Maar naast 'doeltaal = voertaal' en 'reële' conversatie blijft het spelen van gefingeerde situaties noodzakelijk. Er zijn namelijk veel woorden en ook enkele grammaticale constructies die in de klassensituatie zelden of nooit gebruikt worden, maar die de leerling toch nodig heeft in allerlei situaties buiten de school.

4.3.7 Oefenparameter 7: makkelijk tegenover moeilijk – inhoudelijk en talig

Bij vrije(re) spreek- en schrijfoefeningen kan de inhoud die geformuleerd en gearticuleerd moet worden, simpeler of juist moeilijker zijn. Een voorbeeld: de leerling moet aan een fictieve vreemdeling uitleggen hoe die van een bepaalde plek in het centrum van zijn woonplaats naar een andere plek komt. Dat kan simpel zijn, maar ook heel lastig, zelfs in de eigen moedertaal, afhankelijk van de situatie. Dit betekent dan een flinke belasting van de *conceptualizer*.

Daarbij kan het benodigde taalmateriaal makkelijk of moeilijker zijn voor de leerling. Beheerst hij het vocabulaire en de grammatica die nodig is voor deze specifieke communicatieve taak, of moet hij vreselijk zoeken? De factor 'talige moeilijkheidsgraad' is vooral bepalend voor de belasting van de *formulator*. Het vinden van alternatieve formuleringen of omschrijvingen voor woorden of uitdrukkingen die de leerling niet ter beschikking heeft, betekent dan weer extra werk voor de *conceptualizer*: de *formulator* moet zoeken, terwijl de *conceptualizer* moet blijven controleren of de boodschap wel dezelfde blijft als oorspronkelijk bedoeld.

5 Toetsen

5.1 Toetsen versus oefenen

Toetsen lijken vaak sterk op oefeningen. Alleen gaat het bij het toetsen niet om het *oefenen*, maar om het *meten* en het *beoordelen*.

In het dagelijks taalgebruik gaat men vaak slordig om met het verschil tussen oefening en toets. Zo spreken vreemdetaaldocenten vaak van een '*luistertoets*', als ze een luister(vaardigheids)*oefening* bedoelen. Dit zal wel te maken hebben met het feit dat de Cito-luistervaardigheidstoetsen (of de vorm ervan) heel vaak als oefening gebruikt worden.

Nog een berucht – en al sinds de invoering in de jaren zeventig bekritiseerd – voorbeeld van het gebruik van toetsen als oefeningen zijn de toetsen van het Cito voor de centrale examens leesvaardigheid, en met name de meerkeuzevragen daarin. Veel oefenen daarmee kan tot gevolg hebben dat leerlingen meer bedreven raken in het beantwoorden van meerkeuzevragen dan in echte leesvaardigheid in de vreemde taal. Het eerste bezwaar is dat leerlingen met het beantwoorden van (meerkeuze- en andere soorten) vragen niet of te weinig leren zelfstandig vreemdtalige teksten aan te pakken. In het leven buiten de school worden nooit vragen bij vreemdtalige teksten geleverd. Het tweede bezwaar betreft specifiek meerkeuzevragen: die dwingen toetsmakers om moeilijke, vaak abstracte en ver van het bed liggende teksten te zoeken. Bij teksten die beter aansluiten bij de belangstelling en kennis van de wereld van de leerlingen, blijkt het lastig om meerkeuzevragen te maken die discrimineren, dat wil zeggen niet door alle leerlingen goed gemaakt worden.

Docenten moeten natuurlijk hun leerlingen voorbereiden op de toets of het examen. Maar ze horen de leerlingen niet *uitsluitend* op te leiden voor die toets of dat examen. De toets of het examen is een meetinstrument, en wat daar gevraagd en gemeten wordt is niet per se identiek met wat de leerling moet kunnen in de maatschappij.

Heel wat zinvolle leerdoelen zijn moeilijk of niet te meten, vooral leerdoelen op het attitudinele vlak, die dus houdingen betreffen, bijvoorbeeld of iemand een positieve houding heeft tegenover het leren van een bepaal-

de vreemde taal of tegenover de sprekers van die taal. Toch mag een meetbaarheidsprobleem geen reden zijn om zulke doelen maar af te schrijven. In de kerndoelen en de examenprogramma's staan goed meetbare doelen naast minder goed meetbare. Loop ze maar eens na.

In examenprogramma's kunnen naast 'toetsdelen' ook 'handelingsdelen' voorkomen: oefeningen die voor het examen gedaan moeten zijn, maar die geen toets zijn. Hieruit zou je kunnen afleiden dat er bij handelingsdelen niet wordt gemeten en beoordeeld. Maar er wordt wel degelijk gemeten en beoordeeld: er wordt gekeken of de handelingsdelen 'naar behoren' gedaan zijn. Er is dus een voldoende-onvoldoendegrens (cesuur).

5.2 Meten en beoordelen; functies van toetsen

Meten – corrigeren, de fouten vaststellen – is iets technisch en neutraals; het heeft op zichzelf nog geen gevolgen voor degene die getoetst wordt. Wel kunnen bij de toets bepaalde, meestal beperkende, eisen gesteld worden, bijvoorbeeld dat de getoetste bepaalde vormen van hulp niet krijgt, dat hij de taak binnen een bepaalde tijd moet uitvoeren, en dergelijke.

Beoordelen – de fouten wegen en tellen (scoren), de cesuur vaststellen en het cijfer berekenen – vindt plaats op grond van de resultaten van de meting.

Het toetsen heeft twee belangrijke functies en gevolgen voor de getoetste:

1. een *feedbackeffect (diagnostisch effect)*: de getoetste komt te weten wat er wel en wat (nog) niet goed is aan zijn prestatie en wat hij moet doen om die te verbeteren;
2. een *curriculair en/of civiel (maatschappelijk) effect*: de toetsing heeft gevolgen voor de loopbaan van de getoetste in het opleidingssysteem (overgaan of blijven zitten, wel of geen toegang tot een vervolgopleiding) en/of in de maatschappij (wel of niet in aanmerking komen voor bepaalde banen en functies).

Vaak wordt gesproken van *formatief* en *summatief* toetsen. Formatieve toetsing geeft de leerling feedback, summatieve toetsing heeft een curriculair en/of civiel effect. In de schoolpraktijk lopen deze twee functies in elkaar over, want vrijwel altijd worden voor toetsen cijfers gegeven die meetellen voor het rapport, de overgang, uiteindelijk voor het diploma. Duidelijk is wel dat een globale beoordeling met alleen maar een cijfer de leerling weinig feedback geeft. Hoe beter onderbouwd en hoe specifieker (analytischer) de beoordeling is, des te meer heeft de leerling eraan voor zijn verdere leerproces.

Toetsing is ook belangrijk voor de docent, om twee redenen.

1. Het toetsen geeft de docent *feedback* (terugkoppeling) over zijn onderwijs: heeft hij de gestelde doelen weten te bereiken? Natuurlijk is hij daarbij afhankelijk van de leerlingen, maar hij kan ook uit de resultaten conclusies trekken met betrekking tot zijn onderwijs. Misschien heeft hij de doelen te hoog of te laag gesteld, misschien kan hij uit de toetsresultaten afleiden dat hij zijn onderwijs anders moet inrichten om toch de doelen te halen.
2. Het toetsen heeft in de schoolpraktijk voor de docent ook nog een meestal inofficiële, maar niet minder belangrijke functie: *extrinsieke motivatie*. Tegenover onwillige leerlingen zonder intrinsieke motivatie voor het onderwijs en/of het vak hanteren docenten cijfers, rapporten, diploma's vaak als lok- en/of dwangmiddelen. Toetsen is een bekend disciplineringsmiddel: veel (schriftelijke) overhoringen, extra proefwerken, misschien ook strengere beoordelingen dan bij leerlingen of klassen die wel plezierig meewerken. De omgang met deze functie van toetsing vergt veel van de beroepsmoraal van de leraar.

