De bruikbaarheid van de schaalmethode om schrijfvaardigheid te beoordelen

Een bachelor scriptie van Hanneke Liemburg

24-06-2013

Dit onderzoek bespreekt een van de meest betrouwbare methodes om schrijfvaardigheid te beoordelen: de schaalbeoordelingsmethode. Tot nu toe is altijd gedacht dat het niet mogelijk is een tekst te beoordelen met een schaal wanneer de ankerteksten verschilde van te beoordelen teksten wat betreft onderwerp en communicatief doel. Deze vraag staat centraal in dit onderzoek: '*wat is de invloed van de congruentie tussen ankerteksten van een beoordelingsschaal en te beoordelen teksten op de beoordelaarsovereenstemming?*' Om de vraag te kunnen beantwoorden zijn drie tekstsoorten met verschillende onderwerpen en genres door een jury van vijf beoordelaars beoordeeld met dezelfde schaal. Uit het experiment bleek dat de beoordelaarsovereenstemming bij de drie tekstsoorten even hoog was. Dat maakt de beoordelingsschaal een betrouwbaar instrument om schrijfvaardigheid te beoordelen, ook als de ankerteksten en te beoordelen teksten niet congrueren. De uitkomsten van het onderzoek impliceren dat de beoordelingsschaal breder inzetbaar is dan tot nu toe is gedaan, maar omdat in dit onderzoek slechts één beoordelingsschaal is getest, is het niet mogelijk de resultaten te generaliseren.

Inleiding

Eén van de belangrijkste vaardigheden die leerlingen van zowel basis-, middelbare als hogescholen onderwezen krijgen is schrijfvaardigheid. Jaarlijks schrijven leerlingen tientallen opstellen, werkstukken, betogen, beschouwingen, verslagen, essays, papers en scripties. Docenten moeten al die schrijfproducten becijferen en dat is niet eenvoudig. Renkema (1992) zegt hierover: "over tekstkwaliteit lopen de meningen sterk uiteen. Dat weet iedereen die op de middelbare school hetzelfde opstel bij verschillende docenten heeft ingeleverd" (p. 65). De beoordeling van schrijfvaardigheid is blijkbaar ingewikkeld. Dit heeft verschillende redenen. Het eerste probleem heeft te maken met de vraag of je met één schrijfproduct werkelijk iemands schrijfvaardigheid kan meten. Het prestatievermogen van een schrijver is niet constant. Er zijn altijd factoren aanwezig die het schrijfproces kunnen beïnvloeden zoals tijdsdruk, omgevingsgeluiden of persoonlijke factoren, die een negatieve invloed kunnen hebben op de concentratie van de schrijver en op de kwaliteit van zijn werk. Dergelijke factoren zorgen ervoor dat het lastig is om door middel van één tekst een uitspraak te doen over schrijfvaardigheid (McColly, 1970). Ook het onderwerp van de schrijftaak heeft invloed op het prestatievermogen van de schrijver. Elke schrijver heeft meer of minder voorkennis over bepaalde onderwerpen en

dit draagt bij aan de kwaliteit van het schrijfproduct. Deze factoren zorgen ervoor dat één tekst vaak niet representatief is voor schrijfvaardigheid (McColly, 1970). Van den Bergh en Meuffels (2000) vatten de problematiek van het beoordelen van schrijfvaardigheid als volgt samen:

> Het kernprobleem bij het meten van iemands niveau in schrijven . . . is dat de betreffende vaardigheden niet direct observeer- en dus meetbaar zijn, maar moeten worden afgeleid uit iemands handelingen, gedragingen en de concrete resultaten daarvan. Wanneer wij aan personen de dispositie 'schrijfvaardigheid' toeschrijven, dan spreken we daarmee tevens de verwachting uit dat die personen zich, telkens wanneer de situatie daartoe aanleiding geeft, op een soortgelijke wijze zullen gedragen. (p. 122-123)

Deze problematiek gaat over de validiteit van schrijfopdrachten: kan schrijfvaardigheid werkelijk worden getest door één schrijfproduct te beoordelen? Ook de validiteit van oordelen over schrijfvaardigheid is gecompliceerd: meet het tekstoordeel van een beoordelaar werkelijk schrijfvaardigheid? Naast deze validiteitskwesties zijn er problemen met de betrouwbaarheid van tekstoordeel. De betrouwbaarheid heeft te maken met de mate waarin beoordelaars met elkaar overeenstemmen in tekstoordeel.

