Schaalbeoordeling: een hoopvolle methode?

Een onderzoek naar de jurybetrouwbaarheid met een beoordelingsschaal

*Joëlle Bemelman*

Faculteit Geesteswetenschappen, Universiteit Utrecht

Gepubliceerd op 24 juni 2013

**Samenvatting**

Er zijn al decennia lang problemen met het beoordelen van schrijfvaardigheid, omdat beoordelaars het zelden onderling eens zijn over de kwaliteit van een tekst. Er bestaan meerdere methodes om schrijfproducten te beoordelen, waarvan er een in het huidige onderzoek centraal staat: de schaalbeoordeling. De schaalbeoordeling lijkt een valide en betrouwbare methode te zijn, echter een groot nadeel is dat het maken van een dergelijke schaal een arbeidsintensieve klus is. In het huidige onderzoek wordt onderzocht welk effect de congruentie tussen ankerteksten op een schaal en schrijfopdrachten heeft op de beoordelaarsovereenstemming. Het doel van het onderzoek is dus om kijken of er voor meerdere opdrachten een schaal gebruikt kan worden. Het experiment is uitgevoerd door een jury van vijf beoordelaars, ze moesten teksten beoordelen van drie verschillende schrijfopdrachten met een beoordelingsschaal. De beoordelingsschaal is gebaseerd op een van de drie schrijfopdrachten. Een van de andere opdrachten verschilde alleen van onderwerp van de opdracht van de beoordelingsschaal en de laatste opdracht verschilde van onderwerp en communicatief doel. Uit de resultaten blijkt dat de beoordelaarsovereenstemming bij alle drie de schrijfopdrachten even hoog is. Hieruit kan geconcludeerd worden dat een beoordelingsschaal voor meerdere schrijfopdrachten gebruikt kan worden, omdat de schaal voldoende houvast geeft voor de beoordelaars. Dit impliceert dat een beoordelingsschaal in het onderwijs gebruikt kan worden om de schrijfvaardigheid voor leerlingen te beoordelen, omdat er niet voor elke schrijfopdracht een nieuwe schaal ontwikkeld hoeft te worden.

De beoordeling van schrijfvaardigheid staat al decennia lang in de aandacht binnen de wetenschap. Uit veel onderzoek blijkt dat het beoordelen van schrijfvaardigheid een probleem vormt (Huot, 1990; Schoonen, 2005; Van den Bergh & Meuffels, 2000). Schrijfvaardigheid is niet direct observeerbaar en dus ook niet direct meetbaar (Van den Bergh & Meuffels, 2000). Dit is waarschijnlijk dan ook de reden dat er geen specifieke richtlijn of een objectieve maatstaf bestaat over hoe schrijfvaardigheid beoordeeld moet worden. Het probleem met het beoordelen van schrijfvaardigheid is dus een ‘meetprobleem’ (Huot, 1990; Van den Bergh & Meuffels, 2000).

Ondanks dat schrijfvaardigheid niet direct meetbaar is, bestaan er toch twee verschillende soorten methoden. Volgens Cooper (1984) zijn dit directe en indirecte methoden. Bij directe methoden schrijven leerlingen teksten en bij indirecte methoden beantwoorden leerlingen meerkeuzevragen over teksten (Cooper, 1984). Directe methoden zijn beter dan indirecte methoden, omdat ze meer constructvalide zijn (Wesdorp, 1981). Dit houdt in dat directe methoden de schrijfvaardigheid in hogere mate meten dan indirecte methoden, omdat leerlingen bij directe methoden daadwerkelijk schrijven. Bij indirecte methoden wordt de werkelijke prestatie niet gemeten.

Toch is er ook een groot probleem met directe methoden: de beoordelaarsbetrouwbaarheid is laag (Van den Bergh & Meuffels, 2000; Schoonen, 2005). Verschillende beoordelaars zijn het vaak niet eens over de kwaliteit van schrijfproducten. Zo kan één schrijfproduct door verschillende beoordelaars heel gevariëerd beoordeeld worden. Ook zijn beoordelaars het vaak niet met zichzelf eens op verschillende beoordelingsmomenten (Schoonen, 2005).