5.3 Meten

Bij het meten spelen twee begrippen een belangrijke rol:

- *Validiteit* (of geldigheid). De vraag *Is de toets valide?* betekent *Meet je met de toets ook werkelijk wat je wilt toetsen of pretendeert te toetsen?* Bijvoorbeeld: meet je met je gespreksvaardigheidstoets (spreektoets) ook werkelijk gespreksvaardigheid, of met een lees(vaardigheids)toets ook werkelijk leesvaardigheid? Dit lijkt een open deur, maar is dat bij nadere beschouwing zeker niet.
- *Betrouwbaarheid*. De vraag *Is de toets betrouwbaar?* betekent *Komt dezelfde beoordelaar op verschillende momenten bij dezelfde prestatie tot dezelfde meting en dezelfde beoordeling?* En ook: *Komen verschillende beoordelaars bij dezelfde prestatie tot dezelfde meting en beoordeling?*
  Voor betrouwbare toetsing zijn natuurlijk correctie- (meet-) en beoordelingsrichtlijnen nodig. Uit de betrouwbaarheidseis volgt dat een toets geen toets is zonder meet- en beoordelingsinstrument erbij. Hij is dan net zoveel waard als een thermometer zonder schaal: hij wijst wel iets aan, maar er is geen schaalverdeling. Met alleen de opgaven voor de leerling in de hand moet de beoordelaar een slag in de lucht slaan en is betrouwbaarheid ver te zoeken.

5.3.1 Validiteit

Toetsdeskundigen onderscheiden verschillende soorten validiteit. Wij gaan hier voornamelijk in op wat zij de *inhoudsvaliditeit* van toetsen noemen.

Toetsen hangen altijd samen met (leer)doelen. De vraag *Meet je met de toets ook werkelijk wat je wilt toetsen of pretendeert te toetsen?* kan ook zo geformuleerd worden: *Meet je met de toets werkelijk in hoeverre de leerling het doel heeft bereikt dat je hem wilt laten bereiken?* Anders gezegd: om te kunnen

zeggen of een toets valide is, moet je weten wat het leerdoel is waarvoor de betreffende toets geacht wordt het meetinstrument te zijn.

Het leerdoel is in algemene zin af te lezen uit de naam van de toets: een schrijfvaardigheidstoets bijvoorbeeld moet schrijfvaardigheid meten. Maar vervolgens duikt de vraag op: wat versta je precies onder schrijfvaardigheid? Hetzelfde geldt voor alle andere vaardigheden en deelvaardigheden.

Om bij schrijfvaardigheid als voorbeeld te blijven: is een dictee een valide schrijfvaardigheidstoets? De leerling moet inderdaad schrijven, maar krijgt tekstopbouw, alle formuleringen, zinnen en woorden cadeau. Wat van hem gevraagd wordt, is voornamelijk spelling, een deelvaardigheid dus.

Een opstel dan? Dat roept de vraag op wat eigenlijk 'een opstel' is. Uit de opdracht *Schrijf een opstel over milieuvervuiling* kan een heel andere tekstsoort komen dan uit de opdracht *Schrijf een opstel over je mooiste vakantie* – misschien wordt het een betoog, misschien een verslag, misschien een fictioneel, literair verhaal. Dat zijn heel verschillende tekstsoorten met eveneens verschillende tekststructuren, stijlen, soorten woordgebruik enzovoort.

De vraag is dus: welke *tekstsoorten* in welke *rol(len)* voor wat voor *publiek* of *lezer(s)*? Zo blijkt de vraag naar validiteit van toetsvormen direct samen te hangen met de keuze van leerdoelen.

Niet zelden geven praktische omstandigheden de doorslag bij het kiezen van toetsvormen. Neem gespreksvaardigheid. Vaak wordt schriftelijk getoetst met opgaven van het type *Zeg in het Duits/Engels/Frans/... dat je wel een glas limonade lust.* Is dat een valide gespreksvaardigheidstoets? Om praktische redenen (mondelinge toetsing kost heel veel tijd) telt dit soort toetsen vaak voor het gespreksvaardigheidscijfer. Maar een valide toetsvorm kan het natuurlijk nooit zijn, ook niet als je de spelfouten niet meetelt.

Vaardigheid opdelen in deelvaardigheden

In principe is het mogelijk om een vaardigheid *op te splitsen in diverse deelvaardigheden* die samen de vaardigheid moeten representeren. Op basis daarvan kan dan een batterij deeltoetsen worden geconstrueerd. Bijvoorbeeld voor schrijfvaardigheid: een dictee plus een grammaticatoets plus een vocabulairetoets plus een stijltoets plus een tekstopbouwtoets plus ...

Met zo'n toetsbatterij zijn er twee problemen. Ten eerste de validiteitsvraag: is de som van de getoetste deelvaardigheden werkelijk gelijk aan de beoogde vaardigheid, in dit geval schrijfvaardigheid? Ten tweede ontstaat bij invoering als examentoetsvorm wat men noemt een *terugslageffect* van de toetsvorm op het onderwijs: *teaching to the test.* De kans is groot dat veel docenten met hun leerlingen teveel of zelfs alleen nog maar gaan oefenen met de toetsvorm, bijvoorbeeld met grammatica, vocabulaire en andere deelvaardigheden, in plaats van ze teksten te laten schrijven. Daarom heeft men in het Nederlandse vto lange tijd (tot 2007) geen deelvaardigheidstoetsen in het

examen willen hebben. Het beste is dat de toetsvorm zoveel mogelijk lijkt op datgene wat de leerling in het leven buiten de school moet kunnen.

Receptieve vaardigheden

Bij schrijven is het te meten product van de leerling een stuk geschreven taal, bij spreken een stuk gesproken taal. Als je naar de receptieve vaardigheden kijkt, stuit je op een apart probleem: wat is eigenlijk het *product van een receptieve vaardigheid*? Het gaat om het begrijpen van teksten. Maar begrijpen is iets wat in de hoofden van luisteraars en lezers gebeurt, en dat is niet voor anderen waarneembaar en controleerbaar. Daarom moeten leerlingen in een toets iets extra's doen waarmee ze hun tekstbegrip laten zien. Dat kan bijvoorbeeld parafraseren of vertalen zijn, samenvatten, tekst segmenteren, inhoudsvragen beantwoorden. Maar hoe precies en reëel is het beeld dat je met dat soort toetsing van de luister- of leesvaardigheid van de leerlingen krijgt? Het is altijd een indirect beeld.

Er is nog een probleem. Leerlingen die goed of slecht scoren op een toets of batterij toetsen waarvan de deskundigen denken dat er leesvaardigheid mee gemeten wordt, scoren in dezelfde mate goed of slecht op een leesvaardigheidstoets met meerkeuzevragen. Dat hebben taaltoetsdeskundigen ooit vastgesteld bij de invoering van meerkeuzevragen voor leesvaardigheid in het Nederlandse vto in de jaren zeventig. Maar geldt dit ook voor leerlingen die uitgebreid geoefend hebben met meerkeuzevragen en daar een bijzondere behendigheid in ontwikkeld hebben? Dan wordt die bijzondere behendigheid meegemeten. Dat is niet zinvol, want bij teksten die die leerlingen buiten de school lezen, krijgen ze nooit meerkeuzevragen ter beantwoording bijgeleverd. Vergelijk dat met persoonlijkheids- en intelligentietests. Oefening met een bepaalde testvorm leidt over het algemeen tot een hogere score. Maar is de kandidaat door het oefenen met persoonlijkheidstests een andere persoonlijkheid geworden? Of door het oefenen met intelligentietests intelligenter?