Er wordt in het onderwijs geen gestandaardiseerde beoordelingsmethode voor schrijfvaardigheid gebruikt. Iedereen beoordeelt teksten op zijn eigen manier. Hierdoor zijn beoordelaars het dramatisch oneens over de kwaliteit van door hun te beoordelen teksten.

Het gebrek aan objectiviteit van beoordelingen is te wijten aan onduidelijkheid over een geschikte methode om schrijfvaardigheid te beoordelen. Er bestaan verschillende methodes om schrijfvaardigheid te beoordelen. Een van de meest gebruikte is de globale methode. Deze methode wordt in het onderwijs veel gebruikt omdat het beoordelen snel gaat, geen training vereist en dus door iedereen te gebruiken is en voor een redelijk acceptabel oordeel over de tekst in zijn geheel zorgt (Wesdorp, 1981). De globale methode is een snelle methode maar niet erg nauwkeurig. Een beoordelaar leest de tekst één of twee keer als geheel en velt er vervolgens een oordeel over zonder gebruik te maken van vastgelegde beoordelingsrichtlijnen. Er zijn veel verschillende tekstelementen waar beoordelaars rekening mee kunnen houden bij hun beoordeling. Waar de ene beoordelaar de inhoud relevant vindt en bijvoorbeeld let op de mate waarin de communicatieve doelen van de tekst worden bereikt, hecht een andere beoordelaar wellicht meer waarde aan de structuur van de tekst en andere vormelementen zoals grammatica, spelling en interpunctie. Omdat de verschillen in kwaliteit van deze tekstelementen erg groot kunnen zijn, verschillen beoordelaars in sommige gevallen sterk van mening over het niveau van een tekst. Het hangt dus sterk van de individuele beoordelaar af welk niveau aan de tekst toegekend wordt (Renkema, 1992).

De globale methode is hierdoor onbetrouwbaar: verschillende beoordelaars zullen eigen maatstaven gebruiken om een tekst globaal te beoordelen en dit zorgt voor grote verschillen in

oordeel. Dit maakt de beoordelaarsovereenstemming erg laag wanneer de globale methode gebruikt wordt (Wesdorp, 1974). Ook over de validiteit kan weinig positiefs gezegd worden. Het is bij de globale methode nooit met zekerheid te zeggen dat de beoordelaar werkelijk schrijfvaardigheid meet omdat het niet duidelijk is wat er precies is beoordeeld (Van den Bergh en Meuffels, 2000).

Een verklaring voor de lage betrouwbaarheid van de globale beoordeling is dat er veel bijeffecten voorkomen bij deze beoordelingsmethode. Het contaminatie-effect treedt op wanneer beoordelaars schrijvers van de teksten persoonlijk kennen en hun beoordeling hierdoor laten beïnvloeden. Bij de globale methode kan het tekstoordeel makkelijk beïnvloedt worden door persoonlijke factoren vanwege het subjectieve karakter van deze beoordelingsmethode. Het signifisch effect komt voor wanneer het voor beoordelaars niet duidelijk is welke taak ze hebben en welke aspecten van belang zijn. Het gebrek aan regels zorgt er voor dat de kans groot is dat het signifisch effect optreedt bij de globale beoordelingsmethode. Verder kan het sequentie-effect optreden. Dit effect komt voor wanneer beoordelaars veel teksten achter elkaar moeten beoordelen. Eerder beoordeelde teksten kunnen dan invloed hebben op het oordeel voor latere teksten (Wesdorp, 1981). Ook kan hierdoor normverschuiving optreden; de beoordelaar past zijn oordelen aan het gemiddelde niveau van de teksten aan.

De analytische beoordelingsmethode beperkt het sequentie- en het signifisch effect in belangrijke mate. Bij de analytische methode wordt schrijfvaardigheid gemeten door tekstkenmerken van elkaar te onderscheiden en apart te beoordelen. Dit gebeurt door per tekstkenmerk punten toe te kennen. Zo worden zowel inhoudelijke tekstkenmerken zoals communicatief doel als vormkenmerken zoals structuur meegewogen. Het uiteindelijke tekstniveau wordt bepaald door de punten bij elkaar op te tellen. Er zijn veel verschillende analytische schema's ontwikkeld. In grote lijnen meten de meeste analytische schema's hetzelfde. Er wordt bijvoorbeeld aandacht besteed aan structuur, alinea's, communicatieve doelen, stijl en interpunctie (Wesdorp, 1981).