Er zijn meerdere mogelijke oplossingen voor de beoordelaarsproblematiek. Eén hiervan is de globale beoordeling (Wesdorp, 1981; Schoonen & De Glopper, 1992). De kwaliteit van een schrijfproduct wordt dan beoordeeld door er een impressie-oordeel over te geven. De beoordelaar geeft een globaal oordeel op basis van alle relevante aspecten uit de tekst. Een voordeel van deze methode is dat deze snel en eenvoudig te hanteren is. Echter, een van de nadelen is dat het onduidelijk is hoe de verschillende aspecten gewogen worden, waardoor leerlingen geen specifieke feedback krijgen (Schoonen & De Glopper, 1992). Een ander nadeel is dat de methode zwak is als deze door één docent is toegepast. Hierdoor kan een *signifisch effect* optreden, waarbij verschillende beoordelaars verschillende accenten leggen zodat de beoordeling minder eerlijk wordt (Schoonen & De Glopper, 1992). Verder kan er sprake zijn van een *persoonlijke tendens* bij de beoordelaars. Zij kunnen over het algemeen soepel of juist streng beoordelen. Hierdoor kan er een groot verschil ontstaan tussen de resultaten van verschillende beoordelaars (Schoonen & De Glopper, 1992). Een ander effect dat kan optreden bij een globale beoordeling van slechts één beoordelaar is het *contaminatie-effect* (Wesdorp, 1981). Hiermee wordt bedoeld dat de beoordelaar de leerling kent en daardoor wordt ‘besmet’. Een voorbeeld hiervan is als een beoordelaar een leerling ‘straf’ wil geven omdat hij in de les vaak niet oplet. Hierdoor krijgt deze leerling een lagere beoordeling dan een leerling die altijd oplet, terwijl de schrijfproducten van hetzelfde niveau zijn / kwalitatief gelijk. Daarnaast kan er sprake zijn van het *sequentie-effect* (Wesdorp, 1981). Dit houdt in dat de volgorde waarin de schrijfproducten beoordeeld worden mede de hoogte van het oordeel bepaalt. Een te beoordelen product wordt dan vergeleken met een voorgaand product (Schoonen & De Glopper, 1992). Zo kan een product bijvoorbeeld lager gewaardeerd worden wanneer voorgaande producten een hoge kwaliteit hadden dan wanneer deze een lage kwaliteit hadden. Echter, wanneer de globale oordelen van meerdere beoordelaars (een jury) verzameld en gemiddeld worden, kan een groot deel van de bovenstaande negatieve effecten gereduceerd worden. Ook kan er door middel van een jury van beoordelaars een acceptabele betrouwbaarheid bereikt worden (die bij een enkele beoordelaar erg laag is). Dit houdt in dat meerdere jury-oordelen hoog correleren met elkaar over hetzelfde schrijfproduct. Een nadeel van een jury-oordeel is dat het volgens sommige critici wel betrouwbaar maar niet valide is (Wesdorp, 1981). Dit komt doordat de verschillende beoordelaars hun oordeel op verschillende kenmerken? uit de schrijfopdrachten baseren. Beoordelaar X kan bijvoorbeeld meer gericht zijn op stijl terwijl beoordelaar Y meer waarde hecht aan argumentatie. Toch kunnen zij op basis van deze verschillende aspecten tot eenzelfde globaal oordeel komen

Een andere beoordelingsmethode is de primary trait scoring, waarbij de beoordelaar zich richt op het hoofdkenmerk van de tekst (Wesdorp, 1981). In dit geval vindt er een soort globale beoordeling plaats op dit hoofdkenmerk. Wat dit hoofdkenmerk is, hangt af van het doel van de tekst. Deze methode kan gezien worden als een vorm van de globale beoordeling, waardoor het voordeel is dat het snel kan worden toegepast. Echter, deze methode is net zoals de globale beoordeling minder informatief voor de leerlingen: bij bijvoorbeeld een geslaagd betoog krijgt de leerling een hoog cijfer, maar kunnen er veel spelfouten in het opstel zitten. Deze spelfouten worden niet meegerekend, terwijl dit wel als een deel van schrijfvaardigheid gezien kan worden (Schoonen & de Glopper, 1992). Maar is deze methode wel mogelijk? Kan een beoordelaar wel uitsluitend naar een enkel hoofdkenmerk kijken? Uit een onderzoek van Jansen (2010) blijkt dat beoordelaars teksten slechter waarderen als er spelfouten in staan. Het kan echter ook dat beoordelaars de spelfouten onbewust meetellen in hun beoordeling (waardoor deze lager wordt), maar zich naar eigen zeggen op bijvoorbeeld structuur richten. Op dat moment is er al geen sprake meer van een beoordeling op enkel een hoofdkenmerk en kan het *halo-effect* optreden. Dit effect houdt in dat een bepaald tekstkenmerk een dermate grote ‘uitstraling’ heeft dat de beoordeling van andere tekstkenmerken erdoor beïnvloed wordt, zoals dat bij spelfouten te zien is (Schoonen & De Glopper, 1992). Omdat deze methode een vorm van de globale methode is, betekent het dat de betrouwbaarheid en de validiteit van deze methode ook discutabel zijn. Daarnaast kunnen dezelfde effecten optreden als bij de globale beoordeling, al heeft de beoordelaar bij deze methode al iets minder vrijheid dan bij de globale beoordeling (Wesdorp, 1981).

Daarnaast is er de analytische beoordeling. Hierbij geeft de beoordelaar een oordeel per aspect in een schema. In dit schema staat bijvoorbeeld Stijl, Inhoud en Opbouw (Schoonen & De Glopper, 1992). Sommige analytische schema’s zijn erg uitgebreid (rond de twintig categorieën), maar de meeste schema’s bevatten minder dan tien categorieën (Wesdorp, 1981). Het is de bedoeling dat de beoordelaars in het schema aankruisen waar het te beoordelen schrijfproduct aan voldoet. Van den Bergh en Meuffels (2000) nemen aan dat hoe meer kenmerken van het schema in de tekst voorkomen, hoe beter de kwaliteit van de tekst is. Uitleggen adv voorbeelden. De beoordelaars worden met de analytische beoordeling dus erg beperkt in hun beoordelingsvrijheid (in vergelijking met de globale beoordeling). Hierdoor is te verwachten dat de analytische beoordeling betrouwbaarder is. Dit is ook het geval, de beoordelaarsovereenstemming is meestal iets hoger voor analytici dan voor globale beoordelaars (Wesdorp, 1981) Hier zijn de meningen over verdeeld!!! \*er is geen concensus. Een nadeel van deze methode is dat de beoordeling een veeleisende taak is voor de beoordelaars, omdat de instructies die erbij horen vaak intensief en gedetailleerd zijn. De beoordelaars moeten dus zeer geconcentreerd te werk gaan. Een ander nadeel is dat deze methode veel tijd kost. Ook over de validiteit valt iets te zeggen, en dan met name de constructvaliditeit, want de vele categorieën kunnen niet onafhankelijk worden beoordeeld. Bij deze methode wordt het signifisch effect beperkt, omdat alle te beoordelen categorieën van tevoren vaststaan waardoor de beoordelaars geen verschillende accenten kunnen leggen. Wel kan het halo-effect optreden bij de analytische methode (Wesdorp, 1981). Dat betekent dat de validiteit van deze methode niet optimaal is, omdat het construct wat men wil meten (schrijfvaardigheid) niet gemeten wordt (door onder andere het halo-effect).

