

Toetsen van taalvaardigheid

Dames en heren,

Toetsen is niet makkelijk, en toetsen van taalvaardigheid is al helemaal niet makkelijk. Daarover zijn we het allemaal wel eens, denk ik. En, zodra we de zaak meer handen en voeten willen geven, dan komen nog veel meer verschillen in visies, denkwijzen, gewoonte naar voren. Het is zo complex dat ik me vandaag beperk tot één aspect: summatieve toetsing; toetsing om na te gaan of een kandidaat aan vastgestelde criteria voldoet.

Een taaltoets moet vanzelfsprekend betrouwbaar zijn. Dit geldt natuurlijk voor alle toetsen; een kandidaat die vandaag slaagt, moet morgen –bij een vergelijkbare toets- ook slagen. Dit is vaak een technische aangelegenheid. We kunnen de betrouwbaarheid uitrekenen en uitdrukken in een coëfficiënt. Als deze coëfficiënt maar hoog genoeg is, dan is de toets betrouwbaar, en kunnen we zeker zijn dat als we deze, of een hoogst vergelijkbare toets nogmaals afnemen, dat we dan tot hetzelfde oordeel komen.

Betrouwbaarheid zegt iets over de precisie van een toets; bij een betrouwbare toets zijn kleine verschillen in scores te interpreteren, en bij een onbetrouwbare toets kan zelfs aan grote verschillen in cijfers geen betekenis toegekend worden. Om kandidaten te kwalificeren is een betrouwbare toets dus een *conditio sine qua non*. Als het met de betrouwbaarheid niet ok is, dan kunnen we wel ophouden. Dit geldt niet alleen voor toetsen, maar ook voor beoordelingen; als beoordelaars het niet met elkaar eens zijn, dan is er zo'n probleem dat effectieve kwalificering ONMOGELIJK is. Om de invloed van de betrouwbaarheid op het geven van cijfers te adstrueren is het volgende voorbeeld zinvol. Stel, we hebben een heleboel tekstbegrip toetsen. Deze toetsen zijn allemaal identiek, behalve de tekst waarbij de vragen gesteld worden. En, we hebben een heel gelukkige leerling die alle toetsen maakt. De verschillen in cijfers tussen deze toetsen kan nu opgevat worden als een maat voor de betrouwbaarheid waarmee leesvaardigheid gemeten wordt. Immers, het enige dat verschilt

tussen de toetsen is de tekst waarbij de vragen gesteld worden. Als we alle cijfers van deze leerling op een rijtje zetten krijgen we een figuur als de volgende:

HIER FIGUUR 1

Het blijkt dat het cijfer nogal varieert ten gevolge van de tekst waarbij de vragen gesteld worden. Deze ene leerling, die in de ‘gelukkige’ omstandigheid verkeerde duizend tekstbegriptoetsen te mogen maken heeft gemiddeld een 6 gehaald, maar zijn cijfer varieert van een 3 tot een 9, enkel en alleen ten gevolge van de tekst waarbij de vragen gesteld zijn. Maar wat is nu de ‘echte’ leesvaardigheid van deze leerling? Is dat een 3, een 9 of een 6? We weten het niet precies, maar hoewel een 6 de meest waarschijnlijke score voor deze leerling is, moeten we er volgens mij een andere conclusie aan verbinden. We kunnen NIET zonder meer blind varen de op de resultaten van één tekstbegriptoets gebaseerd op één tekst met vragen. Het is dan ook niet zo verwonderlijk dat het Cito voor de examens in de MVT in het voortgezet onderwijs altijd bij meerdere teksten bevraagd!

Nu ben ik in dit voorbeeld er vanuit gegaan dat de toetsen nagenoeg identiek zijn, behalve de tekst dan. Echter, naarmate de verschillen tussen toetsen groter worden, hoe lager de samenhang. We meten dan immers met de verschillende toetsen meer verschillende zaken. Dat beïnvloedt de samenhang tussen de toetsen enorm, en de betrouwbaarheid wordt navenant lager. Als we dit toepassen op ‘leesopdrachten-waarbij-meer-komt-kijken’, meer functionele leesopdrachten, of wat wel praktijk opdrachten genoemd wordt, dan is duidelijk dat die minder hetzelfde zijn. Daardoor zijn de verschillen tussen opdrachten nog groter. We betalen dan een prijs voor de functionaliteit in termen van betrouwbaarheid en waarschijnlijk ook in termen van validiteit. Dit is in feite een ‘giga’ probleem. We kunnen doodeenvoudig niet zo maar toetsen maken die betrouwbaar zijn. Daarvoor moeten deze toetsen uitgetoetst, bijgesteld en vaak nog eens

uitgeprobeerd worden. Ik heb in elk geval nog nooit een toets gezien die in één keer perfect was.