Ondanks de overeenkomsten tussen de analytische schema's bestaat er op gedetailleerder niveau veel onenigheid over welke tekstkenmerken het meest relevant zijn om schrijfvaardigheid te kunnen meten. Analytische schema's kunnen daarom sterk uiteenlopen. Een ander veelgehoord bezwaar is dat een goede tekst meer is dan de som der delen. Het simpelweg bij elkaar optellen van losse tekstkenmerken doet geen recht aan de tekst als geheel en is in die zin beperkt (Wesdorp, 1981). Dit zorgt volgens Wesdorp (1981) ook voor bezwaren wat betreft de validiteit van de analytische beoordelingsmethode. Verschillende losse tekstkenmerken apart beoordelen, zegt weinig over het schrijfproduct in zijn geheel en meet dus geen schrijfvaardigheid.

Hoewel zowel het contaminatie-effect als het signifisch effect bij de analytische beoordelingsmethode minder snel voorkomen, kan het halo-effect wel optreden (Van den Bergh en Meuffels, 2000). Het halo-effect is bij de analytische methode het effect dat het aantal punten voor het ene tekstkenmerk invloed heeft op de beoordeling van het volgende tekstkenmerk. Hoewel te verwachten valt dat de analytische methode betrouwbaarder is dan de globale methode vanwege de concrete maatstaven is uit onderzoek gebleken dat er nauwelijks sprake is van betrouwbaarheidswinst (Van den Bergh en Meuffels, 2000). Ook bij de analytische methode zijn beoordelaars het dus zelden met elkaar eens over de kwaliteit van een tekst.

De schaalbeoordelingsmethode gebruikt een schaal waarin ankerteksten staan die elk een bepaald aantal punten en zo een schrijfniveau representeren. Beoordelen gebeurt door de tekst te vergelijken met de ankerteksten in de schaal en zo punten toe te kennen. Meestal staan er in een beoordelingsschaal vijf ankerteksten die qua niveau verschillen van 'zeer zwak' tot 'zeer sterk'. Als beoordelaar kan je punten toekennen binnen de schaal, maar ook daarbuiten als de kwaliteit van een tekst uitzonderlijk hoog of laag is (Pollmann et al, 2012).

De schaalbeoordelingsmethode heeft veel voordelen vergeleken met de globale en de analytische methode. Zo is de schaalmethode een stuk betrouwbaarder dan de globale en de analytische methode (Van den Bergh en Meuffels, 2000). Beoordelaars geven veel meer dezelfde oordelen over een tekst wanneer zij de beoordelingsschaal gebruiken dan wanneer zij de globale of de analytische beoordelingsmethode gebruiken (Pollman et al, 2012). Dit komt volgens Wesdorp (1974) doordat 'de schaal . . . docenten [helpt] zich te conformeren aan reeds bestaande maatstaven' (p. 21). De ankerteksten gelden als concrete referentiepunten voor tekstniveau en geven zo veel houvast aan beoordelaars. Naast de betrouwbaarheidswinst kunnen met de schaalbeoordelingsmethode meerdere bijeffecten worden voorkomen (Pollman et al, 2012). Het signifisch effect zal niet of zeer beperkt optreden. Wanneer een beoordelaar de schaal goed gebruikt, zullen de verschillen in oordeel tussen beoordelaars veel kleiner worden omdat er met de ankerteksten een duidelijk referentiepunt aanwezig is. Er is bij de schaalmethode ook geen sprake van het sequentie-effect. Omdat elke te beoordelen tekst aan de ankerteksten op de schaal wordt gespiegeld, worden teksten minder onderling vergeleken. Ook normverschuiving is bij de beoordelingsschaal niet aan de orde omdat de schaal zelf een duidelijke norm stelt (Pollman et al, 2012). Naast deze voordelen wordt ook de validiteit bij de schaalmethode beter gewaarborgd dan bij de eerder genoemde beoordelingsmethodes. Door teksten te vergelijken met de schaal wordt in tegenstelling tot de analytische methode naar de tekst in zijn geheel gekeken. Ook zorgt de beoordelingsschaal in tegenstelling tot de globale methode voor concrete referentiepunten die een niveau van schrijfvaardigheid representeren (Wesdorp, 1981).