Een laatste methode is de schaalbeoordeling, waarbij een beoordelaar de schrijfproducten beoordeel**d** aan de hand van voorbeeldopstellen, die in een beoordelingsschaal staan. Een dergelijke schaal bestaat uit een aantal (meestal vijf) voorbeeldopstellen die in kwaliteit oplopen. Aan elk voorbeeldopstel is een score toegekend en (**…)?** fungeert als referentiepunt voor de beoordelaar (Pollmann, Prenger & De Glopper, 2012). De beoordelaars komen tot een oordeel over de te beoordelen opstellen met behulp van deze referentiepunten (ankerteksten) uit de beoordelingsschaal, ook wel ankerteksten genoemd

Een voordeel van de schaalbeoordeling is dat de mate van overeenstemming en de betrouwbaarheid van de oordelen van de beoordelaars hoog zijn, want de beoordelingsschaal geeft hen houvast (Pollmann et al., 2012). De beoordelaars vergelijken de te beoordelen opstellen dus met de ankerteksten uit de beoordelingsschaal, waardoor de oordelen van de verschillende beoordelaars overeenkomen (oh? Altijd?). Een ander voordeel is dat de beoordelaars het werken met een beoordelingsschaal prettig vinden (in het onderzoek van .. vinden ze het prettig) en dat het relatief weinig tijd kost (Pollmann et al., 2012). Daarnaast is er geen sprake van het sequentie-effect. Ook treedt er geen *normverschuiving* op. Dit betekent dat beoordelaars zich onbewust aanpassen aan het niveau van de tekst. Het uitblijven van het sequentie-effect en een normverschuiving wordt veroorzaakt doordat beoordelaars de schrijfproducten constant met de schaal vergelijken en niet de schrijfproducten onderling. Het signifisch effect wordt ook voorkomen, omdat er vaste criteria bij de schaal zijn ontwikkeld waaraan de schrijfproducten moeten voldoen.

Een kanttekening die geplaatst kan worden bij al deze voordelen is dat het ontwikkelen van een beoordelingsschaal een arbeidsintensieve klus is (Pollmann et al., 2012). Daarnaast is het van belang om te weten of teksten met andere onderwerpen en genres vergeleken kunnen worden met één beoordelingsschaal, omdat er dan minder beoordelingsschalen nodig zijn. Het is echter (nog) niet duidelijk of dit mogelijk is, want er is in de literatuur weinig tot niets over te vinden. De schaalbeoordeling vormt dus zowel een praktisch als een theoretisch probleem. Om duidelijkheid te krijgen over of deze problemen op te lossen zijn wordt in het huidige onderzoek gekeken of het mogelijk is om verschillende teksten (met verschillende genres) te vergelijken met één beoordelingsschaal. De term ‘genre’ moet hiervoor gedefinieerd worden.

Volgens Schellens en Steehouder (2008) wordt een tekstgenre gevormd doordat teksten van dezelfde soort overeenkomen met betrekking tot de conventionele functie, inhoud, structuur, stijl en vormgeving. Het communicatieve doel van een tekst is erg belangrijk binnen de genrecontext, omdat het meestal de andere kenmerken van een tekst beïnvloedt. Het doel van een genre ontstaat vanuit een communicatieve behoefte in een bepaalde situatie. Schellens en Steehouder (2008) claimen “dat het belangrijk is om het tekstdoel te achterhalen om te kunnen bepalen tot welk genre de tekst eigenlijk behoort”.

Uit de bovenstaande literatuur vloeit de volgende onderzoeksvraag voort: *Wat is het effect van de congruentie van een beoordelingsschaal en schrijfopdrachten op de beoordelaarsovereenstemming?* Deze vraag is relevant omdat we op deze manier kunnen kijken of er voor meerdere schrijfopdrachten één beoordelingsschaal kan worden gebruikt, waardoor er beduidend minder beoordelingsschalen gecreëerd hoeven te worden. Hiermee zou dus zowel het bovenstaande praktische als theoretische probleem ondervangen worden.

De afhankelijke variabele in het huidige onderzoek is de beoordelaarsovereenstemming en de onafhankelijke variabele is de congruentie tussen de ankerteksten in de schaal en de te beoordelen teksten. In dit onderzoek wordt gewerkt met een schaal die gemaakt is voor een overtuigende tekst. Er worden naast de tekst die oorspronkelijk bij de schaal hoort ook nog twee andere teksten vergeleken: een andere overtuigende tekst en een instruerende tekst. Er wordt dus in totaal met drie verschillende schrijfopdrachten gewerkt. Met behulp van een jury wordt er gekeken wat de beoordelaarsovereenstemming is van elk van deze drie schrijfopdrachten in vergelijking met de schaal. Er is voor deze operationalisatie gekozen omdat op deze manier duidelijk wordt of het beoordelen van verschillende schrijfopdrachten met behulp van een enkele beoordelingsschaal mogelijk is.

Uit de literatuur (Pollmann et al., 2012) blijkt dat een beoordelingsschaal werkt met de bijbehorende schrijfopdracht. Om deze reden is de verwachting dat de beoordelaarsovereenstemming bij de schrijfopdracht die bij de schaal hoort hoger zal zijn dan de opdrachten die van onderwerp (en communicatief doel/genre) verschillen.