Een tweede voorbeeld ontleen ik aan meer experimenteel onderzoek. In deze studie hebben we groepen leerlingen uit de theoretische leerweg (TL) en leerlingen uit de basisberoepsgerichte leerweg (BL) verschillende tekstbegriptoetsen met een aantal meerkeuze-vragen voorgelegd. De uitkomsten heb ik wederom in een grafiek gezet.

HIER FIGUUR 2

De TL-leerlingen hebben gemiddeld de hoogste score behaald op de eerste toets. Op deze toets doen de BL-leerlingen het gemiddeld beduidend minder. Op de tweede toets zijn de gemiddelden echter omgekeerd: de BL-leerlingen hebben het beduidend beter gedaan dan de TL-leerlingen. Toch opmerkelijk; de TL-leerlingen doen het goed op de ene toets en de BL-leerlingen doen het goed op de andere toets! Kennelijk maakt het nogal wat uit WELKE tekstbegriptoets de leerlingen voorgelegd krijgen.

Maar laat ik nu een tipje van de sluier oplichten. De BL-leerlingen waren koks-in-opleiding. De toetsen waren: een traditionele tekstbegriptoets en vragen bij recepten. Dus, de TL-leerlingen doen het goed bij de traditionele tekstbegriptoets, en minder bij de recepten. De koks-in-opleiding doen het juist goed bij de receptentoets, maar minder bij de traditionele tekstbegriptoets. Kennelijk ondervinden de koks-in-opleiding voordeel van hun algemene kennis over recepten en de structuur van recepten. Onze meting van de leesvaardigheid van deze leerlingen wordt dus nogal beïnvloed door de aanwezige kennis; hoe meer kennis van hetgeen je leest, hoe beter je kan lezen. Maar ook de structuur van recepten is van groot belang. Dat is als het ware een soort van tekstschema hoe een recept in elkaar moet zitten. Als we deze ‘vaste’ structuur doorbreken, dan daalt de score van de BL-leerlingen aanzienlijk, zoals in verschillende experimenten aangetoond is. De les die we hieruit kunnen en moeten trekken is dat het van

cruciaal belang is, hoe we het object in kwestie definiëren. En, natuurlijk dat we materiaal zoeken, of maken, dat naadloos bij onze definiëring aansluit.

Met deze verklaring(en) hebben we eigenlijk het terrein van de betrouwbaarheid verlaten, en zijn we ongemerkt naar de validiteit opgeschoven. Validiteit is een paraplubegrip, waaronder diverse aspecten van de validiteit vallen: gezichtvaliditeit, constructvaliditeit, predictieve validiteit, discriminant validiteit, convergente validiteit om er maar een paar te noemen. Elk onderdeel van de validiteit heeft zijn eigen definitie, zijn eigen problemen, en zijn eigen methodes om aan te tonen in hoeverre een specifieke toets in dit opzicht valide genoemd kan worden.

In het algemeen verwijst validiteit naar de vraag of datgene gemeten is wat men daadwerkelijk wilde meten. Toon aan dat deze toets leesvaardigheid meet en er is aangetoond dat deze leestoets constructvalide is. Maar hoe toon je een construct, een niet observeerbaar iets, even aan. Dat is geen sinecure. Sommigen zijn een half mensenleven bezig met het aantonen dat één bepaalde toets valide is.

De meest zwakke vorm van validiteit is zonder meer de gezichtsvaliditeit, of indrukvaliditeit. Dit komt er op neer dat iemand zegt: 'ik vind deze toets valide', en dan is de toets dat ook! Klaar, een valide toets in no time! Helaas, pindakaas, zo simpel is het niet! We hebben ons in dit opzicht al veel te vaak vergist. Een mooi voorbeeld is het centrale examen Nederlands voor HAVO/VWO. In dit examen, waarin alleen tekstbegrip getoetst wordt, speelt het onderdeel samenvatten nog steeds een belangrijke rol. Wat weinigen meer lijken te weten is dat samenvatten vroeger een onderdeel was dat geacht werd schrijfvaardigheid te meten; bij het vrije opstel spelen verschillen in kennis een grote rol, en in dit onderdeel van het examen Nederlands is het niet de bedoeling kennis te meten, maar de schrijfvaardigheid van de leerlingen. Vandaar de samenvatopdracht, want daarin speelt verschil in (voor)kennis veel minder een rol. Nu ken ik geen onderzoek waarin wordt aangetoond dat we ons vergist hebben, en dat samenvatten géén schrijfvaardigheid maar leesvaardigheid meet. Toch

wordt de samenvatting nu geacht leesvaardigheid te meten, en vele deskundigen vinden nù dat de samenvatting leesvaardigheid meet; de samenvatting is dan ook een gezichtsvalide meting van de leesvaardigheid. De enige conclusie die je hier, volgens mij, aan kan verbinden is dat er altijd wel iemand te vinden is die A vindt, of die niet-A vindt. Als je maar lang genoeg zoekt.