Het is echter veel werk een schaal te maken. Volgens Wesdorp (1981) moet voor elke te beoordelen tekst een aparte schaal worden ontwikkeld. Dit is tijdrovend en de voornaamste reden dat de schaalmethode niet veel wordt gebruikt in het onderwijs. Wesdorp (1981) stelt zonder dit te onderbouwen dat het moeilijker wordt ankerteksten in een schaal als referentiepunt te gebruiken wanneer deze teksten een ander genre en onderwerp hebben dan de teksten die beoordeeld moeten worden. Er is echter nog geen onderzoek gedaan naar de vraag of het wel of niet mogelijk is een beoordelingsschaal te gebruiken voor verschillende tekstgenres. Er is hier dus sprake van een hiaat in onze kennis.

Om meer duidelijkheid over deze kwestie te geven, is het van belang de term tekstgenre goed te definiëren. Renkema (2009) hanteert de volgende definitie van tekstgenre: "a particular type of communicative event that had a particular communicative purpose recognized by its users" (p. 82). Deze definitie impliceert dat het communicatieve doel van een tekst bepalend is voor het genre van de tekst. Dudley-Evans (1994) geeft dezelfde kijk op tekstgenre weer: "communicative purpose is, in fact, the defining feature by which a genre . . . is distinguished from other genres" (p. 219). Aan de hand van deze definities wordt in dit onderzoek tekstgenre onderscheiden op basis van communicatief doel. Er worden door Boezeman et al. (1979) verschillende communicatieve doelen onderscheiden waarvan het informatieve doel, het commentariërende doel, het persuasieve doel en het activerende doel volgens hen de belangrijkste zijn. Hoeken (2009) beschrijft de belangrijkste verschillen tussen overtuigende en informatieve teksten:

> In dit verband is het onderscheid tussen sturende en niet-sturende voorlichting van belang. Bij sturende voorlichting [persuasief van aard] verstrekt de voorlichter informatie met als doel dat de doelgroep een bepaalde conclusie trekt. Bij niet-sturende voorlichting [informatief van aard] verstrekken voorlichters informatie zonder daarbij de doelgroep in de richting van een bepaalde conclusie te sturen. (p. 16)

In dit onderzoek worden drie tekstsoorten van twee verschillende genres gebruikt, persuasieve teksten en informatieve teksten. Er zal een experiment worden uitgevoerd om de onderbouwing te verstrekken die Wesdorp (1981) niet gaf en zo de kennis over de bruikbaarheid van de schaalbeoordelingsmethode uit te breiden.

Het onderzoek is relevant omdat de uitkomsten docenten mogelijk betrouwbaardere instrumenten geven om schrijfvaardigheid te beoordelen. Als uit het experiment blijkt dat de schalen voldoende betrouwbaar zijn bij meerdere tekstsoorten heeft dit positieve gevolgen voor de beoordeling van schrijfvaardigheid omdat het makkelijker wordt een betrouwbaar oordeel te geven over de kwaliteit van een tekst. Er hoeft dan niet meer voor elke tekst een nieuwe beoordelingsschaal ontwikkeld te worden. Dit maakt de stap kleiner de schaalmethode te gebruiken om schrijfvaardigheid te beoordelen.

De vraag die in dit onderzoek wordt gesteld is: *wat is de invloed van de congruentie tussen ankerteksten van een beoordelingsschaal en te beoordelen teksten op de beoordelaarsovereenstemming?* De onafhankelijke variabele is de congruentie tussen ankerteksten op de schaal en te beoordelen teksten en zal worden geoperationaliseerd met behulp van een beoordelingsschaal met ankerteksten van het genre 'persuasieve teksten'. Twee van de tekstsoorten die beoordeeld zullen worden hebben ditzelfde tekstgenre. De derde tekstsoort is van het genre 'informatieve teksten'. Dit maakt dat deze laatste tekst fundamenteel verschilt van de andere twee teksten. Behalve de genreverschillen tussen de tekstsoorten, zijn ze ook inhoudelijk anders. De drie tekstsoorten hebben verschillende onderwerpen.