**Methode**

Proefpersonen

Aan dit onderzoek namen vijf proefpersonen deel. Hier is voor gekozen omdat er een jury van vijf beoordelaars nodig was, omdat blijkt dat een jury van vijf beoordelaars ‘ideaal’ is. De proefpersonen in deze jury waren allemaal tweede- of derdejaars studenten van een talige/communicatieve opleiding aan de Universiteit Utrecht. Er is bewust gekozen voor studenten met een taal/communicatieopleiding omdat zij hierdoor voorkennis hebben over taal en communicatie. De verwachting is daarom dat deze proefpersonen de kwaliteit van teksten goed in kunnen schatten. Ook kunnen op deze manier eventuele opleidingsverschillen geminimaliseerd worden. De gemiddelde leeftijd van de proefpersonen is 21.8 jaar. Alle vijf de proefpersonen waren vrouwen. Hiervan waren 4 proefpersonen derdejaars studenten en van één proefpersoon is dit niet bekend. Er is van tevoren aan de proefpersonen gevraagd of ze ervaring hadden met het beoordelen van teksten en dit bleek voor twee van de vijf proefpersonen het geval te zijn.

Materiaal

In dit onderzoek is er gebruik gemaakt van een beoordelingsschaal, de ‘Smikkelschaal’. Deze schaal bestaat uit vijf ankerteksten waar een puntenaantal aan is toegekend die varieert van 70 (zeer slechte tekst) tot 130 (zeer goede tekst). Tussen elke ankertekst zit een standaarddeviatie verschil, dit is te zien in tabel 1.

Tabel 1: scores ankerteksten met standaarddeviaties

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ankertekst | Zeer slecht | Slecht | Gemiddeld | Goed | Zeer goed |
| Standaarddeviatie | -2 SD | -1 SD | M | +1 SD | +2 SD |
| Score | 70 | 85 | 100 | 115 | 130 |

De verschillende ankerteksten, die in de beoordelingsschaal zijn weergegeven, zijn uit een grote steekproef geselecteerd op basis van de scores van 3 beoordelaars. De geselecteerde teksten zijn in de beoordelingsschalen opgenomen en functioneren dus als ‘hulpteksten’ voor de beoordelaars.

Er is een selectie gemaakt van 3 verschillende schrijfopdrachten: Smikkelteksten, Smurfenteksten en Like-teksten. Deze schrijfopdrachten zijn door leerlingen uit groep 8 van verschillende basisscholen uitgevoerd. Om een beeld te geven van de opdracht die de leerlingen kregen voor de tekst worden hieronder in het kort de opdrachten weergegeven:

* *Smikkelopdracht*: Op elke Smikkel-reep zit 1 spaarpunt en je hebt 10 punten nodig voor een gratis CD, maar er zitten geen spaarpunten meer op de wikkels (terwijl de actie nog loopt) en je hebt maar 8 punten gespaard. Stuur een briefje naar de firma SMIKKEL met de 8 punten en 2 wikkels erbij met de vraag of je alsnog de gratis CD mag.
* *Smurfenopdracht*: Bij elke 25 euro die je vader besteedt bij SuperCoop krijg je een smurf. Maar op 15 april zijn de smurfen op (terwijl de actie tot 20 april loopt) en je hebt nog recht op een aantal smurfen. Schrijf een briefje met de kassabon en de vraag of je de smurfen alsnog krijgt.
* *Like-opdracht*: Er komt een nieuwe leerling in de klas die niet weet hoe je in het Nederlands een goede tekst schrijft, hij komt namelijk uit Engeland. Schrijf een brief met tips en adviezen aan Like waarin je uitlegt hoe je in het Nederlands een goede tekst schrijft.

De teksten die door bovenstaande opdrachten zijn geschreven verschillen in hun communicatieve doel. Bij de Smikkelteksten en de Smurfenteksten is het doel overtuigen, terwijl het doel van de Like-teksten instrueren is. Het gaat bij elke tekst dus een stapje verder van de Smikkelschaal af, aangezien de Smurfenteksten qua onderwerp verschillen en de Like-teksten qua onderwerp en doel. De selectie van de teksten bestaat uit 40 teksten van elk tekstonderwerp. Er zijn van alle vijf de categorieën (van zeer slecht tot zeer goed) een aantal teksten geselecteerd, waarbij ongeveer twee derde van de rond het gemiddelde zit. De rest van de teksten zitten onder of boven het gemiddelde, waardoor de steekproef representatief is. Met behulp van de Smikkelschaal gaan vijf beoordelaars (de jury) dus in totaal 120 teksten per persoon beoordelen.

Procedure

*Instructie*

Het experiment is in een enkele sessie afgenomen, de beoordelaars (jury) zaten bij elkaar in een lokaal. Om de proefpersonen bekend te laten raken met de beoordelingsschaal hebben we eerst een training gedaan van ongeveer 20 minuten waarbij de proefpersonen drie Smikkel-voorbeeldteksten kregen die ze moesten beoordelen met een getal op basis van de Smikkelschaal. Ze krijgen een lijstje met criteria (de plus- en minpunten) die gelden voor elk van de vijf ankerteksten. Aan de hand van deze criteria kunnen de beoordelaars het verschil zien tussen de verschillende ankerteksten. In de bijlage zijn de instructies voor de beoordelaars van de Smikkelschaal te vinden.Verder werd uitgelegd dat de beoordelaars de beoordelingscriteria van de schaal moesten gebruiken om te zien waar ze op moesten letten bij het beoordelen van de teksten. We hebben de scores die de beoordelaars gaven aan de voorbeeldteksten ‘klassikaal’ besproken. Op deze manier konden ze alvast oefenen met het beoordelen en kregen ze een beter beeld van wat de bedoeling was. Ook werd duidelijk gemaakt dat de beoordelaars bij elke tekst moesten kijken naar de beoordelingsschaal en aan de hand daarvan een score toekennen. Daarnaast werd vermeld dat ze de teksten niet onderling met elkaar mochten vergelijken, omdat we anders geen goed beeld krijgen van de scores op basis van de beoordelingsschaal.

