

DE BEOORDELING VAN SCHRIJFVAARDIGHEID IN DE TWEEDE FASE

Bert Meuffels

In de les- en examenprogramma's Nederlands voor de Tweede Fase van havo/vwo zal de schrijfvaardigheid van de leerlingen vastgesteld worden aan de hand van een schrijfdossier en een gedocumenteerde tekst; voor havo-leerlingen zal dit plaatsvinden na 2001, voor vwo-leerlingen na 2002. De vraag die ik in deze bijdrage wil beantwoorden, luidt: wordt de schrijfvaardigheid van leerlingen in dit nieuwe programma beter in kaart gebracht dan via het 'traditionele' opstel?

1 Het schrijfdossier en de gedocumenteerde tekst

In het nieuwe programma 'schrijfvaardigheid' in de Tweede Fase moeten leerlingen vanaf de vierde klas gaan werken aan de invulling van het schrijfdossier. Dit dossier bestaat uit een verzameling verschillende typen teksten (zoals een betoog, een uiteenzetting, een brief) die de leerling naar behoren moet hebben afgerond voordat hij aan de afsluitende toets mag deelnemen. Het idee achter dit dossier is dat leerlingen zich ervan bewust moeten worden dat schrijven een proces is waarin stap voor stap geleerd moet worden om verschillende soorten informatie op een systematische manier begrijpelijk en doelgericht over te brengen.

In het nieuwe programma moeten leerlingen niet alleen verschillende typen teksten schrijven, het is ook de bedoeling dat zij leren documentatie te verzamelen, schrijfplannen te maken, proefteksten te schrijven en revisiecommentaar te ontvangen voordat zij de uiteindelijke versie in het dossier mogen stoppen. Als dit schrijfdossier naar het oordeel van de leerkracht adequaat is afgerond, hebben leerlingen naar verwachting voldoende zelfvertrouwen en vaardigheden ontwikkeld om het schoolexamen met vertrouwen tegemoet te kunnen zien.

Op het schoolexamen - niet op het centraal examen zoals oorspronkelijk in de bedoeling lag - zal de gedocumenteerde tekst in aansluiting op het schrijfdossier gaan fungeren als afsluitende toets van de schrijfvaardigheid. Anders dan bij het traditionele opstel waarbij de vwo-leerlingen voor hun eindexamen aan de hand van een tiental titels een hun aansprekend onderwerp mochten kiezen en 'uit het blote hoofd' over het gekozen onderwerp iets informatiefs moesten zien mee te delen, wordt er bij de nieuwe toetsvorm van uitgegaan dat een tekst pas interessant is voor een bepaald publiek als die op tenminste enige documentatie berust. Conform deze gedachtegang moeten de leerlingen op het nieuwe schoolexamen een tekst produceren op basis van documentatie. Wat de keuze van deze documentatie betreft - en dus ook: wat het door de leerlingen te behandelen onderwerp betreft - hebben de leerkrachten een zekere speelruimte: zij kunnen hun leerlingen laten schrijven over een thema dat aansluit bij een van de behandelde onderwerpen uit het schrijfdossier, maar ze kunnen ook

opteren voor een nieuw thema waarover pas documentatie wordt verstrekt op de zitting van het schoolexamen.

2 Klachten over gebrekkige beoordelaarsbetrouwbaarheid

Hoe kan de beoordeling van de schrijfvaardigheid aan de hand van schrijfdossier en gedocumenteerde tekst optimaal verlopen? Dat deze vraag aan de orde moet worden gesteld, vloeit voort uit het gegeven dat objectieve maatstaven voor het vaststellen van iemands schrijfvaardigheid ontbreken – of dat nu gebeurt aan de hand van een opstel of aan de hand van een gedocumenteerde tekst – zodat menselijke beoordelaars (experts; neerlandici dus) ingeschakeld moeten worden om te bepalen in welke mate iemand schrijfvaardig is.

