Jan Van Maele

Instrumenten voor een behoorlijke evaluatie van spreekvaardigheid in de vreemde taal

De natte vinger voorbij

De evaluatie van communicatieve vaardigheden in de vreemde taal is een belangrijke maar delicate kwestie. Tal van studies hebben aangetoond dat beoordelaars grondig van mening kunnen verschillen, wat elk cijfer aanvechtbaar maakt. In dit artikel stel ik een toetsprocedure voor die een behoorlijke, d.w.z. valide en betrouwbare evaluatie van de spreekvaardigheid bij gevorderde studenten nastreeft. In deze procedure neemt de video-opname een centrale plaats in. Deze wordt aangevuld met een beoordelingsschema en een vragenlijst.

Heel lang geleden werd er aan schoolmeesters een onfeilbaar oordeel toegeschreven en, net zoals dit bij priesters het geval is, volgde bekwaamheid automatisch uit wijding. Dit romantische verleden is nu alvast in Vlaanderen vervangen door een situatie waarin een examinator steeds vaker rekenschap en verantwoording verschuldigd is, tot voor de rechtbank (Verstegen 1994). Daarbij toetst de rechter de examenprocedure in vele gevallen aan de beginselen van behoorlijk bestuur, zoals het zorgvuldigheidsbeginsel, het gelijkheidsbeginsel, en het redelijkheidsbeginsel. In toetstaal komt dit erop neer dat een cijfer aangeeft hoe taalvaardig een student is al naar gelang het cijfer *betrouwbaar* is en de interpretatie ervan *valide* (Bachman 1990). Een toets levert betrouwbare scores op wanneer hij nauwkeurig meet. Zo moet een beoordelaar steeds dezelfde punten voor gelijkwaardige antwoorden geven, en moeten beoordelaars die gelijke instructies krijgen ook tot gelijke beoordelingen komen. Een toets is valide in de mate dat er gemeten wordt wat men zegt te meten. Zo boet een spreekvaardigheidstoets waarvan de score wordt meebepaald door acteertalent of vertrouwdheid met het gespreksonderwerp in aan validiteit. Betrouwbaarheid is een noodzakelijke voorwaarde voor validiteit, want een toets kan natuurlijk nooit meten wat hij zegt te meten als hij onnauwkeurig meet.
Naast deze twee hoofdcriteria voor een behoorlijke evaluatie moeten we ook rekening houden met de *praktische uitvoerbaarheid*, en met de *indruksvaliditeit* ('face validity'), die aangeeft of de toets bij de studenten goed overkomt.

Het gesprek als spreekvaardigheidsoefening

Onterechte generalisaties van de behaalde cijfers zijn een vaak voorkomende inbreuk tegen de validiteitsregels. Voorzichtigheid gebiedt te specificeren, maar niemand is geïnteresseerd in al te enge interpretaties ('Je Engelse taalvaardigheid was voldoende om op een druilerige dinsdag acht korte vragen over een voorbereide tekst samenhangend te beantwoorden in een accent dat de leerkracht zonder veel moeite verstond'). Als we op een valide manier toch een zekere generalisatie mogelijk willen maken, doen we er goed aan onze spreekvaardigheidstoets te spiegelen aan de frequentste vorm van verbale communicatie uit het gewone leven, en dat is het gesprek. In tegenstelling tot het interview – nog steeds de klassieke toetsmanier – kenmerkt het gesprek zich door onvoorspelbaar verloop en potentiële symmetrie tussen de gesprekspartners (van Lier 1989). Eigenlijk wordt een toets dus pas echt authentiek op het moment dat beide partijen voor even vergeten dat het om een toets gaat (Underhill 1987). Een belangrijke voorwaarde daarvoor is het vermijden van 'teacher talk': vragen stellen waarop je het antwoord zelf weet, enkel luisteren naar hóe iets gezegd wordt, vermanende blikken bij taalfouten...
Mij lijkt het onmogelijk om de rol van beoordelaar-rechter te combineren met die van gesprekspartner die zich ten volle in de conversatie engageert. De rollen splitsen is een mogelijke uitweg: je verdeelt deze rollen dan over twee personen, die allebei aanwezig zijn, of je verenigt ze binnen één persoon maar dan wel verdeeld in de tijd. Ik koos voor het laatste en gebruik daarbij de *video*.