*Beoordelen*

Tijdens het beoordelen mochten de beoordelaars niet vooruit bladeren zonder een score te hebben gegeven aan de vorige tekst. Verder mochten ze alleen bij hoge uitzondering terugbladeren, omdat de kans anders te groot was dat ze de teksten met elkaar vergeleken. De proefpersonen mochten met het beoordelen een score geven van 0 tot oneindig. Ze kregen eerst de Smikkelteksten, daarna de Smurfenteksten en daarna de Like-teksten om te beoordelen. Binnen de verschillende tekstonderwerpen waren de teksten in een willekeurige niveau-volgorde gelegd, maar wel bij elke beoordelaar dezelfde willekeurige volgorde. Door iedere beoordelaar dezelfde volgorde te geven konden we de teksten op het scoreformulier wat ze kregen nummeren, van 1 tot 120. Ze kregen dit scoreformulier om per tekst de gegeven score op te schrijven. Op die manier hebben een duidelijk overzicht van de scores. De proefpersonen hadden in totaal 100 minuten om de scores van de 120 teksten op het scoreformulier in te vullen. Tussen de verschillende tekstbundels (van 40 teksten) mochten de beoordelaars eventueel een korte pauze houden van 5 minuten.

*Na afloop*

Na het beoordelen werd de proefpersonen gevraagd om een vragenlijst in te vullen. Hierin werden vragen gesteld over waar de beoordelaars vooral op hebben gelet bij het beoordelen en hoe gemakkelijk/moeilijk ze het vonden om de teksten te beoordelen. Ook werden er demografische gegevens gevraagd. Wanneer de proefpersonen al deze vragen hadden beantwoord en het vragenformulier en scoreformulier in hadden geleverd, was het experiment afgelopen.

Data-analyse

Het design van dit experiment bestaat uit een onafhankelijke variabele en een afhankelijke variabele. De onafhankelijke variabele is de congruentie tussen de beoordelingsschaal en de verschillende schrijfopdrachten. Er zijn drie niveaus te onderscheiden: hoog, middel en laag. Dit betekent dat de Smikkelteksten ‘hoog’ overeenkomen met de schaal, de Smurfenteksten komen ‘middel(matig)’ overeen en de Like-teksten komen ‘laag’ overeen met de schaal. De afhankelijke variabele is de beoordelaarsovereenstemming, ofwel de jurybetrouwbaarheid. We willen weten wat de beoordelaarsovereenstemming per niveau is, dus van hoog, middel en laag. Daarom wordt er gekeken naar de Crohnbach’s alpha, als deze rond de .7 of hoger is, betekent het dat de beoordelaarsovereenstemming redelijk/goed is. De Crohnbach’s alpha zal voor het hoge, het middel(matige) en het lage niveau berekend worden. Op deze manier kan de beoordelaarsovereenstemming van elk niveau met elkaar worden vergeleken en hierdoor wordt duidelijk of de verschillen significant zijn. Verder zal worden gekeken of er een beoordelaar is die sterk afwijkt van de rest, omdat deze beoordelaar dan eventueel uit de analyse gehaald kan worden als de overeenstemming minder betrouwbaar is door die beoordelaar.

**Resultaten**

Na een betrouwbaarheidsanalyse blijkt dat de gemiddelde beoordelaarsovereenstemming van de beoordelaars van de Smikkelschaal (*N*= 5) hoog is (*α* = .892). De resultaten laten zien dat geen van de beoordelaars afwijkt van de rest. Er hoeft dus geen beoordelaar uit de analyse gehaald te worden, omdat de Crohnbach’s alpha maximaal tot .894 kan stijgen. Dit verschil is zo klein in vergelijking met de gemiddelde overeenstemming (*α* = .892) dat er geen reden is om een beoordelaar uit de analyses te halen.

Uit de resultaten blijkt dat de overeenstemming van de beoordelaars bij de Smikkelteksten erg hoog is (*α* = .916). Verder blijkt dat de overeenstemming van de Smurfenteksten hoog is (α = .898). De beoordelaarsovereenstemming van de Like-teksten is ook behoorlijk hoog (*α* = .872).

Na het vergelijken van de verschillende teksttypen blijkt dat de overeenstemming tussen de beoordelaars bij de Smikkelteksten en de Smurfenteksten even hoog is, er is namelijk geen significant verschil (*F*(1, 39)= 1.21, *p* = .27). Daarnaast laten de resultaten zien dat de beoordelaarsovereenstemming bij de Smurfenteksten en de Like-teksten ook even hoog is, er is wederom geen sprake van een significant effect (*F*(1, 39)= 1.25, *p*= .24). Verder blijkt dat beoordelaarsovereenstemming tussen de Smikkelteksten en de Like-teksten ook geen significant verschil vertoont (*F*(1, 39)= 1.52, *p* = .09).

Na het berekenen van correlaties van de verschillende beoordelaars blijkt dat de gemiddelde correlatie .632 is. Daarvan is de hoogste correlatie .737 en de laagste correlatie .478.

Uit de evaluatievragen bleek dat vier van de vijf beoordelaars (80%) op het tekstdoel hadden gelet tijdens het beoordelen. Daarnaast kwam ‘structuur’ bij drie van de vijf beoordelaars (60%) terug. Dat betekent dat een groot deel van de beoordelaars (80% en 60%) beoordelaars op dezelfde elementen (structuur en tekstdoel) hebben gelet, wat bovenstaande resultaten versterkt.