Validiteit van deelvaardigheidstoetsen

In alle klassen vóór de examenklassen worden veel toetsen in *deelvaardigheden* afgenomen, omdat iedereen aanneemt dat het apart oefenen van deelvaardigheden bijdraagt tot de beheersing van taalvaardigheden. Nu lijkt het met de validiteit van deelvaardigheidstoetsen, als je die apart bekijkt, een stuk eenvoudiger te liggen dan bij vaardigheidstoetsen. Maar ook dat valt bij nadere beschouwing tegen.

Denk bijvoorbeeld aan een vocabulairetoets. Woordenkennis toets je toch door naar een vertaling van een woord te vragen? Maar bij nadere beschouwing is het bijvoorbeeld de vraag of dat via vertaling moet of mag (bijvoorbeeld als er eigenlijk geen goede vertaling is), en of je woorden contextloos of juist in (mini)contexten – ten minste zinnen – af moet vragen, want zo komen ze in normaal taalgebruik altijd voor.

Welke toetsvorm kies je voor een grammaticatoets? Gatenzinnen natuurlijk. Maar die zijn schriftelijk, en bij nadere beschouwing blijken leerlingen veel grammatica die ze schriftelijk wel beheersen, mondeling niet te beheersen. Blijkbaar vereist mondeling gebruik een hoger beheersingsniveau, dus als mondelinge beheersing het doel is, moet je die grammatica ook mondeling toetsen.

En uitspraak toets je toch door de leerling een stukje te laten voorlezen? Maar voorlezen is iets anders dan normaal spreken. Onder meer kan de uitspraak negatief beïnvloed worden door de schrijfwijze – in het Duits wordt bijvoorbeeld de umlaut of in het Frans een accent over het hoofd of verkeerd gezien – en dat is geen uitspraakfout maar een (voor)leesfout.

Bij alle deelvaardigheidstoetsen blijft het principieel de vraag of je daarmee iets toetst wat de leerling buiten de school moet kunnen. Een deelvaardigheid wordt buiten het vto nooit 'los' bedreven, maar komt daar alleen als geïntegreerd bestanddeel van vaardigheden voor.

Meet je alleen wat je hebt onderwezen?

Nog een heel lastig probleem: vraag je alleen van de leerling wat hij gehad heeft (dus wat jij als docent onderwezen hebt), of meer dan dat? Meestal willen leraren zich wel houden aan de ongeschreven (school)wet dat je alleen vraagt of tenminste alleen fout rekent 'wat de leerling had kunnen weten', dus wat duidelijk tot de leerstof van het voorafgaande onderwijs behoorde. Maar bij taalvaardigheidstoetsen is dat praktisch onmogelijk. Luister- en leestoetsen gaan altijd ergens over, dus doen ze een sterk beroep op de kennis die de leerling over het onderwerp heeft – kennis die hij meestal helemaal niet bij jouw vak, maar bijvoorbeeld thuis heeft opgedaan. Leraren die met allochtone leerlingen werken, weten hier alles van: thuisculturen kunnen flink verschillen. Vergelijkbare problemen gaan spelen zodra je bij spreken en schrijven niet alleen eist dat de leerling in de doeltaal *formuleert*, maar ook de inhoud van wat hij formuleert, *bedenkt* en *ordent*.

5.3.2 Betrouwbaarheid

Als het over betrouwbaarheid gaat, zijn er twee belangrijke factoren: de *toets*(vorm) en de *toetser*, degene die de meting en de beoordeling doet.

De toets

Betrouwbaar is een toets die goed meet en die objectief gescoord kan worden. Het betrouwbaarst zijn toetsen met opgaven waarop maar één antwoord goed is. Zulke gesloten toetsen zijn populair bij leraren én leerlingen: heldere eisen, geen discussies over wat nog wel en wat niet goed is. Over het algemeen zijn deelvaardigheidstoetsen (bijvoorbeeld woordenschat- of grammaticatoetsen) makkelijker betrouwbaar te krijgen dan vaardigheidstoetsen. Maar wat je aan

betrouwbaarheid wint, verlies je aan validiteit – als taalvaardigheid je onderwijsdoel is. Het typische van echte taalvaardigheid is immers dat er bijna altijd meer dan één goed antwoord is of meer dan één goede manier om aan de opdracht te voldoen.

Toch valt ook bij deelvaardigheidstoetsen de betrouwbaarheid weleens tegen. Bij woordenschattoetsen bijvoorbeeld kan het gebeuren dat een leerling een andere vertaling geeft dan die in de geleerde woordenlijst stond, maar die in een bepaalde context ook goed kan zijn. Bij uitspraak is de grens tussen goed en fout vaak niet scherp: klinkt klank X nu wel echt Duits/Engels/Frans/...?

Echte taalvaardigheidstoetsen zijn nog veel lastiger betrouwbaar te maken. Bij de receptieve vaardigheden neemt men graag zijn toevlucht tot meerkeuzevragen, omdat daarbij maar één antwoord goed is. Hoewel – iedereen kent de discussies over de vraag of één van de afleiders niet eigenlijk óók goed is. Dat gebeurt ook bij de Cito-toetsen, terwijl daar toch door allerlei experts eindeloos aan gesleuteld is. En je komt, zoals gezegd, gauw in validiteitsproblemen. Hoe vaak krijgen lezers in het leven buiten de school meerkeuzevragen bij een tekst?

Bij de productieve vaardigheden heb je altijd te maken met een heel complex product dat je moet meten. Een bekend middel om het aantal te meten variabelen te reduceren, is geleide toetsing. De leerling krijgt de inhoud en meestal ook de opbouw voorgegeven van wat hij in de vreemde taal moet zeggen of schrijven. Dan valt die inhoud buiten de meting en de beoordeling – een groot probleem minder, zoals de collega's in het moedertaalonderwijs weten. Maar is een geleide toets wel valide? In normale communicatie buiten de vreemdetaalles kan de leerling zich niet beperken tot het formuleren, maar moet hij ook zelf zorgen voor inhoud en ordening van zijn uitingen.

Globaal versus analytisch meten

Het gevolg van al deze meetproblemen is dat vreemdetaaldocenten nogal eens hun toevlucht nemen tot globale (of holistische) meting. Je kunt (meer) *globaal* meten of (meer) *analytisch*. In het eerste geval ga je op een indruk af; op basis daarvan geef je een indrukscijfer. In het tweede geval deel je de toets in in kleinere onderdelen die onderlinge vergelijking van leerlingenprestaties makkelijker maken, en/of je onderscheidt meerdere aspecten en/of soorten fouten, je telt ze per stuk en geeft ze verschillende gewichten; zo kom je al redenerend en berekenend tot een meting.

Bij vaardigheden, vooral bij gespreksvaardigheid, neigt men vaak tot globale meting en beoordeling, omdat de analytische aanpak lastig en vaak ook tijdrovend is. Maar de nadelen zijn duidelijk:

- Een globale meting zal, omdat die minder nauwkeurig is, eerder subjectief zijn, en dus onbetrouwbaar. Je kunt door veel ervaring – dat is eigenlijk: door erg veel leerlingenprestaties met elkaar vergeleken te hebben – wel

een behoorlijk betrouwbare globale beoordelaar worden, zelfs van zoiets lastigs als gespreksvaardigheid, maar het duurt lang. Beginnende beoordelaars vertonen een extra grote onbetrouwbaarheid.
- De leerling heeft aan een globale meting minder dan aan een analytische meting, omdat hij minder of zelfs helemaal geen gerichte informatie (feedback) krijgt over wat hij moet doen om zijn prestaties te verbeteren.