Hét kernprobleem bij het gebruik van menselijke beoordelaars is dat deze ter zake geachte kundigen het in de praktijk vaak niet met elkaar eens blijken te zijn – een empirisch feit dat al in 1888 door Edgeworth werd vastgesteld, en nadien telkens weer werd bevestigd door tal van onderzoekers in verschillende landen. Maar het blijft niet bij dit gebrek aan *intersubjectieve overeenstemming*. De experts die als beoordelaar optreden, blijken het niet alleen vaak met andere experts niet eens te zijn, maar ook niet met zichzelf: bij herbeoordeling van dezelfde reeks opstellen blijken de oordelen vaak wat anders uit te vallen.

In Nederland werden beide effecten – gebrek aan *intersubjectieve overeenstemming* en gebrek aan *stabiliteit*, kortom gebrek aan beoordelaarsbetrouwbaarheid – onder anderen door Wesdorp (1983) aangetoond in een empirisch onderzoek waarin hij acht ervaren docenten vroeg een groot aantal opstellen van eindexamenkandidaten te beoordelen. Geen enkel opstel uit die grote verzameling werd door de docenten unaniem als voldoende of als onvoldoende beoordeeld: kwalificeerden zeven docenten een opstel als voldoende, dan vond de achtste de kwaliteit onvoldoende, en als zeven docenten de kwaliteit van het schrijfproduct onder de maat oordeelden, dan kon het er volgens de achtste wel degelijk mee door. De docenten moesten de opstellen na een tijdje opnieuw beoordelen: een verschil van drie punten tussen de eerste en tweede beoordeling bleek daarbij niet ongebruikelijk. Bij deze docenten die elk het predikaat 'ervaren' verdienden, kon een eindexamenkandidaat voor zijn opstel bijvoorbeeld een 4 halen (bij de eerste beoordeling), maar evenzogoed een 7 (bij de tweede beoordeling).

Zulke 'alarmerende' resultaten blijven uiteraard niet beperkt tot onze landsgrenzen. In Duitsland bijvoorbeeld voerde Schröter in de jaren zestig van de vorige eeuw het meest grootscheepse onderzoek naar de betrouwbaarheid van opstelbeoordeling uit dat ik ken: liefst 1113 leraren beoordeelden, in groepen van wisselende samenstelling, gezamenlijk een selectie van 617 opstellen uit een totaalbestand van 6135. De resultaten van het onderzoek waarvan Schröter verslag deed in het geruchtmakend boek *Die ungerechte Aufsatzzensur* (1971), toonden volgens hem overduidelijk aan dat de manier waarop in Duitsland opstellen worden beoordeeld, alarmerend is: in 43 percent van de gevallen liep het oordeel over de kwaliteit van een opstel liefst vier punten uiteen; meer dan 10 percent van de opstellen kreeg een cijfer dat uiteenliep van 1 tot 5 of van 2 tot 6, en zelfs zes opstellen kregen een cijfer dat

varieerde van 1 tot 6 (men bedenke dat de Duitse cijferschaal zes categorieën bevat, met de '1' als hoogste en de '6' als laagste cijfer). Bovendien werd *in geen enkel geval* door alle 18 leerkrachten die hetzelfde opstel moesten beoordelen, hetzelfde cijfer gegeven. Kortom, dit alles rechtvaardigde volgens Schröter de kernachtige titel van zijn monografie: opstelbeoordeling kan in redelijkheid niet anders dan onrechtvaardig genoemd worden.

Het beeld dat in deze studies van de betrouwbaarheid van de neerlandicus als opstelbeoordelaar wordt geschetst, stemt niet bepaald vrolijk. Bovendien wordt dit beeld waarin het *testimonium paupertatis* van de neerlandicus als opstelbeoordelaar besloten ligt, bij voortduring in kranten en vaktijdschriften (her)bevestigd, met name in de jaren vijftig en zestig. Zo concludeert Van den Ende in 1954, na analyse van empirisch materiaal, dat opstelcijfers van een 'bandeloze subjectiviteit' getuigen; zo'n tien jaar later claimt De Groot in *Vijven en zessen* (1968) dat in een cijfer eerder de beoordelaar zelf dan de te beoordelen prestatie van de leerling tot uitdrukking wordt gebracht en in 1986 verklaart een onderzoeker de beoordeling van opstellen tot een 'volstrekt willekeurige zaak' en het opstel 'als examenonderdeel immoreel' (*De Volkskrant*, november 1986).