Verder bleek uit de evaluatie dat de beoordelaars het beoordelen van de Like-teksten met de Smikkelschaal relatief moeilijk vonden (*M* = 1.4 met een standaarddeviatie van .548), waarbij ze op een vijf-punt Likertschaal aan moesten geven hoe gemakkelijk/moeilijk ze het beoordelen van de verschillende schrijfopdrachten met de Smikkelschaal vonden. Het gemiddelde van de beoordelaars bij het beoordelen van Smikkelteksten met de Smikkelschaal is 3.6 met een standaarddeviatie van 1.140. Het gemiddelde van de Smurfenteksten met de Smikkelschaal was ook 3.6 met een standaarddeviatie van .548.

**Discussie**

De vraag van het huidige onderzoek was: *Wat is het effect van de congruentie tussen ankerteksten op een beoordelingsschaal en schrijfopdrachten op de beoordelaarsovereenstemming?*  Om deze vraag te beantwoorden is het belangrijk om terug te koppelen naar de resultaten. Uit deze resultaten blijkt dat de gemiddelde beoordelaarsovereenstemming hoog is. Daarnaast bleek dat er geen significante verschillen zijn tussen de drie onderzochte teksttypen met de Smikkelschaal. Dit houdt dus in dat de beoordelaars het bij alle drie de teksten het evenveel met elkaar eens zijn. Voor de opgestelde hypothese betekent dit dat deze incorrect was, omdat er geen verschillen in de beoordelaarsovereenstemming van de drie teksttypen zijn gevonden.

De beoordelingsschaal is dus betrouwbaar en kan voor meerdere schrijfopdrachten gebruikt worden met een ander onderwerp en communicatief doel/genre. Dit houdt in dat beoordelingsschalen voor het onderwijs erg bruikbaar zijn om leerlingteksten te beoordelen. De resultaten zorgen ook voor een positieve kijk op het probleem de beoordeling van schrijfvaardigheid. De beoordelingsschaal lijkt namelijk een objectieve maatstaf te zijn om teksten te beoordelen en daarmee ook de schrijfvaardigheid van een leerling. De beoordelingsschaal is een betrouwbaar meetinstrument, omdat de ankerteksten als referentiepunten fungeren.

Hieruit kan geconcludeerd worden dat een beoordelingsschaal voldoende ondersteuning geeft voor beoordelaars. Dit komt overeen met de literatuur die in het theoretisch kader is genoemd (Pollmann et al., 2012). De ondersteuning die het geeft heeft betrekking op de duidelijkheid van de ankerteksten en de toelichting/criteria daarbij. De beoordelaars kijken naar de tekst als geheel, maar met bepaalde criteria. Ook is er sprake van een concreet voorbeeld (de ankerteksten) die de criteria ondersteunt waardoor de beoordelingsschaal blijkbaar genoeg houvast geeft.

Wat de bovenstaande resultaten impliceren is dat we een enkele beoordelingsschaal kunnen gebruiken voor verschillende schrijfopdrachten. Daarmee geldt het nadeel/probleem dat de inleiding werd genoemd (het ontwikkelen van een beoordelingsschaal is een arbeidsintensieve klus) (Pollmann et al, 2012) niet langer.

Er zijn echter meerdere alternatieve verklaringen voor de resultaten mogelijk. Allereerst kan er een effect zijn van ervaring. De beoordelaars beoordeelden de Like-teksten als laatste, ze hadden daarvoor dus al 80 andere teksten beoordeeld. Het is mogelijk dat ze de beoordelingsschaal hierdoor uit hun hoofd kenden waardoor het beoordelen makkelijker ging. Het is echter nog maar de vraag of de beoordelaars dezelfde overeenstemming zouden hebben als de Like-teksten als enige beoordeeld moesten worden met de Smikkelschaal zonder daarvoor andere teksten beoordeeld te hebben.

Een tweede verklaring van de resultaten kan het volgorde-effect zijn. Zouden de resultaten hetzelfde zijn als de teksttypen in een andere volgorde werden gegeven? Stel dat de beoordelaars met dezelfde Smikkelschaal eerst de Smurfenteksten, daarna de Like-teksten en daarna de Smikkelteksten moesten beoordelen. Het is de vraag of de beoordelaarsovereenstemming dan nog steeds zo hoog zal zijn.

Een derde verklaring kan zijn dat de genreverschillen eigenlijk geen genreverschillen zijn. In het theoretisch kader wordt door Schellens en Steehouder (2008) gezegd dat een “tekstgenre wordt gevormd door teksten van dezelfde soort die zich kenmerken door een overeenkomstige conventionele functie, inhoud, structuur, stijl en (soms) grafisch vormgeving”. Als we deze definitie aannemen is er geen verschil tussen de verschillende teksten die onderzocht zijn in dit onderzoek, ze vallen namelijk allemaal onder het genre ‘brief’ en hebben een vergelijkbare conventionele functie, structuur en stijl. Echter als we een iets explicietere definitie aannemen die zegt dat doel belangrijk voor het identificeren van een genre omdat het vaak de overige kenmerken beïnvloedt, verschillen de teksten wel van elkaar. Dit verschil heeft uiteraard alleen betrekking op de Smikkel- en de Smurfenteksten in vergelijking met de Like-teksten, omdat de Like-teksten een ander doel hebben (instructief) dan de andere twee teksten (overtuigend).

Om te zeggen dat de schaalbeoordeling de ‘ideale’ methode is voor het beoordelen van schrijfvaardigheid, is te ver vooruitgedacht. De reden hiervoor is dat er niet naar alle verschillende teksttypen gegeneraliseerd kan worden. Er zijn maar drie verschillende teksten onderzocht en het is de vraag of deze wel echt van elkaar verschillen (met betrekking tot genre). Voor vervolgonderzoek is het daarom verstandig om beoordelaars teksten met werkelijk verschillende genres te laten beoordelen met één schaal, zoals een sollicitatiebrief, een handleiding en een nieuwsbericht. Op deze manier kan gekeken worden of de verkregen resultaten van het huidige onderzoek houdbaar zijn.