Als het gaat om toetsing van gespreks- en schrijfvaardigheid hebben geleide toetsen voordelen:
- Geleide toetsen leveren meer dan genoeg betrouwbaarheidsproblemen op, maar wel minder dan volledig open toetsvormen. Bij open vormen is immers veel minder voorspelbaar wat de leerling produceert, zodat het veel moeilijker wordt om de leerlingenprestaties te vergelijken – de basis van alle meting en beoordeling. Hoe vaker je dezelfde spreek- of schrijfopdracht door verschillende leerlingen hebt laten doen en hoe sterker geleid hij is, des te scherper hoor of zie je de verschillen in prestatie.
- Met geleide toetsen sluit je het lastige probleem van de meting en beoordeling van de inhoud grotendeels of geheel uit. Dat heeft natuurlijk ook bedenkelijke kanten, namelijk op het punt van de validiteit.
- Geleide toetsen maken het makkelijker om de leerlingenprestaties in kleinere stukjes, namelijk per deelopdracht, te meten. Hoe kleiner de stukjes, des te makkelijker is het de leerlingenprestaties onderling te vergelijken.

Gespreksvaardigheidstoetsen zijn per definitie mondeling en daarom extra lastig. Voor het nakijken van schriftelijke toetsen kun je zoveel tijd nemen als je wilt, je kunt nog eens terugkijken en zo nodig eindeloos op bepaalde details en aspecten vergelijken. Bij mondelinge toetsen gaat het om het gesproken woord, dat meteen vervluchtigd is. Dat vereist grote concentratie en routine – en/of vastlegging op een geluidsdrager. Dat laatste is altijd aan te raden, al was het alleen voor de zekerheid. Voor beginners in het vak van toetsafnemer is het zeer aan te bevelen om het oordeel niet meteen te geven, maar achteraf op grond van de opnamen. Dat kost veel tijd, maar is heel leerzaam.

De toetser

Op welke manier wint de toetser aan betrouwbaarheid? Twee antwoorden:
- doordat de toetser bij elke leerlingenprestatie dezelfde, zo analytisch mogelijke criteria hanteert;
- doordat er niet maar één toetser naar elke leerlingenprestatie kijkt of luistert, maar zoveel mogelijk toetsers.

Het is belangrijk een vaststaand meet- en beoordelingsinstrument te hanteren, waarin zo helder mogelijk wordt aangegeven op welke aspecten (soorten fouten) gelet wordt en hoe die beoordeeld worden. Dan blijven er nog genoeg

problemen over. In het (vreemde)talenonderwijs zijn veel dingen niet simpelweg goed óf fout. Heeft de leerling de essentie van de tekst niet helemaal precies of eigenlijk helemaal niet begrepen? Spreekt de leerling een bepaalde lastige klank heel erg verkeerd of maar een beetje verkeerd uit? Drukt de leerling zich schriftelijk en/of mondeling nou een beetje moeizaam of gewoon heel krukkig uit? De ene docent schat dat anders in dan de andere.

In het Nederlandse onderwijs worden de toetsing én het voorafgaande onderwijs voor een groot gedeelte door dezelfde persoon gedaan. Dat heeft veel voordelen, maar ook een groot nadeel: de docent toetst niet alleen de leerling, maar ook zijn eigen onderwijs. De docent die een goede uitspraak niet zo belangrijk vindt of daar niet zo'n oor voor heeft (misschien is zijn eigen uitspraak ook niet zo geweldig?), doet minder aan uitspraaktraining en -correctie en zal aan een matige of slechte uitspraak van leerlingen niet zwaar tillen. De docent die weinig aan gespreksvaardigheid doet (misschien is hij zelf mondeling niet zo vaardig, of vindt hij het lastig om mondelinge vaardigheden te trainen?), zal minder hoge eisen aan de gespreksvaardigheid van zijn leerlingen kunnen stellen.

Hiertegen is iets te doen door bij de toetsing zoveel mogelijk docenten te betrekken. De toetsing zal daardoor nooit honderd procent objectief worden, maar wel minder subjectief, of zoals men zegt: de toetsing wordt intersubjectief. Nu kun je in de praktijk niet bij elke toets twee, laat staan nog meer docenten inschakelen. Wel kun je vaker met collega's samen beoordelen, daarvan leren en een betrouwbaarder beoordelaar worden, ook als je het weer alleen doet.

5.4 Beoordelen

Kernpunten van de *meting* zijn de vragen: waar let je op, en wat is goed en wat niet? Kernpunten van de *beoordeling* zijn de vragen:
- Hoe *weeg* je de fouten? Sommige fouten tellen zwaar mee, andere licht, weer andere helemaal niet (bijvoorbeeld omdat je ze voor deze leerlingengroep niet relevant vindt en/of omdat de leerlingen 'het niet konden weten': wat fout ging is niet onderwezen en/of is geen leerdoel).
- Waar leg je de *cesuur* (de voldoende-onvoldoendegrens)?

5.4.1 Weging

Beoordeling is weging van gemeten resultaten.

Hoe zwaar reken je het in een leesvaardigheidstoets de leerling aan dat hij een aantal detailvragen mist, maar wel de tekst als geheel begrijpt? Of andersom? Hoe zwaar weegt bij schrijven een spellingfout en bij spreken een uitspraakfout ten opzichte van een grammatica- of andere soort fout?

Bij de productieve vaardigheden speelt onder andere het probleem van de taalvorm (het formuleren) tegenover de inhoud en de opbouw. Bij gespreksvaardigheid wordt vaak gewerkt met een open toetsvorm: een min of meer spontaan, 'vrij' gesprek, waarbij de leerling niet alleen moet formuleren, maar ook de *inhoud* en de *opbouw* van wat hij zegt, zelf moet *bedenken*. Dit terwijl spreken vanwege de tempodruk toch al een stuk moeilijker is dan schrijven. De beoordelaar staat voor de lastige vraag of, en zo ja met welk gewicht, de inhoud mee moet tellen. De ene leerling houdt zich op de vlakte met algemeenheden en voor de hand liggende standaardopmerkingen, terwijl de ander een veel interessanter, maar ingewikkelder verhaal begint en in allerlei formuleringsproblemen terechtkomt. Hoe weeg je in zo'n toets inhoud en taalvorm tegen elkaar af?

Ook als je alleen naar de taalvorm kijkt, zijn de wegingsproblemen niet van de lucht. Hoe zwaar telt een fout woord (lexicale fout) ten opzichte van een foute zinsbouw of uitgang (grammaticale fout)? En ten opzichte van een spelling- of uitspraakfout? Die weging hangt af van de leerdoelen die je voor je leerlingen stelt: gaat het er alleen om dat de leerling zich (mondeling) verstaanbaar kan maken, of gaat het je óók om (mondelinge en/of schriftelijke) correctheid, en zo ja in welke mate en op welke punten (woordkeus, grammatica, uitspraak/spelling)?

5.4.2 Cesuur: cijfers op basis van goede en foute items

Uitgangspunt bij beoordeling is vergelijking van leerlingenprestaties. Iedere beoordeling is relatief. Je vergelijkt altijd prestaties binnen een groep. Die groep kan klein zijn, bijvoorbeeld een klas, of groter, bijvoorbeeld alle havo-examenkandidaten van Nederland. Als de havo-kandidaten door de jaren heen gemiddeld slechter of beter gaan scoren, wordt de cesuur bijgesteld.

Deze subparagraaf gaat over toetsen die per item beoordeeld worden. Dat zijn de meeste deelvaardigheidstoetsen (met name woordenschat en grammatica) en lees- en luistervaardigheidstoetsen die uit losse items (vragen) bestaan, dus niet bijvoorbeeld samenvattingen. Niet alle items hoeven even zwaar te wegen. Neem bijvoorbeeld een proefwerk dat uit verschillende onderdelen bestaat: een woordjesdeel, een grammaticadeel, een leesvaardigheidsdeel. Weegt een verkeerd woord in het woordjesdeel even zwaar als een verkeerde uitgang in het grammaticadeel, of weegt een ja/nee-vraag bij één zin in een leestekst even zwaar als een meerkeuzevraag bij een hele alinea?