Ter relativering van dit negatieve beeld moet opgemerkt worden dat de klaagzangen over de onbetrouwbaarheid van de opstelbeoordelaar niet zelden overdreven zijn. Schröter kon alleen maar tot zijn schokkende conclusies komen doordat hij de opstellen die hij ter beoordeling aan de leerkrachten Duits voorlegde, bewust selectief en dus gecontamineerd samenstelde: met opzet liet hij juist die 617 opstellen uit het totale bestand van 6135 beoordelen die door studenten (die anders dan de leerkrachten vooraf, voor het eigenlijke onderzoek, wél alle 6135 opstellen hadden nagekeken) 'als besonders problematisch empfunden werden', kortom juist die opstellen waarbij zich tussen de studenten 'grosse Beurteilungsunterschiede' manifesteerden; de opstellen echter waarbij sprake was van 'wenige gegensätzliche Meinungen', werden willens en wetens buiten de selectie gehouden. Moet men vervolgens verbaasd staan als de leerkrachten nogal forse verschillen vertonen in hun cijfergeving?

Opstelbeoordeling is niet zo apert onrechtvaardig als sommigen willen doen voorkomen, en evenmin is de beoordeling van opstellen door leraren Nederlands een *volstrekt* willekeurige zaak. Zelf liet ik eens een groep neerlandici, leraren wiskunde en managers dezelfde serie van 24 eindexamenopstellen beoordelen en herbeoordelen. De resultaten wezen duidelijk op de superioriteit van de leraren Nederlands: niet alleen bleken de neerlandici veel stabiel in hun oordeel (gemiddelde afwijking tussen de eerste en tweede beoordeling: .64) dan de leraren wiskunde (gemiddelde afwijking: .81) of de managers (gemiddelde afwijking: .86); ook stemden ze onderling meer met elkaar overeen.

Ook al leiden deze opmerkingen wellicht tot nuancering van het karikaturale beeld van de onbetrouwbaarheid van de opstelbeoordelaar, dat neemt niet weg dat het in het onderwijs vigerende 'ideaal van gelijke kansen' vereist dat we ernaar moeten streven verschillen tussen de beoordelingen van dezelfde leerkracht alsook verschillen tussen docenten onderling te minimaliseren. Mogen we nu verwachten dat de beoordeling van schrijfdossier en gedocumenteerde tekst betrouwbaarder zal verlopen dan

die van het traditionele opstel? Een positief antwoord hierop is afhankelijk van de vraag of storende effecten die bij de beoordeling van opstellen een contaminerende rol spelen, al dan niet zijn uitgeschakeld bij de beoordeling van schrijfdossier en gedocumenteerde tekst.

3 Storende effecten; mogelijke maatregelen

Een voor de hand liggende factor die tot een lage intersubjectieve overeenstemming tussen beoordelaars leidt, is het *signifisch effect*: verschillen in de opvatting (betekenis) van de beoordelingstaak. De een legt in zijn oordeel de nadruk op correct en verzorgd Nederlands, een tweede acht de structuur van doorslaggevend belang voor de kwaliteit, een derde is allergisch voor spelfouten, enzovoort. Bij gebrek aan theorie over relevante kenmerken waaraan een goed schrijfproduct zou (moeten) voldoen, heeft elke beoordelaar een zekere mate van vrijheid om in zijn beoordeling zijn eigen particuliere voorkeuren voor bepaalde tekstkenmerken (of voor het relatieve gewicht van bepaalde kenmerken) te benadrukken, met als onvermijdelijk gevolg dat verschillende beoordelaars van hetzelfde product andere waarderingen toekennen. Deze vrijheid die een beoordelaar heeft bij de zogenaamde globale beoordeling wanneer hij volgens eigen maatstaven en normen opstellen beoordeelt, kan met behulp van een *analytisch beoordelingschema* aan banden gelegd worden: hierbij moet een beoordelaar aan de hand van een lijstje met vooraf gespecificeerde tekstkenmerken als 'stijl', 'taalgebruik', 'opbouw' enzovoort de opstellen analyseren en beoordelen. Maar anders dan verwacht, lijkt analytische beoordeling nauwelijks tot betrouwbaarheidswinst te leiden (Meuffels 1994). Een ander tot nu toe onopgelost probleem betreft de vraag welke tekstkenmerken precies, en hoeveel, in het analytisch beoordelingsvoorschrift moeten worden opgenomen, en hoe de oordelen over al deze tekstkenmerken tot een totaaloordeel gecombineerd moeten worden.