Wat we nù weten, op grond van de gegevens in het centrale examen, is dat het onderdeel samenvatten niet correleert met het andere onderdeel voor tekstbegrip: de tekst met vragen. Dat is gek, want we hebben twee metingen voor één construct: leesvaardigheid. En je zou toch op zijn minst enige correlatie tussen de cijfers voor het ene onderdeel en de cijfers voor het andere onderdeel mogen verwachten. Als die verwachting niet uitkomt, dan meten beide onderdelen niet hetzelfde, en is tenminste één van beiden niet constructvalide. Constructvaliditeit is wat dat betreft oneindig veel belangrijker, en veel meer zeggender dan gezichtsvaliditeit. Eigenlijk, maar dat is mijn particuliere mening: gezichtsvaliditeit zegt meer over degene die zich uitspreekt dan over het instrument in kwestie.

De essentie van validiteit is dat we meten wat we willen meten. En, dat is knap lastig; wat valt er wèl onder en wàt valt er niet onder? Nu weet ik niet precies wat taalvaardigheid is. Ik ben in de luxe situatie dat ik me vaak kan verschuilen achter redelijk academische definities zoals: ‘taalvaardigheid is een relatief constant attribuut wat ik aan personen toe ken’. Maar daar komen we niet zo ver mee. Wat ik wèl gedeeltelijk weet, is wat het niet is: het is geen grammatica, het is géén spelling, het is niet alleen woordkennis, etc. Bij taalvaardigheid staat bij mij begrip voorop. Een taalvaardig persoon zal beter begrijpen wat er bedoeld wordt, welke boodschappen er over gedragen worden, of welke boodschappen er over gedragen moeten worden, dan een minder taalvaardige kandidaat.

Zoeven zagen we dat bij verschillende tekstbegriptoetsen, er enkel en alleen door het variëren van de tekst nogal wat variatie in cijfer ontstaat. Klaarblijkelijk zijn er andere factoren die de score op de toets beïnvloeden. Over het algemeen willen we dergelijke factoren zo veel mogelijk

buitensluiten. We willen immers tekstbegrip evalueren, en geen zaken als kennis van het onderwerp, of kennis van het teksttype, tenminste zo werd lang gedacht. Het idee was dat dergelijke factoren zoveel mogelijk uitgeschakeld meten worden. De huidige definities van lezen zijn iets als het adequaat leggen van verbanden tussen tekstelementen, én het leggen van relaties met kennis over het in de tekst aangesneden onderwerp. Een dergelijke mentale representatie van een tekst gekoppeld aan algemene kennis, wordt wel een situatiemodel genoemd. De omarming van een situatiemodel betekent een andere manier van toetsen; immers ook de relatie met al aanwezige kennis, en integratie daarin moet meegenomen worden. Maar, let wel, het kan nooit de bedoeling zijn dat alleen die kennis centraal staat. Dat moet op een geheel andere wijze getoetst worden dan tekstbegrip. Voor het toetsen van tekstbegrip rekeninghoudend met de kennis van leerlingen loopt nu een promotieproject op de Universiteit Utrecht. In dit project wordt de leesvaardigheid van Basisberoepsgerichte- en Kaderleerlingen getoetst allerlei andere soorten vragen, zoals: met tijdsbalkvragen, vragen waarbij de volgorde van elementen centraal staat, of schema-vragen, waarbij juist de relaties tussen tekstelementen (en algemene kennis) centraal staat.