Bij de meeste toetsen ligt het in de rede om ieder item een gewogen aantal maximaal te behalen scorepunten te geven, op grond van argumenten die aan de leerlingen uit te leggen zijn. Als de behaalde scorepunten bij elkaar opgeteld zijn, ontstaat de vraag naar de cesuur: welke score is nog voldoende, welke niet meer?

Toetsdeskundigen gaan uit van een cesuur om en nabij de 70 procent, dat wil zeggen dat 70 procent van de maximale score behaald moet zijn voor een voldoende. Daarmee stel je je eisen niet te laag, en je hebt genoeg speling boven de cesuur om goed te kunnen differentiëren tussen de voldoendes onderling. Je wilt immers verschil maken, bijvoorbeeld tussen voldoende, ruim voldoende, goed, zeer goed, en uitstekend.

Daarbij nemen we aan dat de toets goed afgestemd is op de doelgroep. Dat wil zeggen: er zitten genoeg moeilijke items in om de slechte leerlingen er uit te sorteren; die moeten een onvoldoende kunnen halen, anders was de toets te makkelijk. Maar hoeveel onvoldoende leerlingen mogen of moeten er eigenlijk zijn? Bij de eindexamens liggen op landelijk niveau de slaagpercentages meestal ergens tussen de 85 en 95 procent. De overheid kan deze percentages beïnvloeden, bijvoorbeeld door de slaag-zakregeling te wijzigen. Maar ook allerlei andere factoren spelen mee: leerlingenstromen die in de trend liggende, maar te moeilijke schooltypes of profielen kiezen, veranderingen in de inhoud of didactiek van vakken, veranderingen in de organisatie van scholen of in de schoolcultuur, in de gemiddelde kwaliteit van leraren, enzovoort. Tussen scholen, vakken en docenten kunnen flinke verschillen optreden in de percentages onvoldoendes, dus gezakten en zittenblijvers.

Stel, je hebt een toets geconstrueerd voor jouw havo 3-klas en hanteert het standaardcriterium '70 procent van de maximale score moet behaald zijn'. In jouw klas blijkt niet 85 of 95 procent, maar slechts 50 of 40 procent van de leerlingen een voldoende te halen. Grote vraag: is je toets moeilijker dan je bedoelde, of hebben deze leerlingen slechter geleerd dan je verwachtte? Of lag het misschien aan jouw onderwijs? Moet je nu de cesuur verlagen, zodat toch bijvoorbeeld 90 procent van je leerlingen een voldoende scoort?

Iedere nieuwe toets voor bijvoorbeeld havo 3 zou, voordat hij echt ingezet wordt, uitgetest moeten worden op een groep leerlingen die representatief is voor alle havo 3-leerlingen in Nederland. Dat is natuurlijk in de onderwijspraktijk ondoenlijk. Ervaren leraren kunnen, min of meer intuïtief, hun toetsen vaak verbazingwekkend goed op de groep afstemmen, maar ook zij moeten de cesuur weleens bijstellen, omdat de toets kennelijk toch te makkelijk of te moeilijk was. Ook bij de eindexamens gebeurt dit, namelijk door de normeringsterm (N-term), waarmee op basis van de scores op de centrale examens de cijferberekening achteraf bijgesteld wordt.

Een betrouwbare methode is het gebruik van gemeenschappelijke toetsen voor een hele jaarlaag en/of het hergebruik van oude, beproefde toetsen, of het gebruik van gestandaardiseerde toetsen uit een toetsenbank.

Rekenkundig intermezzo

Het Nederlandse schoolcijferstelsel (hoogste cijfer een 10, het laagste een 1 suggereert dat de cesuur bij 60 procent ligt: een 6 is voldoende – of zelfs bij 55 procent, als je meegaat met wie een 5,5 ook voldoende vindt. Als je dan bij een

toets die uit 10 items bestaat (woordjes, invullingen, opgaven ...), voor ieder item 1 scorepunt rekent, krijg je een 60-procent- (of 55-procent)cesuur. Dus: 10 goed = een 10; 9 goed = een 9 enzovoort.

Maar als je de cesuur bij 70 procent wilt houden, moet 70 procent goede items een 6 opleveren. Anders gezegd: met 30% fout heb je een 6. Dat wil zeggen dat iedere fout niet 1 (cijfer)punt aftrek oplevert, maar 1,33 (cijfer)punt. Het gevolg is dat een leerling met 7,5 foute items al op een 0 zit. Een leerling met 10 foute items zou 13,3 aftrekpunten moeten krijgen, dus een –3,3. Om dat effect te vermijden, kun je aan het ondereind van de schaal, dus bij een 1, gewoon stoppen met verder omlaag gaan. Als je het verschil tussen de heel slechte prestaties (bijvoorbeeld tussen 7 en 10 fouten op een maximale score van 10) duidelijker tot uitdrukking wilt brengen in het cijfer, dan moet je in het onvoldoendegebied als het ware wat 'indikken', dat wil zeggen meer fouten een punt (dus soepeler) gaan rekenen, zoals te zien is in deze tabel.

aantal gemiste scorepunten van max. 10	aftrek	cijfer	afgerond en niet lager dan een 1	'ingedikt ondereinde van de schaal' (voorbeeld)
0	0	10	10	
1	1,33	8,66	8,5 of 9	
2	2,66	7,34	7,5 of 7	
3	3,99	6,01	6	
4	5,32	4,68	4,5 of 5	5
5	6,65	3,35	3 of 3,5	4
6	7,98	2,02	2	3
7	9,31	0,69	1	3
8	10,64	- 0,5	1	2
9	11,97	- 1,97	1	2
10	13,33	- 3,33	1	1

Schema 5.1 **Voorbeeld: omzetting van aftrekpunten in cijfers**

De behoefte aan precieze differentiatie tussen prestaties is in het gebied van de onvoldoendes meestal minder dan in het gebied van de voldoendes. Er zijn landen waar men in het nationale cijfer- of lettersysteem helemaal geen verschil maakt tussen een betere en een slechtere onvoldoende. Zo bestaat in de Verenigde Staten voor 'onvoldoende' maar één waardering, een 'D'.

5.4.3 Cesuur: cijfers op basis van aftrekpunten

In het taalonderwijs wordt bij de productieve vaardigheden, waarbij de toetsvorm geen opdeling in items toelaat, met name bij schrijfvaardigheid, vaak een variant op het percentagecriterium toegepast, namelijk beoordeling op grond van aftrekpunten voor gemaakte fouten. Die kun je ook weer wegen, bijvoorbeeld voor woordfouten meer aftrekpunten dan voor grammaticafouten.

In feite richt je je dan naar het aantal (gewogen) fouten dat de slechtste leerling gemaakt heeft. Die prestatie geef je een slecht cijfer, bijvoorbeeld een 1, maar je kunt ook wat milder zijn en er een 2 of 3 van maken. Anders uitgedrukt: je rekent zoveel fouten per punt dat de slechtste leerling die 1, 2 of 3 krijgt. Een rekenkundig beter werkende variant hierop is dat je de berekening niet op de slechtste leerling baseert, maar op de leerling die je als nét voldoende inschat; die zet je op een 6, dus op de cesuur. Uitgaand van het aantal fouten per punt voor die net voldoende leerling kun je dan de cijfers voor de overige leerlingen berekenen. Duidelijk mag zijn dat dit een niet erg betrouwbare methode is.