Een tweede factor die verantwoordelijk kan zijn voor de divergentie in oordelen, dus voor onbetrouwbaarheid, is het *halo-effect*: de beoordelaar laat zich, bewust dan wel onbewust, bij zijn oordeelsvorming leiden door zijn (voor)kennis van de schrijver in kwestie. Is deze schrijver een joviale leerling die zich tijdens het schooljaar coöperatief heeft opgesteld, dan zal de beoordeling stellig anders uitpakken dan wanneer het om het opstel van een recalcitrante leerling zou gaan, ook al leveren beiden een uit objectief oogpunt gelijkwaardige prestatie. In theorie kunnen een aantal effectieve maatregelen tegen dit type storend effect opgesomd worden (bijvoorbeeld: ont-doe alle opstellen van voor de beoordelaar bekende en relevante informatie over de schrijvers, of laat alle opstellen beoordelen door volstrekt onafhankelijke beoordelaars die onbekend zijn met de leerlingen in kwestie), maar in de praktijk blijken dergelijke maatregelen meestal onhaalbaar.

Een ander type effect dat een negatieve invloed heeft op de betrouwbaarheid van opstelbeoordeling is de *persoonlijke vergelijking*. Daarmee wordt de voor een beoordelaar min of meer karakteristieke wijze van gebruik van onze cijferschaal bedoeld: de een beoordeelt streng, de ander mild; de een geeft extreme oordelen, de ander brengt slechts kleine nuances in zijn oordelen tot uitdrukking, enzovoort. Ter bestrijding van dit effect hebben buitenlandse onderzoekers, vooral in Zweden en de VS,

voorgesteld om (schrijf)prestaties te beoordelen volgens een vaste, normale verdeling: 1% van de werkstukken moet een 10 krijgen, 6% een 9, 24% een 8, 38% een 7, 24% een 6, 6% een 5 en 1% een 4. Op deze manier worden verschillen tussen beoordelaars in gemiddelde, spreiding en percentages onvoldoende rigoureus tegengegaan. Maar dit voorstel, dat tegenwoordig nog steeds in het Amerikaanse onderwijssysteem wordt toegepast onder de naam 'grading on the curve', tast de validiteit (d.w.z. de deugdelijkheid) van de beoordeling aan: een beoordelaar moet soms differentiaties aanbrengen waar hij ze niet ziet en - omgekeerd - verschillen veronachtzamen die hij wel degelijk aanwezig acht.

