F. ZONDERVAN

JURYBEOORDELING ALS VALIDERINGSCRITERIUM BIJ HET METEN VAN SCHRIJF- EN LEESVAARDIGHEID

Samenvatting

Schrijf- en leesvaardigheid zijn vaag omschreven theoretische begrippen. Om ze te kunnen meten maakt men gebruik van een handelingsvoorschrift, een operationalisering van het begrip-zoals-bedoeld. Van dit handelingsvoorschrift dient men na te gaan of men daarmee inderdaad meet wat men wenst te meten. Met andere woorden men dient de theoretische validiteit ervan te bepalen. We beperken ons hier, bij het meten van schrijf- en leesvaardigheid, tot zakelijke teksten.
Bij schrijfvaardigheidsmeting tracht men na te gaan of het schrijfprodukt bepaalde kwaliteiten bezit. Men kan trachten vast te stellen of het stuk 'goed gebouwd' is, of het 'prettig leesbaar' is, of het 'goed gememoriseerd' kan worden, of het 'goed is samen te vatten'. Hierdoor kan men een indruk krijgen van de kwaliteiten van het schrijfprodukt.
Bij leesvaardigheidsmeting tracht men na te gaan of bepaalde leesdoelen zijn bereikt. Hiertoe bestaan veel verschillende toetsvormen. Het vergelijken van de verschillende toetsresultaten biedt een goede mogelijkheid tot validering.
Een noodzakelijke voorwaarde voor theoretische of begripsvaliditeit van een operationalisering is de betrouwbaarheid van de operationalisering, waarvan verschillende soorten onderscheiden kunnen worden.
Van jurybeoordeling door deskundigen is herhaald vastgesteld dat de betrouwbaarheid hoog is. Deze voorwaarde is noodzakelijk, maar is hij ook voldoende als valideringscriterium bij het meten van schrijf- en leesvaardigheid?
In het onderstaande zal ik laten zien op grond van welke methodologische overwegingen deze (retorische) vraag met *nee* beantwoord moet worden.

Inleiding

De onderhavige lezing was in eerste opzet bedoeld als inleiding voor kandidaten Nederlands in Utrecht, die het eerste semester '78/79 een college over dit onderwerp gaan volgen. De uiteindelijke opzet is afgestemd op een wat breder publiek van Neerlandici met de bedoeling te laten zien welke methodologische aspecten van onderzoek relevant zijn bij het beoordelen van schriftelijk geformuleerde antwoorden op toetsvragen. Het lag niet in de bedoeling een overzicht van de vakliteratuur te geven die na Wesdorp (1974) is verschenen, hoewel die literatuur, en de door Wesdorp niet geraadpleegde Duitse studies[1] zoals Ingenkamp (1970), en recente Duitse studies zoals Tscherner & Masendorf (1974), Heller (1974), Nickel & Wieczerkowski (1974) en Baurmann (1975), de moeite van het bestuderen waard lijken. Een verslag van de literatuurstudie in het kader van het kandidatencollege in Utrecht zal te zijner tijd in de instituutsreeks van 'De Vooys' worden ondergebracht.