Bij dit systeem zijn verschillen in omvang van het te beoordelen product problematisch. De leerling die in een schrijfvaardigheidstoets (veel) meer woorden schrijft dan anderen, loopt daarmee het risico (veel) meer fouten te maken. Er zijn in principe twee manieren om die ongelijkheid te reduceren: je zorgt voor een ongeveer gelijke omvang door de schrijfopdracht sterk(er) geleid te maken en/of door een bepaald minimum- en maximumaantal woorden voor te schrijven, en/of je verrekent het aantal woorden in de beoordeling: wie meer woorden schrijft, mag meer fouten per punt maken.

5.4.4 Cesuur: indrukscijfers, al of niet in combinatie met aftrekpunten

In het vreemdetalenonderwijs wordt ook vaak globaal getoetst; de prestatie wordt dan beoordeeld met een indrukscijfer, zoals heel gebruikelijk is bij gespreksvaardigheid. Indrukscijfers gaan zelden hoger dan een 9 (meestal kan alleen een native speaker een 10 krijgen). Bij indrukscijfers is er geen 70-procentcesuur, maar een veel vagere voldoende-onvoldoendegrens die gebaseerd is op een globale vergelijking tussen (vergelijkbare) leerlingenprestaties.

Meer betrouwbaarheid kun je bereiken, ook bij indrukscijfers, door aspecten te onderscheiden, die afzonderlijk met een indrukscijfer te waarderen, de aspecten een weging ten opzichte van elkaar te geven en op basis daarvan een totaalcijfer te berekenen. Daarmee kun je ook de leerling meer feedback geven over zijn sterke en zwakke punten.

Ook mogelijk is een combinatie van indrukscijfers met beoordeling op goede en foute items of aftrekpunten. Bij schrijfvaardigheid bijvoorbeeld ligt het voor de hand om de *inhoud* en/of de *opbouw* (als die niet voorgegeven zijn) met een indrukscijfer te beoordelen, en de *(taal)vorm* met aftrekpunten.

Wat betreft de taalvorm worden vaak deelvaardigheden onderscheiden:
- uitspraak (bij spreken) of spelling (bij schrijven);
- grammatica: zinsbouw (syntaxis) en woordvormen/uitgangen (morfologie);
- woordenschat/idioom (lexicon).

Plus een algemeen criterium:
- vlotheid/vloeiendheid (bij spreken), stijl (bij schrijven).

Een voorbeeld voor de weging bij schrijfvaardigheid: een woordfout telt als één fout (aftrekpunt), een grammaticafout als een halve fout, een spelling- of interpunctiefout als een kwart fout. Uitspraak en vlotheid/vloeiendheid (bij spreken) en stijl (bij schrijven) worden vervolgens voor de toets als geheel beoordeeld met een indrukscijfer of -score. Hetzelfde kan gelden voor inhouds- en/of ordeningsaspecten. Met die indrukscijfers, al dan niet onderling gewogen, kan de leerling een bepaald aantal bonus- of maluspunten verdienen.

Bij gespreksvaardigheid wordt meestal veel globaler beoordeeld, vaak op basis van een indeling in vier of vijf aspecten waarvoor indruksscores (eventueel met verschillende weging) gegeven worden.

5.5 Gevaren bij toetsing

De moraal van dit verhaal luidt: toetsen is een hachelijke zaak. En dat terwijl de beoordelingen belangrijke effecten voor de leerlingen hebben. Maar als leraar moet je wel. Hier wordt veel gevraagd van de deskundigheid en van de beroepsmoraal van de leraar.

Heel belangrijk is uiteraard dat je criteria en je beoordelingsbeslissingen duidelijk zijn voor jezelf, maar ook voor anderen, namelijk:
- voor de leerlingen, in het belang van een goede feedbackwerking: de leerling moet zo precies mogelijk weten wat wel en niet goed is aan zijn prestatie en wat hij had kunnen doen of nog moet doen om zich te verbeteren;
- voor de collega-beoordelaars, in het belang van een zo objectief mogelijke beoordeling met het oog op het curriculaire en civiele effect.

5.5.1 Pedagogisch cijferen

Beoordeling heeft naast een 'technisch-didactische' ook altijd een emotionele en motivationele kant. De leerling kan door een positieve beoordeling worden gemotiveerd, maar ook door een negatieve beoordeling. Iedere leraar komt weleens in een situatie waarin hij een prestatie minder zou willen belonen, omdat die geleverd is door een leerling die er niets voor gedaan heeft en die met meer inspanning meer had kunnen presteren. Ook klassen als geheel kunnen zulk gedrag vertonen. Als voorbeeld een authentiek citaat uit een

stageverslag in de lerarenopleiding: 'Tip van mijn coach: als je strenger toetst (minder fouten een punt rekent), zal deze klas jouw vak serieuzer nemen en er meer voor doen.' Toetsing lijkt hier te worden gebruikt als disciplineringsmiddel. Kan dat wel?

Ook het omgekeerde komt voor: een leerling heeft een prestatie geleverd die in verhouding tot de norm van de klas en het schooltype niet zo geweldig is, maar waarvoor hij echt z'n best gedaan heeft. Voor die inzet wil je dan wel een wat hoger cijfer geven. Of een leerling is getroffen door een ongeluk of een andere bijzondere omstandigheid en heeft daarom geen normale inzet kunnen leveren, dus je wilt als leraar je hand over je hart strijken: die leerling moet een opsteker krijgen. Kan dat, of is dat onrechtvaardig tegenover de andere leerlingen? Heeft pedagogisch cijferen geen verkeerde gevolgen voor het curriculaire en uiteindelijk ook voor het civiele effect?

Ja, dat heeft het. Het uitgangspunt bij meten en beoordelen is: dezelfde criteria voor iedereen. Als je een factor als inzet mee wilt laten tellen: prima, maar doe dat dan ook bij iedereen. Bijzondere omstandigheden komen per definitie niet bij iedereen voor. En stel dat iemand niet slechts bij één of twee lessen en/of proefwerken afwezig is vanwege ziekte of de gevolgen van een ongeluk, maar vijf of tien of twintig keer? Moet iemand die door omstandigheden een half of heel jaar niet mee kan doen, alle toetsen van die periode cadeau krijgen? Kun je voor die persoon de lat lager leggen dan voor de anderen? Het diploma vermeldt uiteindelijk niet dat er andere criteria gehanteerd zijn.

Pedagogisch cijferen kan eigenlijk niet. Als je het toch doet, doe het dan open en in overleg met je collega's. Bedenk dat de criteria bij pedagogisch cijferen aanzienlijk subjectiever zijn dan bij het gewone cijferen. Bedenk ook dat je er precedenten mee schept, zodat andere leerlingen – terecht – ook eisen kunnen gaan stellen. Openheid en duidelijkheid hierin zijn onontbeerlijk.

5.5.2 Halo-effect

Heel lastig bij toetsing is ook het zogenoemde halo-effect: een bijkomstig verschijnsel overstraalt datgene waar je eigenlijk naar hoort te kijken. Twee handgeschreven brieven hebben dezelfde kwaliteit, maar de ene is keurig geschreven, de andere is nauwelijks leesbaar door de hanenpoten en de doorhalingen. Kijkt de docent ze allebei even mild of streng na? En hoe gaat het bij de gespreksvaardigheidstoets waar de ene leerling zich met charme doorheen slaat en de andere nauwelijks doorheen te slepen is? Leerlingen die altijd al goed scoren en/of van wie leraren denken dat ze goed zijn, krijgen al snel een beter cijfer dan leerlingen bij wie het tegendeel het geval is – vooral bij indrukscijfers natuurlijk. De goede leerling heeft als het ware een bijzonder aureool of 'heiligenschijn' waardoor alles wat hij presteert een extra glans en bij voorbaat een

betere beoordeling krijgt. Bij talen is het trouwens vaak een 'zij', omdat meisjes weleens profiteren van het verschijnsel dat ze veelal goed in talen zijn.