Verder kan er ook niet gegeneraliseerd worden naar schrijfproducten in het algemeen, omdat er uitsluitend leerlingteksten van de bovenbouw van de basisschool zijn beoordeeld. Daarnaast hebben we ook maar een enkele schaal gebruikt. Om een goed beeld te kunnen krijgen of beoordelingsschalen in het algemeen voor meerdere schrijfopdrachten gebruikt kunnen worden zal er meer onderzoek gedaan moeten worden met verschillende schalen.

Daarnaast blijkt uit de evaluatievragen dat twee van de vijf beoordelaars ervaring had met beoordelen en drie hadden geen ervaring. De meeste leerkrachten hebben (nog) geen ervaring met beoordelingsschalen dus de proefpersonen uit het huidige onderzoek zijn niet erg representatief. Er is nog een reden waarom de proefpersonen niet representatief zijn, dat is dat ze allemaal taal- en/of communicatiekundige proefpersonen zijn. Leerkrachten in het basisonderwijs zijn dit (hoogstwaarschijnlijk) niet. In vervolgonderzoek is het daarom van belang om bijvoorbeeld leerkrachten aan het onderzoek deel te laten nemen, zodat de proefpersonen representatief zijn.

Ook kan er nog gekeken worden of er sprake is van een ervaringseffect of een volgorde-effect. Het is van belang om hetzelfde onderzoek nog een keer uit te voeren, maar dan met de te beoordelen teksten in een andere volgorde om een volgorde-effect uit te sluiten. Ook is het interessant om te kijken of dezelfde resultaten verkregen worden wanneer de beoordelaars niet eerst 80 teksten beoordelen, maar gelijk beginnen met een tekst die sterk afwijkt van de beoordelingsschaal.

Wat uit dit onderzoek geconcludeerd kan worden is dat de resultaten hoopvol zijn voor de praktijk en dan in het bijzonder het onderwijs, want er is een hoge beoordelaarsovereenstemming bij verschillende schrijfopdrachten. Een beoordelingsschaal geeft dus voldoende houvast voor de beoordelaars en kan om deze reden voor meerdere schrijfopdrachten gebruikt worden (die verschillen in onderwerp en communicatief doel/genre). Dat betekent dat korte teksten die geschreven zijn door dezelfde doelgroep en dezelfde conventionele vorm hebben goed beoordeeld kunnen worden met dezelfde beoordelingsschaal. Het is dus niet nodig om voor elke schrijfopdracht een nieuwe schaal te maken, waardoor deze schaalbeoordeling (ondanks alternatieve verklaringen en beperkingen aan dit onderzoek) naar een ‘ideale’ methode neigt.

**Referenties**

Cooper, P. L. (1984). *The assessment of writing ability: A review of research*. Princeton, NJ: Educational Testing Service. GRE Board Research Report GREB No. 82-15R/ETS Research Report 84-112.

Huot, B (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research, 60* (2),237-263.

Jansen, F. (2010). Onkunde wordt bestraft. Het effect van systematisch gemaakte afwijkingen van de standaardtaal op de waardering van direct-mailbrieven. *Neerlandistiek.nl, 2*, 1-39.

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

Schellens, P.J. & Steehouder, M.R. (2008). Tekstanalyse: Methoden en Toepassingen. Assen: Uitgeverij van Gorcum.

Schoonen, R. (1991): De evaluatie van schrijfvaardigheidsmetingen. Een empirische studie naar betrouwbaarheid, validiteit en bruikbaarheid van schrijfvaardigheidsmetingen in de achtste groep van het basisonderwijs. [The evaluation of measures of writing ability. An empirical study into the reliability, validity and usefulness of measures of writing ability at grade 6.] Amsterdam: University of Amsterdam. Doctoral dissertation with English summary.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. Language Testing, *22*, 1-30.

Schoonen, R. & Glopper, C.M. de (1992). Toetsing van schrijfvaardigheid: problemen en mogelijkheden. *Levende talen Tijdschrift*, *470*, 187-195.

Van den Bergh, H., & Meuffels, B. (2000). Schrijfvaardigheden en schrijfprocessen. In A. Braet (Ed.), *Taalbeheersing als communicatiewetenschap* (pp. 1–32). Bussum: Uitgeverij Coutinho.

Wesdorp, H. (1981). De evaluatie van de schrijfvaardigheid. In *Evaluatietechnieken voor het moedertaalonderwijs*. 's-Gravenhage: Stichting voor Onderzoek van het Onderwijs.

**Bijlagen**

**Bijlage 1: Smikkelschaal**

**Bijlage 2: Beoordelingscriteria Smikkelschaal**

Man/Vrouw

Leeftijd: ………

Opleiding: ………

Beoordelingservaring: ………

**Scoringsvoorschrift (schaal)**

Leerlingen uit groep 8 van verschillende basisscholen hebben een brief geschreven naar aanleiding van de opdracht Smikkel. In deze opdracht moet de leerling een brief schrijven aan de organisatoren van een spaaractie voor een gratis muziek-cd. De voor de actie benodigde spaarpunten zijn nergens meer te krijgen, hoewel de einddatum van de actie nog niet verstreken is. De leerling heeft nu tot taak de organisatoren van de actie zo ver weten te krijgen dat zij hem toch de cd sturen.