Ter bestrijding van de persoonlijke vergelijking zoekt men tegenwoordig meer zijn heil in schaalbeoordeling, met voorshands licht positieve resultaten (Schoonen 1991). Bij schaalbeoordeling moet een beoordelaar een opstel niet als een op zichzelf staand iets beoordelen, maar dit vergelijken met een reeks in kwaliteit oplopende voorbeeldopstellen (meestal vijf, soms ook drie) die al voorzien zijn van een cijfer. De taak van de beoordelaar bestaat eruit het te beoordelen opstel als het ware een plaats op die schaal te geven. Net zoals bij analytische beoordeling en bij het 'grading on the curve' wordt ook bij schaalbeoordeling de vrijheid van een beoordelaar ingeperkt: elk product moet aan de hand van de gegeven voorbeelden, die al voorzien zijn van een cijfer, beoordeeld worden. Verwacht mag worden dat beoordelaars, doordat ze min of meer van hetzelfde interpretatiekader uitgaan, minder op 'drift' raken zodat de intersubjectieve overeenstemming toeneemt, een verwachting die enigszins bevestigd lijkt te worden door resultaten van empirisch onderzoek. Een praktisch nadeel van schaalbeoordeling is echter, zeker in het licht van de geringe betrouwbaarheidswinst die daarmee te behalen valt, dat voor elk te beoordelen tekstkenmerk en voor elke opdracht telkens nieuwe beoordelingsschalen geconstrueerd moeten worden, een nogal bewerkelijke klus. Per type opdracht moet immers een groot aantal 'proefteksten' verzameld worden die, na te zijn beoordeeld door een jury van deskundigen, kunnen dienen als uitgangspunt van de schaalconstructie. In didactisch opzicht is het bovendien problematisch – wil schaalbeoordeling effectief zijn – dat alle leerlingen over precies hetzelfde onderwerp moeten schrijven.

Weer een ander type effect dat de betrouwbaarheid van de beoordeling kan aantasten, is het *contaminatie-effect in engere zin*. Een docent is, zeker wat het afsluitende examen betreft, direct belanghebbende bij de uitslag en het resultaat van zijn onderwijs; in deze zin is zijn beoordeling gecontamineerd (in engere zin). Voor een leerkracht die zich heeft ingezet en zich moeite heeft getroost om de schrijfvaardigheid van zijn leerlingen op een hoger peil te brengen, is het in psychologisch opzicht moeilijk de werkstukken op het eindexamen strikt onbevangen, onpartijdig en onbevooroordeeld te beoordelen: zijn eigen status en goede naam zijn in het geding. Waarschijnlijk vallen zijn oordelen te positief uit. Afdoende maatregelen ter bestrijding van dit effect (zoals inschakeling van onpartijdige beoordelaars die geen enkel persoonlijk belang bij de uitslag van een beoordeling hebben) zijn in theorie weer gemakkelijker bedacht dan dat ze in de weerbarstige praktijk van alledag kunnen worden uitgevoerd.

Naast het signifiisch effect, het halo-effect, de persoonlijke vergelijking en het contaminatie-effect (in engere zin) kunnen nog tal van andere effecten de kwaliteit

van het opsteloordeel aantasten: *normverschuiving* (oordelen zijn niet stabiel, maar passen zich aan aan het niveau van de groep te beoordelen producten als geheel), *volgorde-effecten* (na vier opstellen van abominabele kwaliteit slaakt de beoordelaar bij een vijfde, middelmatig opstel een zucht van verlichting en kent dit product een te hoge waardering toe), *contaminatie in ruimere zin* (een collega die de opstellen ook beoordeelt, ziet allerlei aantekeningen van de eerste corrector - de oordelen zijn kortom niet onafhankelijk van elkaar) en het *competentie-effect* (verschillen in beoordelingsbekwaamheid). Het zou in dit kort bestek te ver voeren al deze storende effecten te bespreken en een opsomming te geven van maatregelen die getroffen kunnen worden om hun storende invloed te minimaliseren. Daarom volstaan we hier met de constatering dat

- (1) de existentie van de hierboven genoemde storende factoren in empirisch onderzoek naar de betrouwbaarheid en validiteit van opstelbeoordeling is aangetoond (Meuffels 1994)
- (2) de invloed van deze storende factoren op de validiteit (d.w.z. de deugdelijkheid) van de beoordeling onder alle omstandigheden negatief is, en de invloed op de betrouwbaarheid doorgaans negatief
- (3) er geen algemene, generieke maatregel bestaat die alle onderscheiden effecten tegelijkertijd tenietdoet: elk type storend effect moet op zijn eigen, specifieke wijze bestreden worden.