Video en validiteit

Het 'eindexamen' op video opnemen kan natuurlijk enkel als de studenten al met het medium vertrouwd zijn, m.a.w. de proef moet nauw aansluiten bij de tijdens het jaar gebruikte werkmethode (zie Haggstrom 1994 en Van Maele 1994 voor lesvoorbeelden). Zolang de camera immers als een vreemd object ervaren wordt, zullen nogal wat studenten zich geremd of benauwd voelen, hoezeer je ook probeert ze te stimuleren tot hun beste prestatie. Ook logistieke ingrepen kunnen de plankenkoorts mee helpen reduceren. Zo blijft de camera tijdens de taalvaardigheidsproef op zijn gewone plaats in de hoek van het medialokaal staan, van waaruit hij onbemand het hele tafereel registreert. Op geen enkel ogenblik hoeft de student in de lens te kijken.
Door mijn beoordelingstaak uit te stellen tot een later tijdstip kan ik me tijdens de proef helemaal concentreren op het gesprek zelf. Daarbij probeer ik

de macht minder sterk naar mezelf toe te trekken: ten eerste door de student het *onderwerp* vrij te laten kiezen (hij/zij spreekt dus niet noodzakelijk over iets wat mij boeit of over iets waar ik meer van afweet), en ten tweede door de student mee verantwoordelijkheid te laten dragen voor het *verloop* van het gesprek (hij/zij kan het initiatief nemen van het vraag-antwoordpatroon af te wijken door zelf een vraag te stellen; ik van mijn kant onderbreek de student enkel als ik echt nieuwsgierig naar iets ben, of wanneer de student een gememoriseerd nummertje opvoert, of in mijn rol van tijdwaarnemer).
Dankzij de video herwon ik als leerkracht de vrijheid de mondelinge taalvaardigheidsproef tot een authentieker (want representatiever) gebeuren te promoveren, en die authenticiteit heeft op zijn beurt weer bijgedragen tot een meer valide beoordeling.

Video en betrouwbaarheid

De video als hulpmiddel om de mondelinge taalvaardigheidsproef een authentieker karakter te geven, komt de *validiteit* ten goede. Een ander voordeel van de video is dat de *betrouwbaarheid* van het oordeel erop vooruit kan gaan. Doordat de vluchtigheid van het gesproken woord geen belemmering meer vormt, kan het risico van normverschuiving worden beperkt. Behoorlijk examineren veronderstelt immers dat elke student beoordeeld wordt in relatie tot een vooraf omschreven norm, los van de toevallige differentiatie binnen een bepaalde groep in een bepaald jaar (De Neve & Janssen 1992). In de praktijk wil die norm echter wel eens gaan schuiven: een matig gesprek komt sterker over na enkele uiterst zwakke broertjes; de vermoeidheid begint de beoordelaar parten te spelen; vergelijkbare prestaties krijgen met de jaren een milder (of strenger) waardeoordeel; enzovoort. Met behulp van de video probeer ik de normconstante beter te bewaken:

- door *ijking*: vooraleer ik de proeven beoordeel, bekijk ik eerst een paar cesuurgesprekken (9/20, 10/20, 12/20) uit een vroegere jaargang;
- door *condensatie*: alhoewel de gesprekken zelf over minstens een hele maand gespreid zijn, kan de beoordeling binnen korte tijd worden afgerond;
- door *mengeling*: de volgorde van beoordeling hoeft niet noodzakelijk die van de opname te zijn;
- door *zappen*: bij een vermoeden van normverschuiving grijp ik terug naar al beoordeelde fragmenten op andere banden.

Het analytisch beoordelingsschema

Om de spreekvaardigheidsprestatie op video op een behoorlijk wijze te kunnen evalueren, heb ik een *analytisch beoordelingsschema* ontwikkeld. Figuur 1 laat zien dat zo'n schema een aantal dimensies en een aantal vaardigheidsniveaus onderscheidt. Het is de bedoeling dat de beoordelaar zich bij het bekijken van de videoband beurtelings op elk van die dimensies concentreert. Het schema is ontworpen voor een specifieke doelgroep (tweede kandidatuurstudenten Germaanse Talen) die voor de laatste keer in hun opleiding getoetst worden op spreekvaardigheid en vervolgens zonder enige verdere taalvaardigheidstraining een onderwijsbevoegdheid kunnen krijgen) en kan in een andere context enkel als inspiratiebron dienen. Vooralsnog heb ik als dimensies de vier traditionele pijlers gekozen (uitspraak, vlotheid, grammaticale correctheid en woordenschat), aangevuld met een recentere component die ook plaats biedt aan pragmatische, sociolinguïstische en strategische vaardigheden, want het is bekend dat de kwaliteit van iemands spreekvaardigheid niet enkel door taalkundige factoren bepaald wordt.