Kunstrijden op de schaats en beoordeling van iemands taalvaardigheid

Bij wedstrijden kunstrijden op de schaats worden twee onderdelen gereden: 1. de verplichte figuren, die elk een verschillende moeilijkheidsgraad hebben en 2. de Kür, een in harmonie met zelfgekozen muziek gereden geheel bestaande uit sprongen, pirouetten, passen en andere verbindende bewegingen naar eigen keuze. De duur van de Kür bedraagt voor mannen 5 minuten en voor vrouwen 4. De Kür wordt door een deskundige, internationale jury beoordeeld, waarbij elk jurylid onafhankelijk van zijn medeleden, op het teken van de voorzitter, een bord omhoog steekt, waarop het cijfer staat dat hij toekent voor de kwaliteit van de zo juist verreden Kür. Sommering van de individuele cijfers, minus het hoogste en minus het laagste cijfer leveren de jurywaardeing op van de kwaliteit van de Kür.
Dit voorbeeld over kunstschaatsen kan op verschillende onderdelen worden vergeleken met het bepalen van iemands taalvaardigheid. Het rijden van de verplichte figuren in de wedstrijdsituatie is vergelijkbaar met het beantwoorden van meerkeuzevragen (waarvan vaak beweerd wordt, dat ze geen z.g. hogere vaardigheden meten) van bij voorbeeld tekstbegrip in de examensituatie: er kan namelijk objectief worden vastgesteld of een kandidaat een verplichte figuur foutloos heeft gereden en of een kandidaat een meerkeuzevraag foutloos heeft beantwoord. In dit onderdeel van het kunstrijden wordt de techniek van de kandidaten gemeten. Van het beantwoorden van meerkeuzevragen kan echter niet zonder meer gezegd worden dat de techniek van het interpreteren of zo u wilt de techniek van iemands taalvaardigheid wordt gemeten. Hier gaat de vergelijking dus mank.
Maar de vergelijking met het kunstrijden is hiermee niet uitgeput: het rijden van een Kür kan vergeleken worden met het schrijven van een opstel. Men zou kunnen zeggen: het schrijven van een opstel is het in harmonie met een zelfgekozen inhoud samenvoegen van paragrafen, alinea's, zinnen en verbindende elementen tot één geheel.
Ook van dit onderdeel van iemands taalvaardigheid moet de kwaliteit worden vastgesteld in de examensituatie. De vraag is echter: moet dat door middel van een deskundige, onafhankelijk van elkaar, oordelende jury? Kan het ook anders, en waar moet men bij een keuze uit verschillende mogelijkheden dan op letten? Zo er al een keuze gemaakt moet worden.
Als we ons eerst afvragen of de kwaliteit van het rijden van een Kür per se door een jury bepaald moet worden, kunnen we daarna, parallel aan de beantwoording van deze vraag, nagaan of de kwaliteit van iemands schrijfvaardigheid per se door een jury moet geschieden.
Aan welke eisen moet een goede Kür voldoen? *Een mannelijke kandidaat moet zich gedurende vijf minuten schaatsend voortbewegen, waarbij hij een boeiende afwisseling van sprongen, pirouetten, passen en andere verbindende bewegingen uitvoert op de maat van zelf gekozen muziek, en wel zodanig, dat er een harmonieus geheel ontstaat.* Met deze omschrijving is nog niet precies duidelijk wanneer een examinator of jurylid kan zeggen, dàt een concrete Kür eraan beantwoordt en in welke mate hij eraan beantwoordt.
In principe kan iedere leek vaststellen of er sprake is van afwisseling van sprongen, passen etcetera op de maat van de muziek, maar een deskundige zal moeten vaststellen of de afwisselende sprongen, passen etcetera technisch correct zijn uitgevoerd. Daar komt bij dat de begrippen *boeiende afwisseling* en *harmonieus geheel* zich door de adjectieven boeiend en harmonieus niet zo gemakkelijk laten concretiseren. Zeker, de vereiste deskundigheid om uit te kunnen

maken in welke mate er sprake is van technische volmaaktheid van de uitgevoerde sprongen, passen etcetera, zal een steun zijn bij het vaststellen van de mate van boeiendheid en harmonie. Maar ik denk dat het ook voor iedereen aanvaardbaar is als ik zeg dat een zekere mate van subjectiviteit in de bepaling ervan haast niet is uit te sluiten. Wel zal oefening in het zien van boeiende en minder boeiende afwisselingen en oefening in het zien van harmonieuze en minder hamonieuze gehelen de beoordelaar een zekere deskundigheid verschaffen. Deze deskundigheid door oefening verkregen, zal echter nooit zo groot kunnen worden dat willekeurige geoefende beoordelaars het niet meer met elkaar *oneens* zullen zijn. Met andere woorden het is in het geval van kunstrijden voor wat betreft het onderdeel Kür niet goed mogelijk de vereisten voor kwaliteit zo te formuleren dat zelfs een deskundige op het gebied van het kunstrijden *eenduidig* de kwaliteit van een Kür in een cijfer kan uitdrukken. dat de kwaliteit van een Kür in een cijfer kan worden uitgedrukt, zal wel niemand willen betwijfelen.
In een wedstrijdsituatie zal men echter eenduidig de rangorde van de deelnemende kandidaten willen kunnen vaststellen. Bij het onderdeel *verplichte figuren* is dat nauwelijks een probleem, voor zover ik dat als leek op het gebied van het kunstrijden kan beoordelen. Bij het onderdeel *Kür* hoeft een enkele examinator ook geen probleem op te leveren: alle kandidaten komen onder zijn ogen en zijn oordeel zou het eindoordeel kunnen zijn. De toeschouwers zijn bijna altijd minder deskundig en zullen bij kritiek op de uitslag daardoor geen recht van spreken hebben. Het oordeel van deze ene examinator kan echter door andere deskundigen aangevochten worden. Zij kunnen te kennen geven dat de gegeven cijfers niet weergeven wat er bedoeld is: de juiste rangorde in de kwaliteit van de kunstschaatsprestaties.
In de praktijk van de kampioenschappen kunstrijden op de schaats zal men het risico van zo'n eenmansbeoordeling te groot achten en het zekere voor het onzekere willen nemen door een deskundige jury een gezamenlijk oordeel te laten vellen. In theorie is het dan nog altijd mogelijk dat een andere groep deskundigen tot een ander oordeel zou zijn gekomen. Een confrontatie van de verschillende kwaliteitseisen van beide jury's kan er dan toe bijdragen dat er meer helderheid wordt verkregen over de kwaliteitseisen die men aan het rijden van een Kür behoort te stellen.

