijk'. Deze vragen dienden ter ondersteuning van de beoordelingsresultaten door extra inzicht te verschaffen in de beoordelingsprocessen van de beoordelaars.

Data-analyse

In totaal leverde het onderzoek 120

tekstbeoordelingen op per beoordelaar, van in totaal tien beoordelaars: vijf beoordelaars per schaal. Per beoordelingsschaal is de Crombachs alpha over de vijf beoordelaars berekend. Hierbij werd de interbeoordelaars-betrouwbaarheid bepaald per verzameling teksten van één opdracht (40 teksten), waarna deze betrouwbaarheden onderling vergeleken zijn. Om te testen of de betrouwbaarheden onderling van elkaar verschilden zijn K-sample-significantietests uitgevoerd (Hakstian & Whalen, 1976; Feldt, 1980).

Daarnaast is bepaald in hoeverre de schaalkeuze invloed had op de beoordeling. Hiertoe is per schaal van iedere tekst het juryoordeel berekend, waarna de juryoordelen vergeleken zijn via een independent T-toets.

De antwoorden op de evaluatievragen zijn geanalyseerd via een one-way ANOVA, waarbij de drie congruentieniveaus zijn vergeleken. Er is hierbij geen onderscheid gemaakt tussen de twee condities. Naar aanleiding van de resultaten zijn twee post-hoctests (Bonferroni en Scheffe) uitgevoerd.

Resultaten

Smikkelschaal

De vijf proefpersonen in de Smikkelschaal-conditie beoordeelden eerst 40 Smikkelteksten (hoge congruentie), daarna 40 Smurfenteksten (middelmatige congruentie) en toen 40 Like-teksten (lage congruentie). Uit analyse van de data bleek dat de interbeoordelaars-betrouwbaarheid bij alle opdrachten zeer hoog was: bij de hoge congruentie α = .92, bij de middelmatige congruentie α = .90, en bij de lage congruentie α = .87. Om te bepalen of deze waarden significant van elkaar verschilden is een K-sample-significantietest uitgevoerd. Hieruit bleek dat de overeenstemming van de beoordelaars bij alle opdrachten even hoog was: de betrouwbaarheden bij de hoge en de middelmatige congruentie verschilden niet (F(1,39) = 1.21, p = .27), evenals de betrouwbaarheden bij de hoge en de lage congruentie (F(1,39) = 1.52, p = .10) en de betrouwbaarheden bij de middelmatige en de lage congruentie (F(1,39) = 1.25, p = .24).

Smurfenschaal

In de Smurfenschaalconditie werden eerst 40 Smurfenteksten (hoge congruentie), daarna 40 Smikkelteksten (middelmatige congruentie) en toen 40 Like-teksten (lage congruentie) beoordeeld. In deze conditie werden grotendeels gelijksoortige resultaten behaald. Uit de data-analyse bleek dat de interbeoordelaarsbetrouwbaarheid bij alle opdrachten zeer hoog was: bij de hoge congruentie α = .91, bij de middelmatige congruentie α = .92, en bij de lage congruentie α = .84. Wederom is een K-sample-significantietest uitgevoerd om te bepalen of deze waarden significant van elkaar verschilden. Hieruit bleek dat de betrouwbaarheden bij de hoge en de middelmatige congruentie even hoog waren (F(1,39) = 1.23, p = .26), evenals de betrouwbaarheden bij de hoge en de lage congruentie (F(1,39) = 1.64, p = .06). Opvallend is dat de betrouwbaarheden tussen de middelmatige en de lage congruentie wel significant verschilden; de interbeoordelaars-betrouwbaarheid bij de Smikkelteksten was hoger dan bij de Like-teksten (F(1,39) = 2.03, p = .02).

Wanneer de Smurfen- en Smikkeldata samengevoegd worden en als één groep worden vergeleken met de Like-data, blijkt de beoordelaarsovereenstemming bij de hoog-middelmatigcongruente teksten significant hoger te zijn dan bij de laagcongruente teksten (F(1,79) = 1.79, p < .01). De beoordelaars waren het dus is mindere mate eens over de laagcongruente schrijfproducten dan over de schrijfproducten die qua teksttype overeenkwamen met de ankerteksten.

Vergelijking van de schalen

In beide condities zijn dezelfde 120 schrijfproducten beoordeeld. Om een eventueel effect van de schaalkeuze op de beoordeling te detecteren, zijn de beoordelingen bij de verschillende schalen met elkaar vergeleken. In beide condities is per tekst het juryoordeel berekend, waarna deze gemiddelde oordelen

van de beide schalen met elkaar zijn vergeleken. Hieruit bleek dat de beoordelingen niet verschilden tussen de twee condities ($F(1,238) = .89$, $p = .98$). De schaalkeuze lijkt in het huidige onderzoek dus geen invloed te hebben op de beoordeling.