Het analytisch beoordelingsschema

dimensions	5 native-like	4 high school teacher	3 second-year student	2 remediation needed	1 negative
pronunciation	4	3	2	1	0
fluency	4	3	2	1	0
structural accuracy	4	3	2	1	0
vocabulary	4	3	2	1	0
success in communication	9	7	5	3	1

Opdat een analytisch beoordelingsschema behoorlijk zou werken, is het ten eerste belangrijk dat elke dimensie duidelijk omschreven wordt, bijvoorbeeld met een opsomming van de aandachtspunten. Zo horen bij 'success in communication' in dit schema de volgende focusvraagjes: 'Slaagt de student erin de boodschap over te brengen?', 'In welke mate is de student daarbij afhankelijk van uw steun?', 'Wekt de taalvorm enige wrevel of irritatie bij u op?', 'Kan de student zich bij eventuele taalmoeilijkheden behelpen?', enz. Ten tweede moeten de niveaus gedefinieerd worden, met voorop een bepaling van de grens tussen slagen en zakken. In het voorgestelde schema ligt de cesuur tussen de niveaus 2 ('remediëring vereist') en 3 ('voldoende voor een kandidaat in de Germaanse Talen'), want daar gaat het erom of een student de nodige taalvaardigheid bezit om zonder uitdrukkelijke remediëring na de studie als model voor een klas te kunnen functioneren. Wanneer een student de <th> onveranderlijk als /t/ of /d/ uitspreekt, betekent dat bijvoorbeeld dat hij of zij niveau 3 niet haalt voor de uitspraakdimensie, want bij zo'n student ontbreken twee van de fonemen van het Engels en bijsturing is in zo'n geval werkelijk noodzakelijk. Tenslotte moet elk vakje in het schema aan een cijfer gekoppeld worden, wat gepaard gaat met vragen als: 'Weegt "success in communication"

zwaarder dan de andere dimensies?', en: 'Is de afstand tussen elk niveau even groot?'
Het antwoord op die vragen zal mee afhangen van de doelstellingen van de leerkracht en van het niveau van de studenten. Ook de empirische vakliteratuur die onderzoekt waar moedertaalsprekers nu precies over struikelen in gesprekken met anderstaligen en waar ze het meeste belang aan hechten, kan de leerkracht helpen bij het afwegen. Figuur 1 geeft een mogelijke invulling bij een maximumcijfer van 25 punten.
Ik ben nog aan het nagaan in hoeverre dit beoordelingsschema deugt en of ook collega's zich erin kunnen vinden. Toch is het al duidelijk dat zo'n analytisch schema de betrouwbaarheid beter bewaakt dan de holistische beoordeling waarbij intuïtief één globaal cijfer toegekend wordt. Daar bestaat immers het gevaar dat de evaluatie volledig bepaald wordt door een enkele factor (zoals de <th> of zelfs het afwijzen van de inhoud) en dat bij verschillende beoordelaars bovendien telkens andere factoren bepalend zijn (Beheydt 1992).