Naast deze systematisch-storende factoren zijn er natuurlijk nog tal van andere, min of meer 'triviale' factoren op te sommen die afbreuk kunnen doen aan de kwaliteit van het oordeel: iemands stemming, humeur, de mate van vermoeidheid, interesse, enzovoort. Dit soort factoren echter opereert, anders dan het gros van de hiervoor genoemde factoren, grotendeels op bewust niveau: ieder weldenkend mens wéét immers dat je geen opstellen moet beoordelen wanneer je extreem vermoeid bent, of slaapdrunk, heftig geëmotioneerd, enzovoort. Een beoordelaar kan zichzelf dus gemakkelijk wapenen en effectieve voorzorgsmaatregelen treffen tegen dit type storende effecten. Maar welke maatregelen zou hij in 's hemelsnaam moeten treffen om het effect van, bijvoorbeeld, contaminatie in engere zin teniet te doen - gesteld al dat hij zich bewust is van het bestaan ervan?

Ligt het nu voor de hand om bij het schrijfdossier en de gedocumenteerde tekst een hogere beoordelaarsbetrouwbaarheid te verwachten dan bij het traditionele opstel? Tot op heden ontbreken 'harde' empirische gegevens over de beoordelaarsbetrouwbaarheid die bij dit type opdrachten haalbaar is. Dat laat echter onverlet dat het - gezien het algemene, systematische karakter van de onderscheiden storende effecten - niet realistisch is om te veronderstellen dat een hogere betrouwbaarheid *zonder meer* (dat wil zeggen zonder extra inspanningen om specifieke storende effecten te minimaliseren) tot de mogelijkheden behoort. De tweede vorm van onbetrouwbaarheid leidt echter tot meer positieve conclusies voor het schrijfdossier en de gedocumenteerde tekst: vergeleken met het 'traditionele' opstel is daarbij inderdaad sprake van een substantiële verbetering.

4 Scorebetrouwbaarheid

Scorebetrouwbaarheid is ongetwijfeld wat minder bekend bij 'het grote publiek' dan de hierboven besproken beoordelaarsbetrouwbaarheid. Dat is niet zo verwonderlijk als men beseft dat beschouwingen over deze vorm van betrouwbaarheid vaak nogal technisch-esoterisch van karakter zijn en als gevolg daarvan grotendeels beperkt blijven tot wat genoemd wordt 'de kring van vakgenoten'. De praktische consequenties zijn bij deze vorm van onbetrouwbaarheid echter niet minder dramatisch dan die bij de beoordelaarsonbetrouwbaarheid.

Waar gaat het bij scorebetrouwbaarheid om? Schrijfvaardigheid is een hypothetisch construct: wanneer wij iemand schrijfvaardigheid toeschrijven, dan veronderstellen wij bij die persoon een dispositie (een attribuut) aanwezig die aan het geobserveerde gedrag ten grondslag ligt en die zich onder vergelijkbare omstandigheden ook op een andere tijd en plaats zal manifesteren. Met andere woorden: wie het begrip 'schrijfvaardigheid' hanteert en iemand een bepaald niveau in die vaardigheid toeschrijft, doet daarmee impliciet dan wel expliciet een *voorspellende* uitspraak over hoe de persoon in kwestie in soortgelijke situaties en omstandigheden zal presteren. In het empirisch onderzoek naar schrijfvaardigheid van de afgelopen decennia wordt dan ook in het gros van de gevallen uitgegaan van de (niet-getoetste) aanname dat de gegevens die verkregen zijn met behulp van één schrijfo opdracht zonder meer geldig zullen zijn voor alle, soortgelijke opdrachten.

Deze cruciale aanname staat of valt echter met de hoogte van de correlatie (d.w.z. de mate van samenhang) tussen de prestaties op verschillende schrijftaken. Als de meting van schrijfvaardigheid van taak tot taak weinig varieert en de metingen onderling consistent zijn, dan levert dat een adequate, precieze indicatie op van iemands schrijfvaardigheid (dus een hoge scorebetrouwbaarheid). Als deze metingen echter van taak tot taak sterk uiteenlopen, dan kunnen bezwaarlijk generaliserende uitspraken gedaan worden over 'de' schrijfvaardigheid van een persoon (dus een lage scorebetrouwbaarheid).