Evaluatievragen

Ter ondersteuning van de beoordelingsresultaten is de beoordelaars gevraagd om per congruentieniveau aan te geven hoe gemakkelijk/moeilijk het was om de schrijfproducten te beoordelen. Er werd een significant verschil gevonden tussen de evaluaties ($F(2,27) = 46.02$, $p < 0.01$). Uit beide post-hoctests bleek dat beoordelaars het moeilijker vonden om de laagcongruente teksten met de schaal te vergelijken dan de hoog- en middelcongruente teksten ($p < 0.01$). Bij de hoog- en middelcongruente teksten was de moeilijkheidsgraad van de beoordeling gelijk ($p = .40$ (Scheffe), resp. $p = .55$ (Bonferroni)).

Discussie

Het doel van het huidige onderzoek was om vast te stellen wat de invloed was van de congruentie tussen schrijfproducten en ankerteksten op de beoordelaarsovereenstemming. Verwacht werd dat bij een middelmatige congruentie (gelijk teksttype, afwijkende opdracht) de beoordelaarsovereenstemming even hoog zou zijn als bij een hoge congruentie. Daarnaast werd verwacht dat bij een lage congruentie (afwijkend teksttype, afwijkende opdracht) de beoordelaarsovereenstemming lager zou zijn dan bij een hoge congruentie.

De eerste hypothese wordt volledig ondersteund door de data: zowel in de Smikkel- als in de Smurfenschaalconditie was de beoordelaarsovereenstemming even hoog bij de teksten die middelmatig met de schaal congrueerden als bij de teksten die hoog met de schaal congrueerden. Bovendien was de betrouwbaarheid in alle gevallen zeer hoog ($\alpha \geq .90$). Dit impliceert dat het niet noodzakelijk is dat na te kijken schrijfproducten bij dezelfde opdracht horen als de ankerteksten, zolang de teksten behoren tot hetzelfde teksttype.

De resultaten zijn minder eenduidig ten aanzien van de tweede hypothese. Binnen de Smurfenschaalconditie wordt de hypothese ondersteund: in deze conditie werd een verschil gevonden in de beoordelaarsovereenstemming. Opvallend was echter dat dit verschil niet gemeten werd tussen de hoog en laag congruerende teksten (zoals je zou verwachten gezien de grotere tegenstelling tussen deze twee soorten teksten), maar tussen de middelmatig en laag congruerende teksten. Uit verdere analyse bleek echter dat wanneer hoog en middelmatig congruerende teksten samengevoegd werden, de beoordelaarsovereenstemming bij deze tekstgroep als geheel hoger was dan bij de laagcongruente teksten. Wanneer het teksttype van de schrijfproducten dus niet overeenkomt met het teksttype van de ankerteksten, lijken deze schrijfproducten niet met gelijke betrouwbaarheid beoordeeld te kunnen worden met de betreffende schaal als hoger congruerende opdrachten.

Dit zou verklaard kunnen worden door de structuurverschillen tussen de schrijfproducten en de ankerteksten ten gevolge van de verschillen in teksttype: wanneer de teksten te zeer verschillen kan het lastiger zijn voor de beoordelaars om ze te vergelijken. Hierdoor krijgen beoordelaars mogelijk meer ruimte om hun eigen invulling te geven aan de vergelijking tussen ankertekst en schrijfproduct, en kunnen persoonlijke opvattingen over de kenmerken van een goede tekst een grotere rol gaan spelen. Wanneer beoordelaars zich door verschillende opvattingen laten leiden, kan dit een lagere overeenstemming tot gevolg hebben (Wesdorp, 1981). De beoordelingsmethode gaat hiermee in feite meer lijken op globale beoordeling.

De resultaten binnen de Smikkelschaalconditie zijn daarentegen in strijd met de hypothese. In deze conditie werd namelijk geen enkel significant verschil gevonden in de beoordelaarsovereenstemming tussen laag congruerende teksten en de hoog en middelmatig congruerende teksten (hoewel de resultaten significantie benaderden). Dit suggereert dat ankerteksten met een afwijkend

teksttype wel genoeg ondersteuning bieden bij de beoordeling van schrijfproducten. Dat zou betekenen dat een beoordelingsschaal niet alleen gebruikt kan worden voor schrijfproducten met hetzelfde teksttype, maar ook voor schrijfproducten met een ander teksttype.

Belangrijk om op te merken is dat de beoordelaarsovereenstemming in de Smurfen-schaalconditie bij de laagcongruente teksten nog steeds zeer hoog was (α = .84), al was hij lager dan bij de andere congruentieniveaus. Men dient zich dus af te vragen welke mate van betrouwbaarheid in de onderwijspraktijk acceptabel is. Dit vraagstuk zal echter in het huidige onderzoek niet verder worden behandeld.