De afhankelijke variabele is de beoordelaarsovereenstemming. De teksten zullen worden beoordeeld door een jury van vijf beoordelaars. Als de onderlinge overeenstemming over de kwaliteit van een tekst hoog is, is er sprake van een betrouwbaar gemiddeld oordeel. De beoordelaarsovereenstemming tussen de drie teksten zal worden vergeleken. Dit zal duidelijk maken of de schaal ook betrouwbaar is wanneer teksten die een ander onderwerp en/ of communicatief doel hebben dan de ankerteksten worden beoordeeld.

Verwacht wordt dat minder congruentie tussen ankerteksten en te beoordelen teksten leidt tot een lagere beoordelaarsovereenstemming. Als de congruentie tussen ankerteksten en te beoordelen teksten laag is, verschillen de teksten sterk van elkaar. Hierdoor wordt het moeilijk voor beoordelaars de ankerteksten als referentiepunt te gebruiken. Bij gebrek aan een goed referentiepunt, zullen beoordelaars wellicht terugvallen op hun eigen maatstaven. Zo is er in feite weer sprake van een globale beoordeling. De beoordelaarsovereenstemming zal in dat geval erg laag zijn bij de derde tekstsoort. Bovendien zou het signifisch effect hier in sterke mate terug kunnen komen. Ook het contaminatie-effect en het halo-effect zouden opnieuw kunnen optreden (Wesdorp, 1974).

Methode

Proefpersonen

De teksten werden beoordeeld door een jury bestaande uit vijf personen. De geselecteerde beoordelaars waren allen ouderejaars studenten die een aan taal of communicatie gerelateerde studie volgden aan de Universiteit. De beoordelaars waren dus allen hoogopgeleid en deelden een talige achtergrond. Ze hadden daarom veel kennis over taal en ervaring met teksten zodat verwacht kon worden dat ze goed in staat waren de kwaliteit van een tekst in te schatten. Alle beoordelaars waren vrouwen. De gemiddelde leeftijd van de beoordelaars was 21,8 met een standaarddeviatie van 1,09. Kennis over de achtergrond van de beoordelaars zal eventuele

discrepanties in de resultaten mogelijk kunnen verklaren. Ook biedt deze informatie wellicht ruimte voor discussie.

Materiaal

Er werd een reeds bestaande beoordelingsschaal gebruikt: de smikkelschaal. Deze beoordelingsschaal is terug te vinden in Bijlage 1. In de schaal stonden vijf ankerteksten die geschreven waren door kinderen uit groep 8. Dit waren overtuigende teksten met de smikkelrepen als onderwerp. Het doel van deze schrijftaak was een gratis cd te bemachtigen ondanks een tekort aan spaarpunten. Door smikkelrepen met actiepunten te kopen kon een cd gespaard worden maar op de laatste twee wikkels zaten geen actiepunten meer. Door een brief te schrijven moesten de makers van Smikkel ervan overtuigd worden alsnog een cd op te sturen.

De vijf teksten in de schaal representeerden een bepaald schrijfniveau, van 'zeer zwak' tot 'zeer sterk'. Elk niveau stond gelijk aan een bepaald aantal punten. De ankerteksten op de beoordelingsschaal zijn geselecteerd door een jury van drie deskundige beoordelaars. Zij hebben voor de ankerteksten gekozen omdat ze kenmerkend waren voor de bijbehorende schaalpunten. Daarnaast zijn de ankerteksten geselecteerd omdat de beoordelaarsovereenstemming bij deze teksten hoog was.

Er is gekozen voor drie verschillende tekstsoorten. De smikkeltekst is hierboven al beschreven en was een overtuigende tekst met de smikkelrepen als onderwerp. De tweede tekstsoort was ook overtuigend en had smurfen spaarpoppetjes als onderwerp. Kinderen moesten voor deze schrijfopdracht een tekst schrijven om de winkel Supercoop ervan te overtuigen dat zij recht hadden op smurfen van de smurfenactie omdat hun ouders voldoende geld aan boodschappen hadden besteed maar de smurfen in de winkel op waren. De derde tekstsoort heeft een informatief tekstdoel. De kinderen hadden voor deze tekst de schrijfopdracht een nieuwe klasgenoot uit Engeland in te lichten over het onderwijs in Nederland.

Van elke tekstsoort zijn er 40 geselecteerd. Er is gekozen voor een normaalverdeling van de tekstniveaus voor een representatieve steekproef. Er is daarom één 'zeer zwakke' en één 'zeer sterke' tekst geselecteerd, vijf 'zwakke' en 'sterke' teksten en 14 teksten die op of net boven of onder het gemiddelde niveau zitten.