In diverse studies is de samenhang tussen verschillende schrijfo opdrachten nader onderzocht, met als onveranderlijke conclusie dat de correlatie tussen prestaties op verschillende schrijfo opdrachten bedroevend laag is: prestaties op de ene taak hebben vrijwel *geén* voorspellende waarde voor de prestaties op een andere taak (Schoonen 1991). Deze geringe voorspellende waarde doet zich gelden ongeacht het beoordeelde aspect ('inhoud', 'structuur' of 'stijl') en ongeacht de wijze van beoordelen (globale beoordeling, analytische beoordeling of schaalbeoordeling).

Voor die lage scorebetrouwbaarheid zijn een aantal verklaringen geopperd. Een verklaring schrijft deze toe aan de omstandigheid dat een schrijftaak als een opstel eigenlijk uit slechts één testitem bestaat. Kan een complexe vaardigheid als schrijfvaardigheid ooit betrouwbaar gemeten worden met behulp van slechts één testitem? Kan men, om een parallel te trekken, concluderen dat iemand over onvoldoende rekenvaardigheid beschikt als hij niet binnen vijf seconden kan uitrekenen hoeveel 23 maal 46 is? Natuurlijk niet. In de eerste plaats wordt met deze ene vraag/opgave wel een bijzonder beperkt deel van mogelijke rekenvaardigheden gemeten: de steekproef

aan gemeten vaardigheden is zo klein dat generalisatie naar alle rekenvaardigheden uitgesloten is. In de tweede plaats kan een vaardigheid als rekenen met behulp van één enkele opgave onmogelijk op betrouwbare wijze gemeten worden. Aangenomen dat dit alles voor rekenvaardigheid geldt, is dit dan niet a fortiori van toepassing op zo'n complexe vaardigheid als schrijven? Vanuit het hier geschetste perspectief manoeuvreert een enkel opstel "*the student in a position of a gambler who risks all on a single throw of the dice*" (Noyes 1945). Kortom, een persoon die goed is in het schrijven van verhalende opstellen, hoeft nog helemaal geen meester te zijn in het schrijven van betogen. Afhankelijk van de specifieke tekstsoort, de communicatieve taak, het onderwerp enzovoort, fluctueren de schrijfsprestaties van een individu en dat maakt dat men van diens prestatieniveau geen adequaat beeld verkrijgt wanneer men dat beeld baseert op slechts één opstel. Pas wanneer een schrijver meer teksten schrijft, mag een redelijke betrouwbaarheid verwacht worden. Schattingen van het aantal benodigde taken laten onveranderlijk zien dat er enorme aantallen schrijfoopdrachten afgenomen moeten worden voordat van een betrouwbare meting gesproken kan worden. Deze schattingen lopen uiteen van elf tot liefst veertig!

Al met al kan gesteld worden dat met de introductie van het schrijfdossier een eerste, bescheiden stap is gezet om te voldoen aan in ieder geval de minimumeisen voor een meting van schrijfvaardigheid.

Literatuur

- Meuffels, B. (1994). *De verguisde beoordelaar*; opstellen over opstelbeoordeling. Amsterdam: Thesis Publishers.
- Noyes, E.S. (1945). *Report on the first six tests in English composition*. New York: College Entrance Examination Board.
- Schoonen, R. (1991) *De evaluatie van taalvaardigheidsmetingen*. Diss. UvA.
- Schröter, G. (1971). *Die ungerechte Aufsatzzensur*. Bochum: Verlag Kamp.
- Wesdorp, H. (1983). Gelijke kansen in centrale examens. In: W.C. Weeda (red.), *Examens in discussie: Een bundel opstellen voor J. W. Solberg*. Groningen: Wolters-Noordhoff.