Implicaties voor het basisonderwijs

Op basis van de bovenstaande data kan geconcludeerd worden dat de mate waarin schaal en opdrachten met elkaar congrueren geen of weinig invloed heeft op de kwaliteit van de beoordeling: zowel bij een hoge als bij een lage congruentie is de interbeoordelaars-betrouwbaarheid zeer hoog. Voor het onderwijs is dit goed nieuws: via schaalbeoordeling kunnen schrijfproducten niet alleen verantwoord (resp. betrouwbaar) worden nagekeken, maar de methode behoudt enkele belangrijke voordelen van de globale beoordelingsmethode. Zo heeft een (enigszins geoefende) beoordelaar niet veel tijd nodig om tot een oordeel te komen, zeker niet wanneer hij vertrouwd is met de schaal (Pollmann et al., 2012). Daarnaast dwingt de constante vergelijking met de ankerteksten de beoordelaar om naar de tekst als geheel te kijken, en niet naar de afzonderlijke onderdelen. Hiermee wordt het validiteitsprobleem van analytische beoordeling (resp. dat een tekst niet gezien kan worden als een optelsom van zijn deelkenmerken) grotendeels ondervangen (Barkaoui, 2011).

Naast het feit dat het gebruik van een schaal betrouwbare oordelen oplevert, blijkt uit vergelijking van de oordelen in de verschillende condities eveneens dat de schaalkeuze geen invloed heeft op de beoordeling. Een zorgvuldig samengestelde schaal zou dus, onafhankelijk van de opdracht aan de hand waarvan hij geconstrueerd is, bruikbaar moeten zijn voor alle soorten opdrachten.

Alternatieve verklaringen

De bovenstaande implicaties klinken zeer hoopgevend voor de toekomst van schrijfvaardigheidseducatie. Wel zijn er enkele alternatieve verklaringen te bedenken voor de verrassend hoge interbeoordelaars-betrouwbaarheid bij de laagcongruente teksten. Er zijn twee alternatieve benaderingen van waaruit deze resultaten te verklaren zijn: de overeenkomsten tussen de beoordelaars, en de overeenkomsten tussen de verschillende schrijfopdrachten.

De overeenkomsten tussen de beoordelaars die deelnamen aan dit experiment waren erg groot: ze volgden allen een taalgerelateerde studie, waren ongeveer even oud en hadden allemaal geen ervaring met het nakijken van schrijfproducten. In een realistische beoordelingssituatie worden schrijfproducten daarentegen nagekeken door (basisschool)-docenten, die onderling zeer afwijkende achtergronden kunnen hebben (leeftijd, opleidingen, nakijkervaring) en daarmee zeer uiteenlopende opvattingen kunnen hebben over de eigenschappen van een goed schrijfproduct. Het is dus nog maar de vraag of de interbeoordelaarsbetrouwbaarheid in een realistische situatie even hoog zou zijn. Een factor die hierbij wellicht eveneens meespeelt is de training die de beoordelaars hebben gehad voorafgaand aan het experiment, waarbij de kernvoorwaarden van een goede tekst zijn beschreven en er geoefend is met de beoordeling van een aantal teksten. Hierdoor werd in hoge mate bepaald hoe de beoordelaars gingen nadenken over de teksten. Bovendien werden in beide condities de laagcongruente teksten als laatste nagekeken, direct nadat er 80 schrijfproducten waren beoordeeld die in hogere mate overeenkwamen met de schaal. Dit kan een sturende werking hebben gehad op de manier waarop de laagcongruente teksten beoordeeld werden, doordat de beoordelaars vooral gingen letten op de kenmerken die ook terug te vinden waren in de voorgaande

schrijfproducten, zoals briefconventies en structuur. Dit wijkt af van een natuurlijke nakijksituatie, waarin beoordelaars (resp. docenten) geen gelijkwaardige nakijkervaring hebben, en dus niet op een exact gelijke manier gestuurd worden in hun denkprocessen.

In eerste instantie gingen we ervan uit dat de hoogcongruente en de laagcongruente teksten sterk van elkaar verschilden op basis van het verschil in teksttype. Bij de teksttypen *verzoeken* en *instrueren* horen immers verschillende structuren (Pander Maat, 2002). De teksttypen lijken wellicht echter meer op elkaar dan in eerste instantie werd verwacht. Zo behandelt Pander Maat (2002) de teksttypen *verzoeken* en *instrueren* in dezelfde paragraaf, omdat beide typen betrekking hebben op de handelingen van de lezer. In beide gevallen probeert de schrijver de intenties van de lezer te beïnvloeden op een niet-dwingende wijze (Brinker, 1992). Hierbij wordt wel opgemerkt dat de *instructieven* verschillen van de *directieven* (resp. *verzoeken*) door hun specifieke context: "Die context kenmerkt zich door een wel-willen-maar-niet-kunnen: de lezer is in principe bereid de handeling uit te voeren, maar het ontbreekt hem aan de handelingskennis die daarvoor nodig is." (Pander Maat, 2002, 278). Dit heeft een duidelijk effect op de tekststructuur: bij *verzoeken* ligt de nadruk op de argumentatie achter het belang van de handeling, en bij *instrueren* op de uitleg van de handeling en de stappen hiervan.