Dit soort effecten kan natuurlijk zowel in positieve als in negatieve richting werken. Ze kunnen sterker optreden naarmate meting en beoordeling globaler zijn.

5.5.3 De leraar die zijn eigen onderwijs toetst

Stel: een vreemdetaalleraar vindt gespreksvaardigheid lastig om te oefenen. Vroeger, toen spreekvaardigheid alleen in het examenprogramma stond en de docent zelf kon bepalen wanneer hij daarmee begon te oefenen, werd het eerste spreekvaardigheidscijfer pas in de voorexamenklas of zelfs pas in het examenjaar gegeven. Sinds in 1993 de gespreksvaardigheid als één van de basisvormingskerndoelen genoemd werd, wordt de gespreksvaardigheid vanaf jaar 1 geoefend en getoetst. De leraar van ons voorbeeld doet er toch weinig aan. Er kan bij komen dat hij de taal die hij onderwijst, zelf niet zo makkelijk spreekt. De beoordelingscriteria voor de gespreksvaardigheidstoetsen zijn open en vaag. Als hij die toetsen zelf afneemt en beoordeelt, zal hij geen hoge eisen stellen. Hij beoordeelt namelijk niet alleen de prestaties van de leerlingen, maar ook – of misschien vooral – zijn eigen (gespreksvaardigheids) onderwijs.

Om onbetrouwbaarheid in de toetsing tegen te gaan heeft de Nederlandse overheid verordend dat de helft van het examencijfer in een landelijk, centraal examen bepaald wordt. Maar bij de moderne vreemde talen is voor dat centrale examen slechts één vaardigheid uitgekozen, de leesvaardigheid. Er is weinig overlap met de andere onderdelen van het vak, namelijk de drie overige vaardigheden (plus eventueel literatuur), die in het schoolexamen getoetst worden. Bij veel andere vakken is meer overlap tussen centraal examen en schoolexamen.

Dit kan ertoe leiden dat sommige vreemdetaaldocenten heel veel oefentijd besteden aan leesvaardigheid, misschien vrijwel uitsluitend in de examenvorm: Cito-toetsen. En omdat ze de andere vaardigheden minder oefenen, stellen ze in het schoolexamen ook navenant lagere eisen. Zo houd je als docent je onderwijsresultaten, althans je examencijfers, wel op peil. En dat geldt niet alleen voor het examen, maar ook voor de rapportcijfers van alle klassen daarvóór. Het is niet uitgesloten dat de ene docent, de ene sectie of de ene school niet alleen in de beoordeling, maar ook in het onderwijs heel andere accenten legt dan de andere. Toch ziet een 6 er op de ene cijferlijst net zo uit als op de andere.

De overheid benadrukt sterk de autonomie van de scholen. Ze moeten het onderwijs allemaal op hun eigen manier vormgeven. De overheid heeft tegelijkertijd de taak om de kwaliteit van dat onderwijs en het eindniveau van de leerlingenprestaties te bewaken, en dat doet ze vooral via kerndoelen en

eindtermen in de examenprogramma's. Daarbinnen hebben docenten, secties en scholen veel vrijheid. Wie weinig eisen stelt aan zijn eigen onderwijs, zal ook weinig eisen stellen aan zijn leerlingen. Niet alleen van de overheid, met name de onderwijsinspectie, maar ook van de docenten, secties en scholen zelf wordt veel professionaliteit gevraagd om dit ongewenste effect zoveel mogelijk te vermijden.

5.5.4 Meer toetsen dan je onderwijst, en meer onderwijzen dan je toetst

Een ander groot probleem van toetsing is de verhouding tussen de onderwijsdoelen en de toetsing. 'Toets wat je onderwijst en onderwijs wat je toetst' – daar zit veel in. Het is niet eerlijk om in de toets leerlingen te beoordelen op zaken die je ze niet bijgebracht hebt. Maar deels is dat onvermijdelijk. Omgekeerd is het fair dat je zoveel mogelijk onderwijst wat je toetst. Alleen moet je daaruit niet de conclusie trekken dat het genoeg is te onderwijzen wat je toetst. Je moet meer onderwijzen dan je kunt toetsen.

Deze nogal filosofische uitspraken vragen om uitleg.

Meer toetsen dan je onderwijst

Waarom kun je niet altijd vermijden dat je zaken meetoetst en meebeoordeelt die je niet onderwezen hebt?

Bij taalvaardigheidstoetsen gaat het altijd ergens over. Dus heeft de leerling behalve taalkennis en taalvaardigheid ook kennis van de wereld nodig. Die heeft hij voor het allergrootste deel opgedaan buiten jouw les en zelfs buiten de school. Dat wordt dus problematisch als je leerlingen met verschillende thuisculturen in je klas hebt, of die nu autochtoon of allochtoon zijn.

Verder kun je denken aan allerlei vaardigheden die een leerling in allerlei andere situaties kan hebben verworven of spontaan heeft ontwikkeld en die het resultaat op jouw toetsen behoorlijk kunnen beïnvloeden. Denk bijvoorbeeld aan de vaardigheid om snel de hoofdpunten uit een tekst te halen, ook als jij dat in jouw lessen zelden of nooit oefent.

Soms kun je je onderwijs op dit vlak (deels) aanpassen en bijvoorbeeld meer aan samenvatten doen.

Omgekeerd kun je ervoor zorgen dat je toetsvormen zo weinig mogelijk elementen bevatten die jij niet onderwijst. Een voorbeeld: sommige toetsvormen doen een beroep op specifieke karaktereigenschappen terwijl jouw onderwijs weinig of niets bijdraagt aan de vorming daarvan. Als je voor gespreksvaardigheid in het schoolexamen kiest voor een open gespreksvorm, en misschien zelfs als beoordelingscriterium opneemt: 'Neemt vaak/soms/weinig het woord', dan beoordeel je de spraakzaamheid en de assertiviteit van de kandidaten mee. Dan hoort dus zoiets als 'spraakzaamheidstraining' of assertiviteitstraining ook op jouw onderwijsprogramma te staan.

Meer onderwijzen dan je toetst

Waarom is het niet genoeg om te onderwijzen wat je toetst?

Heel wat zinvolle leerdoelen zijn moeilijk of niet te meten. Het is heel verleidelijk voor een docent om zich in zijn onderwijs te concentreren op, of zich zelfs te beperken tot, het makkelijk toetsbare, vooral woordenschat en grammatica. Ook veel leerlingen vragen daar direct of indirect om. Ordeproblemen kunnen je ertoe brengen om veel leer- en maakwerk op te geven en daar je toetsen mee te vullen. De vaardigheidstoetsen kunnen een bijna lege formaliteit worden. Want hoe kun je een leerling een onvoldoende geven voor bijvoorbeeld gespreksvaardigheid als je die bijna nooit oefent?

Denk ook eens aan allerlei doelen die niet in examenprogramma's staan en die je toch als school, als sectie en/of als docent kunt nastreven. Bijvoorbeeld dat leerlingen leren om samenvattingen te maken, dat ze zich openstellen voor vreemdtalige culturen en mensen, dat ze leren samen te werken bij groepswerk enzovoort. Doelen in de sfeer van de algemene vaardigheden en in de sociale en attitudinele sfeer zijn moeilijk of niet te toetsen. Gelukkig ervaren veel docenten het als een verarming alleen te onderwijzen wat je toetst. De school leidt, als het goed is, niet op voor de toets of het examen, maar voor het leven. De school en de hele maatschappij kunnen niet zonder toetsen en examens, maar het blijven gebrekkige hulpmiddelen om vast te stellen wat iemand kan.