De affectvragenlijst

In de toetsliteratuur kijkt men vaak neer op indruksvaliditeit, omdat het feit dat een student de indruk heeft te maken te hebben met een goede toets helemaal niets toe kan voegen aan de discussie rond de *werkelijke* validiteit van die toets (Stevenson 1985). Vanuit die benadering is indruksvaliditeit louter schijnvaliditeit, een produkt uit de kosmeticarekken. Zo'n dédain vind ik onterecht: een zogenaamd 'uitstekende' toets die op grote scepsis botst bij de studenten, zal immers in de praktijk evenmin werken als een toetstechnisch onding met de nodige 'test appeal'. Door die indruksvaliditeit ook te gaan meten, komt de leerkracht heel precies te weten welke aspecten van de toets de studenten wel en welke ze niet aanvaarden, en dat is informatie met een hoge signaalwaarde voor de lesgever. Ik heb de voorbije twee jaar dan ook alle studenten meteen na hun mondelinge taalvaardigheidsproef Engels een affectvragenlijst (figuur 2) in laten vullen. Die vragenlijst is een licht aangepaste versie van de questionnaire die Scott (1986) als onderzoeksinstrument heeft gebruikt. Vanuit verschillende invalshoeken peilen veertien stellingen naar de indruksvaliditeit van de pas afgelegde proef: vinden de studenten zo'n toets wel *nodig* (stelling 7), vinden ze hem *betrouwbaar* (2 en 6), *valide* (1, 8, 9 en 11), *prettig* (10, indirect ook 4, 12 en 13), en hoe *succesvol* denken de studenten dat ze waren (5, indirect ook 14). De meeste resultaten worden hieronder besproken. Ik wil wel opmerken dat zo'n vragenlijst waardeloos wordt wanneer de studenten denken dat ze maar beter kunnen opschrijven wat de leerkracht graag wil horen, namelijk hoe fantastisch de toets wel was!

- **De studenten over noodzaak, validiteit en welgevallen**

Uit de vijfpuntsschalen van de affectvragenlijst blijkt dat de studenten de mondelinge taalvaardigheidsproef valide, prettig, en van het hoogste belang vinden. Die statistische resultaten worden bevestigd door de vrije commentaar bij de diverse stellingen (bijv. 'I had the possibility to lead the conversation to what I wanted to say', 'It was quite nice to have a chat with my teacher about ordinary stuff that is on my mind'). Toch merken velen ook op dat de zenuwen hun parten hebben gespeeld en dat ze tien minuten eigenlijk niet genoeg vinden om te laten zien wat ze in hun mars hebben. Deze laatste verzuchting is slecht nieuws voor het al hyperdrukke leraarsbestaan. Daar komt nog bij dat andere mogelijk tijdbesparende formules, zoals het groepsgesprek, massaal door mijn studenten afgewezen wor-

De affectvragenlijst
Affect questionnaire

name: date:

It is important to emphasize that the opinions you give will not interfere in any way with your evaluation.

Please mark with an 'X' the description on the agreement/disagreement scale which best expresses your opinion. Note that in some questions you are asked to explain your response.

A: agree; B: somewhat agree; C: neutral; D: somewhat disagree; E: disagree

	A	B	C	D	E
1. I believe that this oral test is an accurate evaluation of my ability to speak English. (Please explain why/why not.)	0	0	0	0	0
2. If I took the oral test from another teacher, I would get a quite different score.	0	0	0	0	0
3. I felt nervous before the test.	0	0	0	0	0
4. I felt nervous during the test.	0	0	0	0	0
5. I believe I did well on the oral test. (Please explain why/why not.)	0	0	0	0	0
6. If I took the same test with the same teacher on another day, the result would be the same.	0	0	0	0	0
7. I believe that oral tests are necessary in English courses.	0	0	0	0	0
8. I believe I had an adequate opportunity to demonstrate my ability to speak English. (Please explain why/why not.)	0	0	0	0	0
9. The oral test was too short.	0	0	0	0	0
10. I liked the oral test. (Please explain why/why not.)	0	0	0	0	0
11. I understood what I was supposed to do during the oral test.	0	0	0	0	0
12. I feel more comfortable when I take an oral test with another student.	0	0	0	0	0
13. I feel more comfortable when I take an oral exam in a group.	0	0	0	0	0
14. I thought the oral test was too difficult. (Please explain why/why not.)	0	0	0	0	0

den. Dit is een mooi voorbeeld van hoe de affectvragenlijst een signaalfunctie vervult: als ik in de toekomst de toetstijd per student zou willen verminderen, moet ik dat duidelijk met de nodige omzichtigheid en verantwoording doen, zo niet, dan zal de indruksvaliditeit van mijn toets gevoelig dalen. Voorts leren de reacties dat je het examenresultaat het beste eerst kunt vergelijken met de observaties bij de tijdens het jaar gehouden gespreksoefeningen, teneinde de extremen weg te filteren die eerder een effect zijn van de zenuwen dan van taalvaardigheid. Ten derde blijkt dat de studenten zo al overtuigd zijn van de noodzaak van een spreekvaardigheidsproef zonder dat ik daar in de klas een peptalk voor hoef te houden.