Toch zijn er meer manieren om de classificering van teksten te benaderen. Zo is de term 'tekstgenre' op meerdere manieren op te vatten: als verwijzend naar het doel van de tekst, de vorm/het medium, het onderwerp of een combinatie van deze zaken (Pander Maat, 2002). Binnen dit kader vielen alle teksten in dit onderzoek binnen het genre 'brief', wat verwijst naar de vorm van de teksten. Dit heeft gevolgen voor de tekststructuur, daar alle teksten een conventionele opening, afsluiting en globale briefindeling dienden te hebben (Brinker, 1992). Omdat 'structuur' bij de training als een van de belangrijkste beoordelingscriteria werd genoemd, is het niet onwaarschijnlijk dat de overeenkomsten op dit punt tussen ankerteksten en laagcongruente teksten de beoordelaars geholpen hebben bij hun vergelijking.

Tot slot dient opgemerkt te worden dat de beoordeling van de laagcongruente teksten wellicht meer leek op globale beoordeling dan op schaalbeoordeling. Uit de evaluatie achteraf bleek dat beoordelaars het moeilijker vonden om de laagcongruente teksten met de schaal te vergelijken dan de hoog- en middelcongruente teksten. Hierdoor is het mogelijk dat de beoordelaars de laagcongruente teksten minder met de schaal hebben vergeleken, en in hogere mate over zijn gegaan op globale beoordeling. Dat de interbeoordelaarsbetrouwbaarheid hierbij desondanks vrij hoog was is te verklaren aan de hand van het teksttype van de laagcongruente teksten (*instructief*). De structuur van deze teksten was vrij simpel: naast een opening en een afsluiting bestonden de schrijfproducten voornamelijk uit een opsomming van adviezen. Hierdoor zijn beoordelaars mogelijkerwijs meer geneigd om te letten het aantal adviezen en de lengte van de tekst, en komt men zo tot een betrouwbaar oordeel. Tekstlengte correleert bij globale beoordeling immers vaak hoog met het toegekende cijfer (Huot, 1990; Lee et al., 2009). Dit biedt echter geen garantie dat teksttypen die minder geschikt zijn voor een dergelijke globale beoordelingswijze eveneens op betrouwbare wijze beoordeeld zouden kunnen worden bij lage congruentie met de schaal.

Beperkingen en aanbevelingen

Eerder hebben we aanbevelingen gedaan voor de toepassing van schaalbeoordeling in het basisonderwijs op basis van de behaalde resultaten. Of deze resultaten echter verantwoord te generaliseren zijn naar de praktijk blijft een lastige kwestie; het huidige onderzoek heeft een aantal beperkingen. Zo is een belangrijke tekortkoming in dit onderzoek de keuze van de beoordelaars: dit waren studenten, geen docenten. Zoals al eerder behandeld werd, schetsen de resultaten door de vele gelijkenissen tussen de beoordelaars wellicht een positiever beeld van de schaalbeoordelingsmethode dan in de praktijk zou blijken. Om vast te stellen of de methode

werkelijk tot een grote verbetering in de beoordelaarsovereenstemming zal leiden in het basisonderwijs, zal hij in de praktijk getest moeten worden onder docenten met zeer uiteenlopende achtergronden. Een belangrijke factor hierbij is de invloed van training, die mogelijkerwijs een noodzakelijke voorwaarde is voor het bereiken van een hoge beoordelaars-overeenstemming (Bacha, 2001). Er zal vastgesteld moeten worden hoeveel training met de schaal er nodig is om tot een hoge betrouwbaarheid te komen en of de toepassing van schaalbeoordeling hiermee nog steeds haalbaar is.

Een tweede beperking van het huidige onderzoek is het beperkte aantal geteste teksttypen: alleen *directieve* en *instructieve* teksten zijn aan bod gekomen. Op basis van de hiervoor besproken overeenkomsten tussen de gekozen teksttypen, is er nog geen garantie dat vergelijking van andere teksttypen, zoals *informeren* en *betogen*, eveneens tot een hoge beoordelaarsovereenstemming zou leiden. Voornamelijk de effecten van de schaalkeuze verdienen verdere aandacht. Het huidige onderzoek impliceert dat de schaalkeuze geen invloed heeft op de beoordeling, maar dit was enkel bij vergelijking van twee schalen met hetzelfde teksttype. Is het ook mogelijk om directieve teksten na te kijken met een instructieve schaal? Dergelijke vragen zullen in de toekomst gesteld moeten worden.