Operationaliseren, begripsvaliditeit en betrouwbaarheid

In het voorgaande zijn de nu volgende vaktermen zonder genoemd te zijn toch al gebruikt: 1. *operationaliseren:* het zodanig formuleren van een begrip in operaties (handelingen), dat door toepassing daarvan precies kan worden bepaald of, en zo ja, in welke mate er sprake is van dat begrip; 2. *begripsvaliditeit:* het aan de orde stellen van de vraag of de beschrijving van de handelingen die voor de meting van het begrip verricht moeten worden, aan zijn doel beantwoordt. Met andere woorden, meet men met behulp van de operationalisering datgene wat men wil meten?; 3. *betrouwbaarheid van de meting:* hierbij maken we onderscheid tussen a) inter-examiner-betrouwbaarheid (de overeenstemming tussen beoordelaars) en b) intra-examiner-betrouwbaarheid (de overeenstemming binnen één beoordelaar, als hij een prestatie volgens het handelingsvoorschrift twee keer moet beoordelen). Deze overeenstemming tussen en binnen beoordelaars zegt niets over het feit of de metingen al dan niet aan een vooropgesteld doel beantwoorden, dus niets over de begripsvaliditeit van de volgens de

operationalisering tot stand gekomen oordelen.
Bij een beschouwing van verschillende toetsvormen spelen de begrippen operationaliseren, begripsvaliditeit en betrouwbaarheid een belangrijke rol. Toetsvormen zijn operationaliseringen van vaardigheden (abstracte begrippen); toetsvormen zijn de instrumenten waarmee men de mate bepaalt waarin een bepaalde vaardigheid aanwezig is. Een uitspraak over de waarde van de meting is afhankelijk van de validiteit en de betrouwbaarheid ervan.

Schrijf- en leesvaardigheid

In de titel worden schrijf- en leesvaardigheid in één adem genoemd. Waarom jurybeoordeling in verband gebracht wordt met schrijfvaardigheid, ligt voor de hand: er moeten schrijfprodukten worden beoordeeld en de overeenstemming tussen beoordelaars schiet over het algemeen tekort. Waarom jurybeoordeling ook met leesvaardigheid in verband gebracht wordt, komt doordat leesvaardigheid nogal eens geoperationaliseerd wordt met behulp van essay-vragen (korte en/of lange antwoordvragen), zodat ook hier schrijfprodukten moeten worden beoordeeld.
Hoe moet iemands schrijfvaardigheid c.q. leesvaardigheid nu worden vastgesteld? Hiervoor is het nodig dat de vaardigheid operationeel wordt gedefinieerd, liefst zo concreet dat een deskundige aan de hand van de gegeven omschrijving kan bepalen òf en zo ja, in welke mate die vaardigheid bij iemand ontwikkeld is. Met andere woorden er wordt een goede operationalisering verlangd.
Een voorbeeld van een te abstract gebleven operationalisering: *Iemand geeft van een voldoende schrijfvaardigheid blijk als hij over een zelfgekozen inhoud paragrafen, alinea's, zinnen en verbindende elementen zodanig op papier kan zetten dat een boeiend, samenhangend en coherent geheel ontstaat.*
Ook hier zullen deskundigen het eens kunnen worden over wat men zou kunnen noemen de 'technische aspecten' van schrijfvaardigheid, zoals daar zijn: een goede spelling, grammaticale zinnen kunnen schrijven, een logische opbouw in een betoog kunnen maken, geen feitelijke onjuistheden vermelden en meer van dergelijke eenvoudig vaststelbare kenmerken van schrijfprodukten. Een dergelijke techniek zou in een 'verplichte-figuren-onderdeel' apart kunnen worden getoetst. Het moet mogelijk zijn hierover voorstellen te doen, waarin staat welke elementen erin zouden moeten worden opgenomen en wat het relatief gewicht van die elementen is. Bovendien moet het mogelijk zijn dat één van dergelijke voorstellen een meerderheid van de huidige Neerlandici mee krijgt.
Naast het 'verplichte-figuren-onderdeel' zal het vrije, creatieve onderdeel moeten staan, waarbij het op speelse wijze kunnen omgaan met de verworven techniek beoordeeld moet worden: originaliteit, stijl, belangwekkendheid, nieuwswaarde, boeiendheid enzovoort. Hierbij zijn verschillende opsommingen mogelijk van allerlei begrippen die stuk voor stuk aspect kunnen zijn waaronder beoordeeld moet worden. In Heller (1974) bij voorbeeld wordt melding gemaakt van vijf dimensies die bij opstelbeoordeling empirisch zouden zijn aangetoond: "Ideen, innere Form (Gliederung), Lebendigkeit (Originalität), Sprachrichtigkeit und Wortwahl (Flüssigkeit)." De dimensies 'Gliederung' en 'Sprachrichtigkeit' zijn hier de meer technische aspecten, de overige drie dimensies moeten tot de meer vrije vaardigheden van een schrijver gerekend worden. Om echter met dergelijke begrippen in de dagelijkse lespraktijk uit de voeten te kunnen is een nauwkeurige operationalisering ervan noodzakelijk.