Procedure

Het experiment vond plaats in een kleine kantoorruimte. De vijf beoordelaars zaten aan één grote tafel. Ook was de onderzoeker aanwezig om uitleg te geven en toezicht te houden. De beoordelaars kregen voorafgaand aan het onderzoek een beoordelaarsinstructie. In deze instructie staat informatie over het onderzoek en uitleg over de schaal en de drie tekstsoorten.

Tijdens de instructie werden drie voorbeeldteksten beoordeeld en daarna besproken om de beoordelaars te laten wennen aan de schaal. Daarnaast werd in de instructie aangegeven wat er van de beoordelaars werd verwacht; de beoordelaars dienden een oordeel te geven over de teksten door deze te vergelijken met de ankerteksten in de schaal. Aan de hand van de vergelijkingen, kenden ze punten toe aan de tekst. De beoordelaars moesten vooral letten op het doel van de tekst en in hoeverre dat werd bereikt. Ze mochten hun score van een tekst achteraf niet aanpassen, omdat het niet de bedoeling was dat de teksten onderling zouden worden vergeleken. Er werd echter wel toegestaan dat beoordelaars hun scores één of twee keer achteraf aanpasten, zolang het daarbij bleef. De beoordelaarsinstructie is terug te vinden in Bijlage 2. De instructie nam ongeveer 20 minuten in beslag. Na de instructie waren er nog 100 minuten over voor de beoordelingen. Alleen de teksten die door elke beoordelaar waren beoordeeld, werden meegenomen in de data-analyse. De drie tekstsoorten werden om de beurt beoordeeld; eerst de smikkelteksten, vervolgens de smurfenteksten en als laatste de Like-teksten. De teksten zijn gerandomiseerd zodat ze niet lineair opliepen in kwaliteit. In totaal duurde het experiment twee uur.

Na afloop van de beoordelingen vulden de beoordelaars persoonsgegevens en een evaluerend vragenformulier in. Dit formulier bevatte vragen over de wijze waarop de beoordelaar beoordeelde. Er stonden bijvoorbeeld vragen in over hoe streng de beoordelaar zichzelf vond, waar de beoordelaar vooral op heeft gelet en hoe moeilijk de beoordelaar het vond om scores toe te kennen aan de verschillende tekstsoorten. Het evaluatieformulier zal mogelijk duidelijkheid geven wanneer één van de beoordelaars sterk afwijkt van de andere.

Data-analyse

Na het experiment zijn er scores van vijf beoordelaars beschikbaar over 120 teksten die zijn beoordeeld aan de hand van de Smikkelschaal. De jury is n=5.

Om te kijken wat de invloed is van de congruentie tussen ankerteksten van een beoordelingsschaal en te beoordelen teksten op de beoordelaarsovereenstemming zal de beoordelaarsovereenstemming per tekstsoort worden berekend door middel van de SPSS-functie Cronbachs Alpha. Als de beoordelaarsovereenstemming hoog is, betekent dit een betrouwbaar gemiddeld oordeel. Dat zou de schaal tot een geschikt beoordelingsinstrument maken. Om de onderzoeksvraag te kunnen beantwoorden zal gekeken worden of de beoordelaarsovereenstemming verschilt bij de drie tekstsoorten.

Resultaten

De beoordelaarsovereenstemming was bij de Smikkelteksten en de Smurfenteksten even hoog ($F(1, 39)= 1.21$, $p= .27$). Ook tussen de Smurfenteksten en de Liketeksten is geen verschil in beoordelaarsovereenstemming gevonden ($F(1, 39)= 1.25$, $p= .24$). Ten slotte is er ook geen verschil gevonden in beoordelaarsovereenstemming tussen de Smikkeltekst en de Liketekst ($F(1, 39)= 1.52$, $p= .09$). De gemiddelde Cronbach's Alpha van de drie teksten is $\alpha= .89$.