Algemene conclusie

Via het huidige onderzoek is een stap vooruit gemaakt in het verbeteren van de beoordeling van tekstkwaliteit. Er is vastgesteld dat een overeenkomst van 100% tussen schrijf-opdrachten en ankerteksten niet noodzakelijk is, en dat beoordelingsschalen zodoende op efficiënte wijze toegepast kunnen worden in het basisonderwijs. Welke beperkingen hieraan verbonden zijn dient verder onderzocht te worden. Er is echter goede hoop voor de toekomst: betrouwbare beoordeling van schrijfvaardigheid is geen onmogelijk streven meer.

Noten

[1] Te weten Nederlandse Taal en Cultuur, Communicatie- en Informatiewetenschappen of Taal en Cultuurstudies.

[2] Het vaststellen van enkele beoordelings-criteria is ook bij schaalbeoordeling belangrijk, om er zo voor te zorgen dat alle beoordelaars de schaal op gelijke wijze interpreteren (Bacha, 2001). Beoordelaars dienden vooral te beoordelen in welke mate het tekstdoel werd bereikt, door te letten op inhoud, duidelijkheid en structuur. Deze criteria zijn afgeleid uit de toelichting bij de schalen en sluiten zodoende aan bij de indeling van de ankerteksten.

Literatuur

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, *29*, 371–383.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*, 279-293.

Bergh, H. van den & Meuffels, B. (2000). Schrijfvaardigheden en schrijfprocessen. In Braet, A. (red.), *Taalbeheersing als communicatiewetenschap. Een overzicht van theorievorming, onderzoek en toepassingen.* Bussum: Coutinho.

Brinker, K. (1992). *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden.* Erich Schmidt Verlag: Berlin.

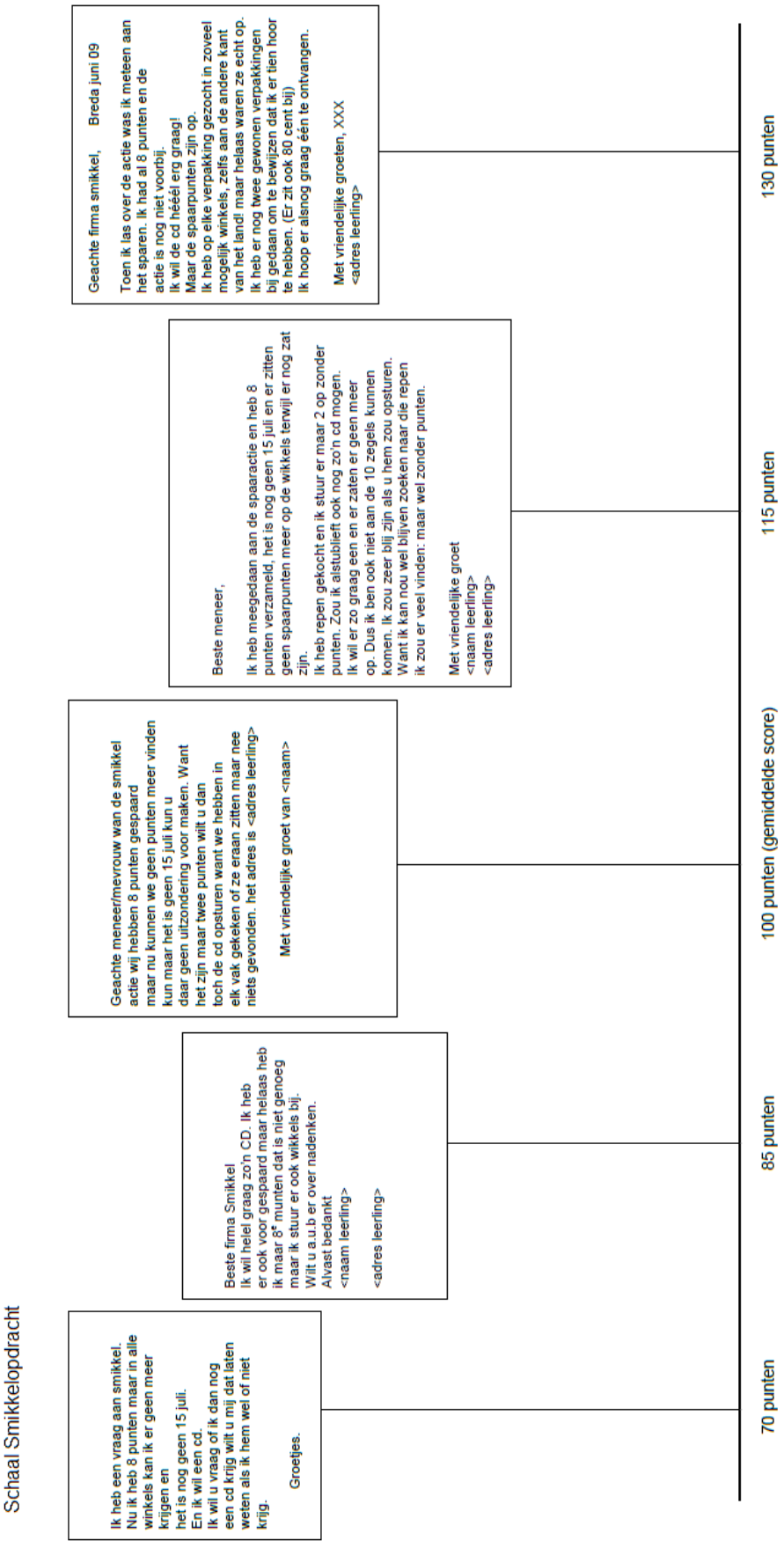
Feldt, L.S. (1980). A test of the hypothesis that crobach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*, *45*, 99-105.