- **De studenten over betrouwbaarheid**
  Zoals aangegeven belichten de stellingen in de affectvragenlijst twee facetten van betrouwbaarheid, namelijk overeenstemming tussen beoordelaars (stelling 2) en stabiliteit van prestatie en beoordeling (stelling 6). Tot mijn verrassing geloven slechts 10 van de 65 studenten dat ze bij een andere leerkracht een ander cijfer zouden krijgen terwijl interbeoordelaarsovereenstemming toch een heel gevoelig punt is in de toetsliteratuur.
  De stabiliteit van prestatie en van beoordeling hebben een erg hoge indruksvaliditeit bij de studenten. Opnieuw kunnen de antwoorden als signaal opgevat worden en kan de leerkracht besluiten om de betrouwbaarheidsfactor niet meer aan te kaarten (want de studenten zien hier toch geen problemen) of om juist aan te tonen hoe ongegrond hun vertrouwen wel is.

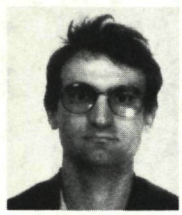
Besluit

De zorgvuldigheids-, gelijkheids- en redelijkheidseisen die gesteld worden aan het besturen, gelden blijkens de rechtspraak ook voor het examineren, waar ze als validiteits- en betrouwbaarheidseisen bekend staan. Door gebruik te maken van de video-opname, aangevuld met het analytisch beoordelingsschema en de affectvragenlijst, probeer ik de spreekvaardigheid te evalueren zonder voortdurend de vinger nat te hoeven maken.

Literatuur

- Bachman, L.F., *Fundamental considerations in language testing*, Oxford: Oxford University Press, 1990.
- Beheydt, L., Evaluatie van schrijfvaardigheid, in: A. Helbo (red.), *Evaluatie en taalonderwijs. Liber amicorum aangeboden aan Frans Van Passel*, Bern: Peter Lang, 1992, 325-338.
- Haggstrom, M., Using a videocamera and task-based activities to make classroom oral testing a more realistic communicative experience, in: *Foreign language annals* 27 (1994), 161-175.
- Lier, L. van, Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversation, in: *TESOL Quarterly* 23 (1989), 489-508.
- Maele, J. Van, Diagnostische evaluatie in de spreekvaardigheidsklas met behulp van video, in: *Werkmap voor taal- en literatuuronderwijs* 72 (1994), 173-180.
- Neve, H. De, & P.J. Janssen, *Succesvol examineren in het hoger onderwijs*, Leuven: Acco, 1992.
- Scott, M.L., Student affective reactions to oral language tests, in: *Language Testing* 3 (1986), 99-118.
- Spolsky, B., Policy issues in testing and evaluation, in: *Annals of the American Association of Political and Social Sciences* 532 (1994), 226-237.
- Stevenson, D.K., Pop validity and performance testing, in: Y.P. Lee e.a. (red.), *New directions in language testing*, Oxford: Pergamon, 1985, 111-118.
- Underhill, N., *Testing spoken language*, Cambridge: Cambridge University Press, 1987.
- Verstegen, R., Rechtsbescherming in onderwijsverband, in: L. Van Hoesberghe (red.), *Student(en)recht. Sociale en juridische gids voor de student hoger onderwijs*, Leuven: Acco, 1994, 119-131.

Jan Van Maele



Geboren 1964. Studeerde Germaanse Talen aan de KU Leuven en behaalde een Master's in TESOL aan Central Connecticut State University. Was drie jaar werkzaam aan de universiteit van Wuhan, China. Is sinds 1992 verbonden aan de vakgroep NT2 van het Instituut voor Levende Talen (KU Leuven) en aan de kandidatuuropleiding Germaanse Talen van de KU Brussel, waar hij is ingeschakeld bij de colleges Engelse taalbeheersing. Bereidt een proefschrift voor over beoordelingsaspecten van spreekvaardigheid in de vreemde taal.
Adres: KU Brussel, Vrijheidslaan 17, 1080 Brussel.