Discussie

De invloed van de congruentie tussen ankerteksten van een beoordelingsschaal en te beoordelen teksten op de beoordelaarsovereenstemming is onderzocht met behulp van een experiment waarbij drie teksten met verschillende onderwerpen en genres werden beoordeeld door middel van een beoordelingsschaal. Uit de geanalyseerde data is gebleken dat de beoordelaarsovereenstemming bij de drie tekstsoorten even hoog was. De congruentie tussen ankerteksten van een beoordelingsschaal en te beoordelen teksten heeft dus geen invloed op de beoordelaarsovereenstemming. De verklaring voor deze resultaten is dat een ankertekst een concreet voorbeeld geeft van een tekstniveau en beoordelaars zo voldoende houvast geeft om tot een goed oordeel te komen. Op deze manier helpt een schaal beoordelaars zich aan te passen aan bepaalde maatstaven voor schrijfvaardigheid Wesdorp (1974). Uit dit experiment is gebleken dat de schaal, ook als de congruentie tussen de ankerteksten en te beoordelen teksten laag is, beoordelaars helpt zich aan deze maatstaven te conformeren.

Deze uitkomst heeft grote gevolgen voor de praktijk. In de inleiding is besproken dat de beoordelingsschaal ondanks haar hoge betrouwbaarheid niet veel wordt gebruikt in het onderwijs. Het kost namelijk veel tijd een beoordelingsschaal te ontwikkelen en Wesdorp (1981) nam aan dat je een schaal niet kon gebruiken voor teksten met een ander onderwerp en tekstgenre dan de ankerteksten. De uitkomsten van dit onderzoek spreken Wesdorps bezwaren tegen. De hoge beoordelingsovereenstemming bij de drie teksten impliceert dat je een beoordelingsschaal kan gebruiken voor meer dan één tekst. Dit zou betekenen dat het onderwijs op grote schaal gebruik kan gaan maken van de schaalbeoordelingsmethode voor schrijfvaardigheid. Een docent heeft voortaan genoeg aan één schaal om standaard te gebruiken voor alle teksten die beoordeeld moeten worden.

Er zijn echter ook alternatieve interpretaties van de resultaten denkbaar. Het is mogelijk dat de operationalisering van het experiment invloed had op de resultaten. Zo zou de tekstvolgorde bij het experiment van invloed geweest kunnen zijn. In het experiment werden eerst de Smikkelteksten beoordeeld, vervolgens de Smurfenteksten en als laatste de Like-

teksten. Hierdoor hadden de beoordelaars al veel teksten beoordeeld en dus ervaring met de beoordelingsschaal nog voor zij aan de Like-teksten begonnen. Dit kan bijgedragen hebben aan de hoge beoordelaarsovereenstemming. Wellicht was de overeenstemming van de Like-teksten veel lager geweest wanneer die als eerste waren beoordeeld.

Een andere alternatieve verklaring brengt ons terug bij de genreverschillen die zijn besproken in de inleiding. Renkema (2009) onderscheidde tekstgenres op basis van hun communicatieve doel. De Like-tekst verschilde van de Smikkel- en de Smurfentekst vooral wat betreft communicatief doel en onderwerp. Maar wellicht zijn de genreverschillen tussen de drie teksten niet zo groot als van tevoren was aangenomen. Pander Maat (2002) definieert tekstgenre aan de hand van teksthandelingen. Volgens hem hoort bij iedere groep teksthandelingen een communicatief doel maar staan deze groepen in een tekst niet altijd logisch bij elkaar en is het bijvoorbeeld ook mogelijk dat een tekst met een informatief tekstdoel persuasieve teksthandelingen bevat. Dit wordt een impliciete strategie genoemd. Door deze strategie te gebruiken kan een doel worden bereikt zonder de lezer hier expliciet op aan te spreken. De theorie van Pander Maat (2002) zou in dit onderzoek kunnen impliceren dat de drie teksten in werkelijkheid niet zoveel verschillen. Zo zitten er wel degelijk persuasieve elementen in de Like-teksten; de kinderen willen hun nieuwe klasgenoot niet alleen informeren maar hem vaak ook ergens van overtuigen, bijvoorbeeld dat het een goed idee is om zinnen te beginnen met een hoofdletter. Vooral in deze door kinderen geschreven teksten is het niet ondenkbaar dat de teksthandelingen soms contrasteren met het tekstdoel. Dit zou betekenen dat de beoordelaarsovereenstemming zo hoog was omdat de drie tekstsoorten in hoge mate overeenkwamen met de ankerteksten. Er is dan in werkelijkheid niet geëxperimenteerd met verschillende genres. Dit zou in vervolgonderzoek grondiger uitgewerkt moeten worden. Er zou een betere analyse van tekstgenre moeten komen op basis waarvan daadwerkelijk van elkaar verschillende tekstsoorten geselecteerd kunnen worden. Dit zou werkelijk aantonen wat de invloed is van congruentie tussen ankerteksten op een beoordelingsschaal en te beoordelen teksten op beoordelaarsovereenstemming.