Hajer, M., Meestringa, T., Leeuw, B. van der, Prenger, J., Glopper, K. de, & Dijk, G. van (2012). Genre, geletterdheid en vaktaalontwikkeling. In de Jong, N., Juffermans, K., Keijzer, M. & Rasier, L. (red.), *Papers of the Anéla 2012 Applied Linguistics Conference*. Delft: Eburon.

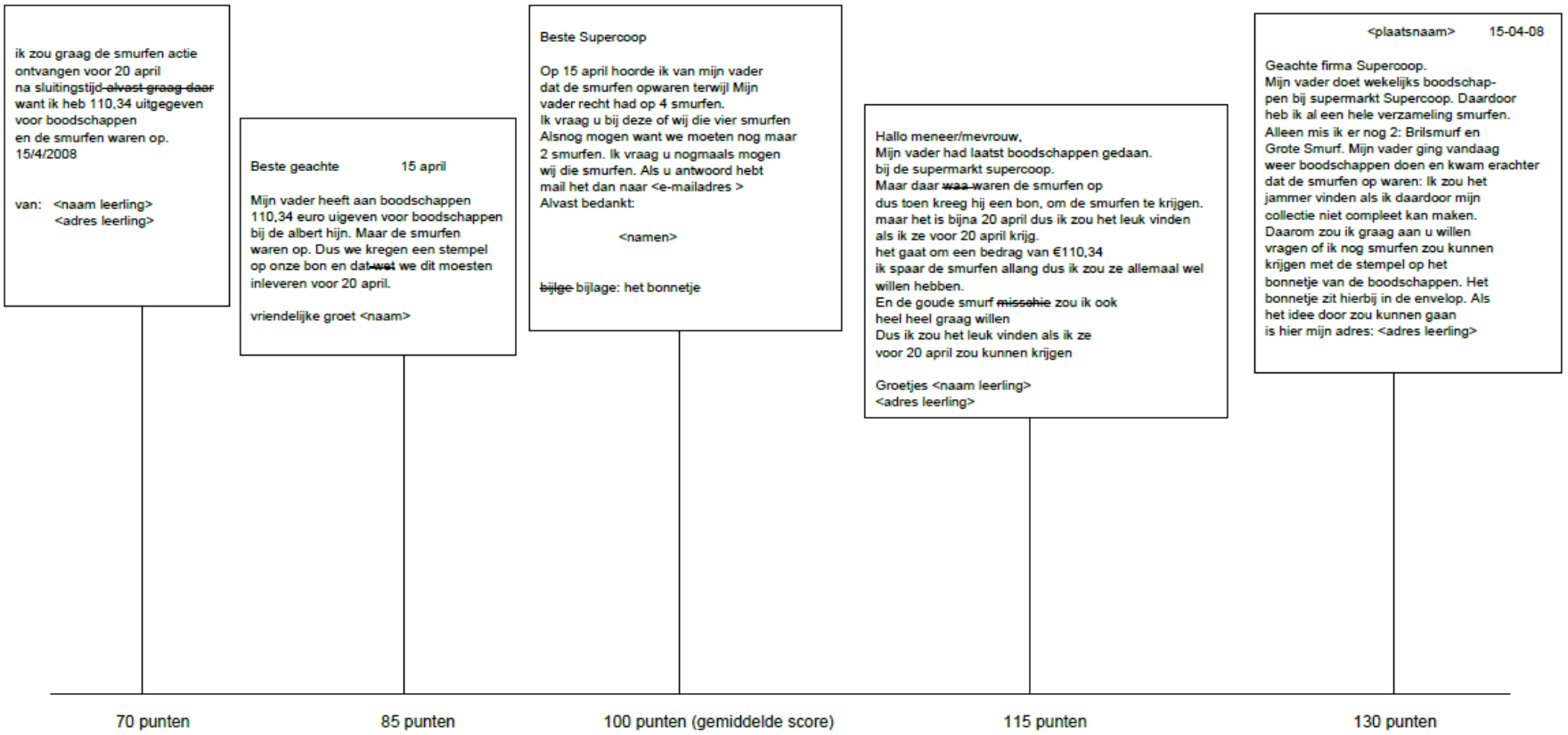
Hakstian, A.R. & Whalen, T.E. (1976). A K-sample significance test for independent alpha