Jouw taak is om 35 van deze leerlingteksten te beoordelen met behulp van de onderstaande schaal. Deze schaal is voorzien van ankerpunten, dat zijn prototypische teksten behorend bij die score. Op de volgende pagina is aangegeven wat de plus- en minpunten zijn van de ankerpunten. Lees dit goed door.

Geef vervolgens elke leerlingtekst een plaats op de schaal door ze te vergelijken met de ankerteksten en lees de bijbehorende score van deze schaal af. De schaal loopt door van 0 punten tot ondeindig, dus je kunt ook punten geven onder de 70 of boven de 130. Beoordeel de teksten zo nauwkeurig mogelijk. Zet voor elke tekst de score rechtsbovenin de hoek. Als je een gegeven score nog wilt aanpassen na het lezen van andere teksten dan is dat mogelijk, maar doe dit wel bij hoge uitzondering.

**Toelichting bij de ankerpunten:**

**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).

**Bijlage 3: Schrijfopdrachten**

**Schrijftaak: SMIKKEL**

Stel je voor ...

Op de SMIKKEL-repen, die je weleens eet, heb je zien staan dat je gratis muziek CD’s kunt krijgen.

Er staat:

SPAAR VOOR EEN GRATIS muziek CD!!

Dit moet u doen:

Op de wikkels van de repen zit 1 spaarpunt.

Spaar 10 punten en stuur deze in een voldoende

gefrankeerde envelop naar:

SMIKKEL SPAARACTIE

postbus 3333

1273 AD Etten-Leur

Stuur ook EUR 0,80 aan postzegels mee voor de

portokosten. Vermeld duidelijk uw naam, adres,

postcode en woonplaats en de gratis (GRATIS!)

muziek CD wordt dan zo spoedig mogelijk

thuisgestuurd.

Deze actie loopt tot 15 juli 2009.

Je hebt 8 punten bij elkaar gespaard, maar nu kun je nergens meer repen met punten vinden. Op de repen in de winkels zit geen spaarpunt meer, hoewel het nog geen 15 juli is. Je kunt je 10 punten dus niet bij elkaar krijgen. Toch wil je zo'n CD hebben.

Je stuurt daarom je 8 punten op en doet er 2 hele wikkels zonder punt bij.

**Opdracht**

Schrijf een briefje dat je meestuurt met je punten en

wikkels. Vertel waarom je geen 10 punten kunt opsturen.

Overtuig de firma SMIKKEL ervan dat jij er niets aan kunt doen dat je geen 10 punten hebt, maar dat je wél die muziek-­CD wilt ontvangen.

Zorg ervoor dat ze je de muziek-CD toch sturen!

Schrijf daarna de envelop.

**Schrijftaak: SMURFEN**

Stel je voor ...

Je vader doet wekelijks inkopen bij de supermarkt (SuperCoop) om de hoek. Voor elke vijfentwintig euro ontvangt hij een mini-Smurf (verpakt in een zakje). Je hebt inmiddels al een hele verzameling maar je mist nog twee smurfen: de Brilsmurf en Grote Smurf. Natuurlijk hoop je ook op de Gouden Smurf. Als je de Gouden Smurf hebt, maak je kans om een digitale fotocamera te winnen. Elke vrijdag kijk je nieuwsgierig in de boodschappentas welke Smurfen je bij je verzameling kunt zetten.

In de folder van Supermarkt “SUPERCOOP” staat:

SPAAR SMURFEN EN MAAK KANS OP EEN PRACHTIGE

DIGITALE CAMERA!

Zo werkt het:  
Bij elke 25 euro aan boodschappen, ontvangt u een zakje met een Smurf. Zo kunt u uw kinderen verrassen met een prachtige verzameling van deze populaire poppetjes.

Heeft u een Gouden Smurf, dan maakt u kans op een digitale camera van 4.0 megapixels!

Stuur uw Gouden Smurf in een gefrankeerde envelop naar:

SMURF SPAARACTIE

postbus 3333

1273 AD Etten-Leur

Stuur ook EUR 0,80 aan postzegels mee voor de

portokosten. Vermeld duidelijk uw naam, adres,

postcode en woonplaats en u krijgt vóór 1 mei 2008

bericht of u een camera hebt gewonnen

Deze actie loopt tot 20 april 2008.

Je hoort op 15 april van je vader dat de zakjes met Smurfen op waren. Hij heeft een kassabon meegekregen met een stempel, zodat je later de Smurfen alsnog kunt krijgen. Hij heeft in totaal EUR 110,34 aan boodschappen uitgegeven. Je besluit de kassabon op te sturen met de vraag om vóór 20 april nog zakjes met Smurfen te kunnen ontvangen. Je hoopt natuurlijk op een Gouden Smurf!

**Opdracht**

Schrijf een briefje dat je meestuurt met de kassabon van 15 april. Leg uit dat bij SuperCoop om de hoek de Smurfen op waren.

Overtuig de firma SUPERCOOP ervan dat je vóór de sluitingsdatum van de actie de zakjes met Smurfen wilt ontvangen. Schrijf daarna de envelop.

**Schrijftaak: LIKE**

Stel je voor…

Volgende week komt een nieuwe leerling bij jullie in de klas: Like.

Like is in Nederland geboren maar heeft een jaar in Engeland gewoond. Het onderwijs in Engeland is heel anders dan in Nederland. Like weet dan ook niet precies hoe je in het Nederlands een goede tekst schrijft. Want daarbij moet je op heel veel dingen letten.

**Opdracht**  
Schrijf nu een brief aan Like waarin je uitlegt hoe je in het Nederlands een goede tekst schrijft. En dus een hoog cijfer voor Taal krijgt.

Geef Like in je brief zo veel mogelijke tips en adviezen.

*Beste Like,*