Toetsvormen

De mate waarin bovengenoemde schrijfvaardigheidsaspecten op positieve wijze aanwezig kunnen zijn in te beoordelen schrijfprodukten, kan op verschillende manieren beoordeeld worden: a) door slechts één beoordelaar; b) door een jury van beoordelaars.

Binnen groep a) kan men de volgende methoden onderscheiden:
1. beoordeling met behulp van een door voorbeeldopstellen geijkte schaal. De voorbeeldopstellen staan voor wat betreft één of meer aspecten al op een cijferschaal gerangschikt. De beoordelaar plaatst de te beoordelen opstellen vergelijkenderwijze op deze schaal.
2. beoordeling met een door het slechtste en het beste opstel (voor één of meer aspecten) geijkte schaal: het slechtste opstel krijgt het cijfer 1, het beste cijfer 10. Hiertussen moeten de overige te beoordelen opstellen volgens hun kwaliteit worden geplaatst.
3. beoordeling met behulp van een analytisch schema: hierin zijn de verschillende, relevant geachte, aspecten zodanig geoperationaliseerd dat per aspect een bepaald aantal punten kan worden gegeven als in een opstel die aspecten in bepaalde mate aanwezig zijn.
4. beoordeling aan de hand van ervaring en intuïtie: min of meer analytisch of globaal.

Binnen groep b) kan men de volgende methoden onderscheiden:
1. beoordeling met door voorbeeldopstellen geijkte schalen.
2. beoordeling met een schaal die 'geijkt' is door de twee meest extreme opstellen uit de te beoordelen serie.
3. beoordeling met een analytisch schema.
4. beoordeling via paarsgewijze vergelijking: deze in Mellenbergh (1974) beschreven methode lijkt zeer bevredigende resultaten op te leveren qua betrouwbaarheid, is echter nogal arbeidsintensief. Wanneer bij voorbeeld 10 opstellen met deze methode beoordeeld moeten worden, zijn er per beoordelaar 45 verschillende tekstparen te vergelijken. Er moet per tekstpaar een nul of een één worden toegekend. Het maximaal te bepalen aantal punten per opstel hangt af van het aantal juryleden. Voor de lespraktijk van de leraar Nederlands is deze methode dus niet van praktisch nut. Zij kan echter in onderzoeksituaties gebruikt worden om betrouwbare resultaten te verkrijgen.
5. beoordeling aan de hand van ervaring en intuïtie: min of meer analytisch of globaal.

Het hier gegeven overzicht is niet volledig. Het is immers op veel meer manieren mogelijk schrijfvaardigheid te operationaliseren. Men denke maar aan de zogenaamde objectieve schrijfvaardigheidstoetsen waarvan in Wesdorp (1974) melding wordt gemaakt, of aan een operationalisering van schrijfvaardigheid door middel van een cloze-test die bij een steekproef van het bedoelde publiek zou kunnen worden afgenomen.