coefficients. *Psychometrika*, *41*, 219-231.
Huot, B. (1990). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.
Inspectie van het onderwijs (2010). *Het onderwijs in het schrijven van teksten. De kwaliteit van schrijfonderwijs in het basisonderwijs*. Utrecht: Inspectie van het onderwijs.
Lee, Y., Gentile, C., & Kantor, R. (2009). Toward automated multi-trait scoring of essays: investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, *31*, 391–417.
Mullis, I. V. S. (1984). Scoring direct writing assessments: what are the alternatives? *Educational Measurement: Issues and Practice, 3*, 16–18.
Oudenhoven, J.P.L.M. van den (1983). *Onderwijsongelijkheid en evaluatieve feedback.* Apeldoorn: Van Walraven.
Pander Maat, H.L.W. (2002). *Tekstanalyse*. Bussum: Coutinho.
Pollmann, E., Prenger, J., & Glopper, C.M. de (2012). Het beoordelen van leerlingteksten met behulp van een schaalmodel. *Levende Talen Tijdschrift, 13*, 15-24.
Renkema, J. (1987). *Tekst en uitleg. Een inleiding in de tekstwetenschap*. Dordrecht: Foris.
Schoonen, R. (1991). *The evaluation of writing assessments. An empirical study into the reliability, validity and utility of writing assessments in final grade of elementary school.* Amsterdam: Universiteit van Amsterdam/SCO.
Schoonen, R., Vergeer, M. & Eiting, M. (1997). The assessment of writing ability. Expert readers versus lay readers. *Language Testing, 14*, 157-184
Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing, 22*, 1-30.
Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs. Een inventarisatie van beoordelingsmethoden voor de stelvaardigheid, het begrijpend lezen, de spreek-, luister- en discussievaardigheid. '*s-Gravenhage: Staatsuitgeverij.

Bijlage 1: Smikkel- en Smurfenschaal



Schaal Smurfenopdracht



Bijlage 2: Toelichting bij de Smikkel- en Smurfenschaal

Toelichting bij de ankerpunten (*Smikkelschaal*):

De gemiddelde ankertekst (100 punten) bevat de volgende plus- en minpunten:
Pluspunten:

- Het probleem wordt min of meer beschreven: de schrijver heeft niet genoeg smikkelpunten kunnen sparen, terwijl de actie nog niet is afgelopen.
- Er wordt een duidelijke vraag gesteld.
- Het adres van de afzender wordt vermeld.
- De brief bevat een conventionele aanhef en afsluiting.

Minpunten:

- Er wordt niet beschreven dat er ook twee losse wikkels zijn meegestuurd.
- De leerling gebruikt lange zinnen, die door de zwakke opbouw (grammaticaal) moeilijk te begrijpen zijn.
- Grammatica, spelling, interpunctie en opmaak zijn zwak.

De tekst met 115 punten is beter dan deze gemiddelde tekst omdat:

- het probleem helder is beschreven: schrijver heeft nog niet genoeg spaarpunten verzameld en er zijn geen punten meer in de winkel te krijgen, terwijl de actie nog loopt;
- het duidelijk is dat er daarom twee wikkels zonder punten zijn opgestuurd;
- er een duidelijke vraag is gesteld;
- de formele organisatie van de brief goed is: bevat goede aanhef + afsluiting, adres is vermeld;
- spelling en interpunctie prima zijn.

Een minpunt van de brief is dat de structuur niet zo goed is. Halverwege de brief valt de schrijver in herhaling en wordt de boodschap warriger, mede door een grammaticale ontsporing. Ook is de toon van het verzoek niet zo gepast.

De tekst met 130 punten is nog beter dan de tekst met 115 punten omdat:

- de brief heel overtuigend is: het probleem is adequaat beschreven en de vraag is duidelijk gesteld;
- de structuur goed is waardoor de inhoud begrijpelijk overkomt;
- de organisatie van de brief goed is: formele aanhef + afsluiting, adres is vermeld;
- de tekst grammaticaal goed is en qua spelling en interpunctie voldoende.