5.6 Tips bij toetsing

1 *Zorg voor duidelijke criteria*
Doe dit zowel voor je collega's als voor de leerlingen. Bedenk: de opgaven vormen niet meer dan de helft van de toets – een thermometer zonder schaal. Een toets is pas een toets met inbegrip van de verwachte antwoorden en van de beoordelingscriteria: het meet- en beoordelingsinstrument (de correctie- en becijferingsrichtlijnen).

2 *Meet en beoordeel zo weinig mogelijk globaal*
Hoe complexer de (deel)vaardigheid die je wilt toetsen is, des te sterker wordt de neiging tot globaal beoordelen. Maar dat heeft twee belangrijke nadelen:
1 De leerling krijgt uit de globale beoordeling – bijvoorbeeld alleen een cijfer zonder toelichting hoe dat tot stand gekomen is – weinig of geen feedback over wat er goed of fout was en hoe hij zijn prestaties kan verbeteren.
2 De betrouwbaarheid – in vergelijking met collega's en met jezelf op andere tijdstippen en/of met andere leerlingen – komt in gevaar.

3 *Meet en beoordeel zoveel mogelijk analytisch*
Hoe complexer de te toetsen (deel)vaardigheid is, des te belangrijker wordt het dat je analytisch meet en beoordeelt. Dat maakt het makkelijker leerlingenprestaties onderling te vergelijken en leerlingen bruikbare feedback te geven. Manieren om dat te bereiken:
1 complexe en omvangrijke opdrachten structureren, dat wil zeggen indelen in deelopdrachten (items) en/of deelaspecten, die per stuk te beoordelen zijn;
2 fouten categoriseren in soorten en de verschillende foutencategorieën, deelopdrachten en aspecten een weging geven ten opzichte van elkaar.

4 *Gebruik zoveel mogelijk gezamenlijke en/of gestandaardiseerde toetsen*
Dat zijn toetsen die afgenomen worden bij en/of eerder uitgeprobeerd zijn op zoveel mogelijk leerlingen van dezelfde doelgroep.

5 *Beoordeel zoveel mogelijk samen*
Je kunt de betrouwbaarheid van je meting en beoordeling verhogen door zoveel mogelijk communicatie met collega-beoordelaars. In principe zou elke complexere toets niet door één, maar minstens door twee mensen beoordeeld moeten worden. Dat is natuurlijk in de praktijk niet doenlijk. Wel te doen is het volgende:
1 Zorg dat je met je collega's duidelijke en gedetailleerde afspraken hebt over meting (correctie) en beoordeling (scoring, weging, normering/cesuurvaststelling) van overhoringen, proefwerken en repetities, schoolexamens in alle klassen van alle schooltypes.
2 Laat af en toe jouw correctie en beoordeling becommentariëren door een ervaren collega.
3 Neem af en toe de tijd om met alle sectiecollega's oefenbeoordelingssessies te houden: je beoordeelt samen een paar leerlingenproducten (overhoringen, proefwerken, schoolexamens); dat wil zeggen iedereen beoordeelt ze individueel, en vervolgens spreek je samen de beoordelingen door. Daarbij kom je allerlei interessante dingen tegen; je leert er veel van.

6 *Vergelijk leerlingenprestaties*
Je kunt de betrouwbaarheid van je meting en beoordeling verhogen door te zorgen dat je de leerlingenprestaties onderling goed kunt vergelijken.
Bij een schriftelijke toets: lees de geschreven producten (of ten minste een aantal ervan) eerst globaal door voor een eerste indruk. Kijk ze dan na aan de hand van je criteria (meting). Bij meer omvangrijke en complexe toetsen is de vergelijking makkelijker als je eerst onderdeel 1 bij alle leerlingen nakijkt, dan onderdeel 2 enzovoort. Maak ten slotte een tabel van de scores en bereken de cijfers (beoordeling).

Bij een mondelinge toets (gespreksvaardigheid): neem de toetsen (of ten minste een aantal) op een geluidsdrager op en ga er op dezelfde manier doorheen als door een schriftelijke toets. Vermijd in elk geval als beginner zoveel mogelijk de situatie dat je meteen na de eerste toets(en) een beoordeling moet geven: je hebt dan nog geen vergelijkingsmogelijkheden. Dit is enorm tijdrovend, maar voor de beginner in toetsing zeer aan te raden als je geen beoordelingsongelukken wilt maken.

7 *Toets wat je onderwijst, maar onderwijs niet alleen maar wat je toetst*

8 *Maak geen misbruik van je beoordelingsbevoegdheid*
Juist omdat meting en beoordeling van (vreemde)taalvaardigheid zo moeilijk en discutabel zijn, is het belangrijk dat je als leraar geen misbruik maakt van de beoordelingsbevoegdheid die de maatschappij in je handen legt, met zulke belangrijke consequenties voor de leerlingen. Je hebt daarmee een zekere mate van macht. Het is dus onderdeel van je beroepsmoraal dat je jezelf steeds de volgende vragen stelt:

1 Laat ik mij, ook in de sfeer van de minder controleerbare overhoringen, niet verleiden tot heimelijk gebruik van toetsing voor disciplinaire doeleinden?
2 Vaar ik als beoordelaar geen eigen koers, waardoor ik andere eisen aan de leerlingen stel dan collega's aan mijn eigen school of aan andere scholen? Bijvoorbeeld door mild te oordelen over aspecten die ik zelf moeilijk te onderwijzen vind?
3 Ben ik zo eerlijk en zorgvuldig mogelijk in mijn beoordelingen, en kan ik die beoordelingen altijd verantwoorden tegenover de leerling, diens ouders, de collega's, de schoolleiding, de overheid (inspectie)?

6 Curriculum en leergang

6.1 Curriculum, leerplan, examenprogramma, kerndoelen, programma van toetsing en afsluiting

Het curriculum van een opleiding (schooltype) of een (school)vak is het geheel van doelstellingen, opeenvolging van leeractiviteiten en toetsen van die opleiding of dat vak. Een leerplan is een document waarin dat curriculum beschreven wordt. In veel landen vaardigt de overheid leerplannen uit waarin ze voorschrijft wat er in elk vak in de verschillende schooltypes in de verschillende klassen gebeuren moet. Dat heeft uiteraard een sturende werking op de leergangen, maar de leerplannen zijn altijd abstract genoeg om ruimte te laten voor verschillende concrete invullingen. Vandaar dat veel landen daarnaast een leermiddelencontrole hebben: leergangen en andere leermiddelen moeten goedgekeurd zijn om gebruikt te mogen worden in de scholen, althans voor zover die openbaar zijn en/of uit openbare middelen betaald worden.

Zo niet in Nederland. Leerplannen waarin de overheid voorschrijft hoe het onderwijs ingericht moet of kan worden, zijn er wel geweest, maar ze waren vaak erg globaal en werden weinig nageleefd. Leermiddelencontrole door de overheid is politiek niet bespreekbaar. Deze traditie heeft alles te maken met de negentiende-eeuwse schoolstrijd, waarin de confessionele bevolkingsgroepen (katholieken en gereformeerden) hun recht op eigen, 'bijzondere' scholen bevochten, naast de openbare. De 'vrijheid van onderwijs' stond al sinds 1848 in de grondwet, maar pas in 1917 werd de schoolstrijd afgesloten met een politiek akkoord en een grondwetswijziging, waarbij de bijzondere scholen dezelfde financiële rechten kregen als de openbare.

Met de vrijheid van onderwijs werd en wordt, iets preciezer, eigenlijk alleen de vrijheid van inrichting of vormgeving van het onderwijs bedoeld. De overheid is namelijk, behalve voor de financiering, ook verantwoordelijk voor de globale organisatie van het onderwijs (schooltypes, profielen en vakkenpakketten) en voor de resultaten, dus voor het niveau van de examens. Daarover waakt de inspectie. De uitgereikte diploma's hebben immers een civiel