Dit onderzoek is te beperkt om deze alternatieve verklaringen met zekerheid te kunnen uitsluiten. Er is maar één beoordelingsschaal gebruikt in het experiment, waardoor niet met zekerheid kan worden gesteld dat de uitkomsten niet te maken hadden met de specifieke schaal. Zoals in de methode besproken is de smikkelschaal ontwikkeld door een jury van drie deskundige beoordelaars. Het valt daarom te verwachten dat deze schaal van hoge kwaliteit is en beoordelaars veel ondersteuning biedt door middel van duidelijke ankerteksten. In de praktijk zullen beoordelingsschalen vaak ontwikkeld worden door docenten. Zij hebben over het algemeen minder ervaring met het maken van een beoordelingsschaal en zijn daardoor mogelijk minder goed in staat geschikte ankerteksten te selecteren. Wellicht is de

beoordelaarsovereenstemming een stuk lager bij andere beoordelingsschalen. Dit zou in vervolgonderzoek getest moeten worden door het onderzoek uit te voeren met meerdere schalen.

Een andere beperking is dat de beoordelaars van het experiment dezelfde universitaire en talige achtergrond hebben. Dit kan invloed hebben gehad op de hoge beoordelaarsovereenstemming. Deze beoordelaars hadden door hun hoge opleiding en talige achtergrond veel ervaring met taal en teksten, waardoor ze goed in staat waren het niveau van een tekst in te schatten. Het is mogelijk dat beoordelaars die lager opgeleid zijn en minder ervaring met taal hebben minder goed in staat zijn teksten te beoordelen en dit zou kunnen leiden tot een lagere beoordelaarsovereenstemming. In vervolgonderzoek zou hetzelfde experiment moeten worden uitgevoerd met een meer gevarieerde groep beoordelaars. Ook de achtergrond van de schrijvers van de drie tekstsoorten en de ankerteksten is een beperking. De teksten zijn geschreven door kinderen uit groep 8. Er zijn alleen korte teksten gebruikt bij het experiment. Het is niet duidelijk of de beoordelingsschaal ook bruikbaar is voor lange, moeilijker teksten. Hier zou in vervolgonderzoek mee geëxperimenteerd kunnen worden. Samenvattend is het in vervolgonderzoek dus nodig de flexibiliteit van de beoordelingsschaal nogmaals te testen door te experimenteren met andere beoordelingsschalen, teksten en beoordelaars.

De drie hierboven beschreven beperkingen geven problemen voor de generaliseerbaarheid van het onderzoek omdat het nog onzeker is of de beoordelingsschaal ook betrouwbaar is in andere situaties. Desondanks zijn de uitkomsten van het onderzoek veelbelovend. De resultaten van het experiment geven goede hoop dat de beoordelingsschaal niet hoeft te worden beperkt tot één tekst maar ook gebruikt kan worden om andere tekstsoorten te beoordelen. Dit onderzoek kan beschouwt worden als een goede stap in de richting van het oplossen van de problemen die de beoordeling van schrijfvaardigheid met zich meebrengt.

Literatuur

Boezeman, L., van Noord, L., Verburg, M. (1979). *Over doel en publiek. Een verkennend onderzoek naar het communicatief aspect van zakelijk schrijven in het voortgezet onderwijs* (doctoraalscriptie). Nederlandse taal en cultuur, Faculteit Geesteswetenschappen, Universiteit Utrecht, Utrecht.

Dudley-Evans, T. (1994). Genre analysis: an approach to tekst analysis for ESP. In M. Coulthard. *Advances in written tekst analysis.* (pp. 219-228) (1e druk). Londen: Routledge.

Hoeken, H., Hornikx, J., & Hustinx, L. (2009). *Overtuigende teksten: Onderzoek en ontwerp*. Bussum: Coutinho.

Pander Maat, H. (2002). *Tekstanalyse. Wat teksten tot teksten maakt.* Bussum: Coutinho

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

Renkema, J. (1992). Doen uw teksten hun werk goed? *De geletterde mens*, 57-73.

Renkema, J. (2009). *Discourse, of Course. An overview of research in discourse studies*. Amsterdam: John Benjamins.

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.

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