De tekst met 85 punten is slechter dan de gemiddelde tekst omdat:

- het probleem onvoldoende duidelijk wordt: niet duidelijk dat de actie nog loopt en er geen wikkels meer zijn en dat de schrijver daarom te weinig punten heeft;
- de vraag niet heel duidelijk is gesteld en de lezer zo niet wordt overtuigd om de cd ook echt op te sturen;
- de brief grammaticale fouten bevat (bv. maar acht*ste* punten) en ook qua spelling en interpunctie niet foutloos is.

De tekst met 70 punten is nog slechter dan de tekst met 85 punten omdat:

- het probleem onvoldoende duidelijk wordt: niet duidelijk dat het om een spaaractie gaat en dat er 2 wikkels zonder spaarpunten zijn meegestuurd;
- er geen adres is opgegeven, zodat de cd niet opgestuurd kan worden;
- de structuur niet goed is: informatie is niet in een logische volgorde weergegeven;
- het niet voldoet aan de briefconventies: er is geen aanhef en geen ondertekening;
- de brief grammaticaal en qua interpunctie niet in orde is (de spelling is wel ok).

Toelichting bij de ankerpunten (*Smurfenschaal*):

De gemiddelde ankertekst (100 punten) bevat de volgende plus- en minpunten:

- Het probleem is beschreven, maar dat had nog helderder gekund: waarom had de vader recht op de smurfen?
- Er wordt een duidelijke vraag gesteld, maar deze is wel twee keer genoemd – waarom? Dit maakt de brief niet overtuigender.
- De brief bevat een conventionele aanhef en afsluiting.
- De brief is gericht aan de lezer: 'ik vraag u …'
- De zinnen lopen grammaticaal goed, het taalgebruik is formeel en er zitten geen spellingsfouten in.

Minpunten:

- Inhoudelijk bevat de brief nog te weinig informatie om echt overtuigend te zijn.
- Het adres is niet genoemd (alleen een e-mailadres), waardoor de smurfen niet daadwerkelijk toegestuurd kunnen worden.
- Het hoofdlettergebruik is niet in orde.

De tekst met 115 punten is beter dan deze gemiddelde tekst omdat:

- er meer informatie is gegeven: zo is het bedrag expliciet beschreven, wordt de einddatum van de actie genoemd en betrekt de briefschrijver emoties in de brief om de vraag kracht bij te zetten (doet allang mee met de actie, zou heel heel graag de gouden smurf willen hebben etc.). Het probleem wordt hierdoor duidelijker en de brief in zijn geheel overtuigender;
- de briefconventies voldoende zijn: er is een aanhef en een afsluiting. Dit had wel formeler gekund;
- de organisatie van de inhoud ok is: eerst wordt het probleem beschreven, dan de vraag gesteld met de onderbouwing waarom dit zo belangrijk is.

Een minpunt van de brief is dat het taalgebruik niet zo gevarieerd is, er worden zelfs hele zinnen herhaald ("dus ik zou het leuk vinden als ik voor …"). Ook zijn de zinnen erg kort en niet altijd logisch afgebroken.

De tekst met 130 punten is nog beter dan de tekst met 115 punten omdat:

- de brief overtuigend is: het probleem is heel helder beschreven, de lezer geeft veel extra relevante informatie, de vraag is heel duidelijk gesteld;
- de structuur prima is: de zinnen lopen goed en zijn logisch opgebouwd, er wordt goed gebruik gemaakt van voegwoorden. Er missen wel alinea's;
- de briefconventies in orde zijn: er is een formele aanhef, de datum en het adres is genoemd. Er mist wel een afsluiting, waardoor de brief niet helemaal af lijkt;
- het taalgebruik in de brief is gevarieerd en de stijl is formeel. Er zijn geen fouten in de grammatica, spelling en/of interpunctie.

De tekst met 85 punten is slechter dan de gemiddelde tekst omdat:

- het probleem is wel genoemd, maar er is geen duidelijke vraag gesteld over het opsturen van de smurfen. Ook ontbreekt overtuigende informatie. Het is dus maar de vraag of de smurfen worden toegestuurd en al helemaal of dit voor 20 april zal zijn;
- de aanhef niet goed is: "beste geachte";
- spelling en hoofdlettergebruik niet helemaal goed gaat.

De tekst met 70 punten is nog slechter dan de tekst met 85 punten omdat:

- het inhoudelijk onder de maat is: er wordt te weinig en foutieve/onduidelijke informatie gegeven: bv. "smurfenactie", "na sluitingstijd". De brief is dus niet duidelijk;
- de brief structuur mist. Het is maar 1 zin en de informatie is daarbinnen ook niet logisch georganiseerd;
- het niet voldoet aan de briefconventies: er is geen aanhef en geen ondertekening. Ook staat de datum op de verkeerde plek;
- er geen gebruik is gemaakt van hoofdletters aan het begin van de zin. Er zijn geen fouten in de grammatica of spelling gemaakt.