Begripsvaliditeit en betrouwbaarheid

Zoals hierboven is aangegeven is het mogelijk op verschillende manieren het begrip schrijfvaardigheid te operationaliseren. Welke van deze manieren is de

beste? Om hierop te kunnen antwoorden hebben we meer informatie nodig over de begripsvaliditeit en de betrouwbaarheid van de metingen volgens de verschillende methoden. Voorop staat de begripsvaliditeit van de toetsvorm: een theoretische verhandeling moet aan een breed publiek van deskundigen (leraren Nederlands) aannemelijk maken, waarom bepaalde aspecten relevant geacht worden voor een goede schrijfvaardigheid, hoe en in welke verhouding ze gemeten zullen moeten worden. Dat is echter makkelijker gezegd dan gedaan. De leraren Nederlands hanteerden tot nu toe zó zeer verschillende doelstellingen, dat de enquêtering over de doelstellingen die zij noodzakelijk achten in het moedertaalonderwijs, geen bevredigende resultaten heeft opgeleverd in het SVO-project van het RITP.
Een mogelijkheid om uit deze impasse te geraken en tot een hogere overeenstemming tussen leraren Nederlands te komen over het begrip schrijfvaardigheid, is hen te overtuigen via een goed opgebouwde cursus schrijven, waarin een leerpsychologisch verantwoorde schrijfprocedure wordt gegeven. Ook deze weg heeft tot nu toe niet tot de nodige overeenstemming geleid. Voorlopig ziet het er dus naar uit, dat de vraag naar de begripsvaliditeit van schrijfvaardigheid niet bevredigend kan worden beantwoord. We zullen het dus moeten doen met de relatief lage begripsvalide methoden zoals hierboven opgesomd.
De begripsvaliditeit is afhankelijk van de betrouwbaarheid van de operationaliseringen: is de betrouwbaarheid van de meetmethoden laag, meet de ene beoordelaar iets anders dan de andere, of meet dezelfde beoordelaar bij een herhaalde beoordeling iets anders, dan kan de begripsvaliditeit van de operationalisering niet hoog zijn. Is de betrouwbaarheid echter hoog, zijn de beoordelaars stabiel in hun oordeel en zijn zij het onderling in hoge mate eens, dan is dat op zichzelf nog geen garantie voor begripsvalide metingen. Als de operationalisering bij voorbeeld alleen maar beoordeling van de dimensie 'Sprachrichtigkeit' vraagt, kan de betrouwbaarheid zeer hoog zijn, maar de validiteit van het begrip schrijfvaardigheid is hiermee zelfs op het eerste gezicht niet veel waard.
Een hoge betrouwbaarheid is dus een noodzakelijke voorwaarde voor een hoge begripsvaliditeit, omdat daarmee tevens de bovengrens voor de validiteit gegeven is. Tot nu toe heeft het empirisch onderzoek laten zien dat jurybeoordelingen, de toetsvormen uit categorie b), betrouwbaarder meten dan beoordelingen door slechts één beoordelaar, de toetsvormen onder categorie a).
Voorlopig zal in onderzoek naar schrijf- en leesvaardigheid nog geen keuze gemaakt mogen worden uit de hiervoor opgesomde toetsvormen. Het is aan te bevelen zoveel mogelijk verschillende toetsvormen in zowel onderwijs als onderzoek te gebruiken, opdat gegevens verkregen worden voor de evaluatie van zowel de betrouwbaarheid als de begripsvaliditeit van operationaliseringen van de begrippen schrijf- en leesvaardigheid.

Afdeling Taalbeheersing
Instituut "De Vooys"
Rijksuniversiteit Utrecht

Noot

(1) Dat Wesdorp de Duitse literatuur van vóor 1974 niet heeft geraadpleegd, is geen omissie in zijn dissertatie, omdat Ingenkamp in 1970, voorlopig als eenling, in een overzichtartikel het probleemveld voor de Duitse onderzoekers en leraren beschrijft en daarbij ook teruggaat op de door Wesdorp bestudeerde Angelsaksische literatuur.

Literatuur

Bauermann, J.
1975 'Aufsatzbenotung und Reihenfolge-Effekt', *Psychologie in Erziehung und Unterricht*, 22, p. 181 22, p. 181-185.

Heller, K.
1974 'Zur Problematik der Leistungsbeurteiling in der Schule', *Psychologie in Erziehung und Unterricht*, 21, p. 105-124

Ingenkamp, K.
1970 'Zur Problematik der Zensurgebung', *Die Deutsche Schule*, 62, p. 438-456.

Mellenbergh, D.
1974, *Schaalmethoden*, Utrecht, (syllabus van de afdeling M & S. Rijksuniversiteit).

Nickel, H. & W. Wieczerkowski.
1974 'Einflusse auf die Beurteilung von Schüleraufsätzen", K. Heller (ed.). *Leistungsbeurteilung in der Schule*, Heidelberg.

Tscherner, K. & F. Masendoff
1974 'Analyse von Schülerbeurteilungen und Zeugnisnoten bei einzelven Lehrern', *Psychologie in Erziehung und Unterricht*, 21, p. 135-149

Wesdorp, H.
1974 *Het meten van de produktief-schriftelijke taalvaardigheid*, Purmerend.