Ook bij het meten van schrijfvaardigheid staan we voor een vergelijkbaar dilemma. Probeer u maar eens een tekst te schrijven van 2 A4 over het functioneren van de overheid ten tijde van Toetanchamon. Ik durf er gif op in te nemen dat vrijwel niemand dat lukt. Gewoon omdat we te weinig kennis hebben van dat onderwerp (of misschien spreek ik wel alleen voor mezelf). Dus: ook bij schrijven speelt kennis een enorm belangrijke rol. Iemand die veel weet over een onderwerp kan daar makkelijker over schrijven van iemand die daar weinig van weet. Echter, en dat is natuurlijk de ellende, we willen niet de verschillen in kennis evalueren, maar juist de verschillen in schrijfvaardigheid (en dan doel ik natuurlijk niet op spelling en zo). Hiervoor zijn talloze dingen uitgetoetst. Het beste lijken nog opdrachten à la gericht schrijven, waarbij de schrijver de beschikking krijgt over achtergrondmateriaal waar hij gebruik van kan maken. Of, in een

project waarbij het vinden van informatie apart op het programma staat voordat er geschreven gaat worden. Er zijn dus aparte maatregelen nodig voordat we verschillen in schrijfvaardigheid zichtbaar kunnen maken.

Maar, stel dat we dat allemaal gedaan hebben, en dat onze leerlingen geschreven hebben, dan nog moeten we de schrijfsels beoordelen. En dat is nog altijd lastig. Velen wijzen voor beoordeling naar het Europese Referentie Kader (ERK). Daarin zijn immers duidelijke omschrijvingen van taalvaardigheidsniveaus te vinden, die een houvast bij het beoordelen zouden kunnen bieden. Het Europese Referentie Kader wordt regelmatig als panacee tegen een hele waslijst aan problemen naar voren gebracht. Het ERK is een heel loffelijk streven, waarvan het belang niet onderschat kan worden. In feite wordt met het ERK getracht een meetlat voor taalvaardigheid te maken. Met een dergelijke meetlat zouden we, door de objectieve beschrijvingen van de onderscheiden niveaus, een objectiever oordeel over een taalprestatie moeten kunnen geven. Met een dergelijke meetlat zouden we beter moeten kunnen communiceren over het niveau van taalvaardigheid, want door de objectieve beschrijvingen weten we allemaal waar we het over hebben. Als ik zeg dat ik Duits op B1 niveau beheers, dan weet iedereen meteen hoe goed ik in Duits ben, om maar iets te zeggen. Maar,, er is altijd een maar. Dat beoordeling met het ERK tot meer overeenstemming leidt tussen beoordelaars is zonneklaar; de omschrijving van het niveau geeft de beoordelaars meer houvast dan wanneer zij zonder een dergelijke beschrijving moeten oordelen over de taalvaardigheid van kandidaten. Echter, het is nog maar de vraag of verschillende groepen docenten op dezelfde manier gestuurd worden in hun oordeel over de taalprestaties.

In een aantal experimenten hebben Nederlandse docenten Engels, en Engelse docenten Engels, dan wel Nederlandse docenten Duits en Duitse docenten Duits dezelfde schrijfprodukten van leerlingen beoordeeld. Hierbij beoordeelden zowel de Nederlandse en Engelse docenten Engels teksten die derde- en vijfdeklassers in het Engels geschreven hadden. En evenzo

beoordeelde Nederlandse en Duitse docenten Duits Duitse teksten van derde- en vijfdeklassers. In een dergelijk experiment kunnen verschillende hypothesen getoetst worden. Ten eerste zouden de docenten het meer met elkaar eens moeten zijn als zij van een niveaubeschrijving, zoals het ERK, gebruik zouden kunnen maken. Dit bleek zonneklaar. Ten tweede, zouden de teksten van derdeklassers lager beoordeeld moeten worden dan die van vijfdeklassers. Dit bleek veel duidelijker bij oordelen zonder het ERK dan bij de beoordelingen met het ERK. Dus, zonder gebruik te maken van het ERK kan wel, en met het ERK kan veel minder goed een onderscheid gemaakt worden tussen de prestaties van derde- en vijfdeklassers in het Engels respectievelijk Duits.

Ten derde zouden de oordelen van de Nederlandse docenten Engels en Duits meer overeen moeten komen met die van hun collegae in het buitenland bij beoordeling met het ERK. Dit bleek lang niet altijd het geval. Over het algemeen bleken Duitse docenten Duits en Engelse docenten Engels coulanter in hun oordeel dan hun Nederlandse collegae. Dit bleek vooral te gelden bij het ERK. In de volgende figuur heb ik dit gevisualiseerd.

HIER FIGUUR 3

Dus, ondanks dat het ERK helpt, weten we nu zeker dat buitenlandse docenten op andere zaken letten bij de beoordeling van schrijfprodukten dan hun Nederlandse collegae, die naar het zich laat aanzien meer gefocust zijn op correctheid dan op begrijpelijkheid.

Een punt van geheel andere orde is de mate waarin een onderscheid gemaakt kan worden tussen leerders. Is het acceptabel als we na vier jaren onderwijs gemiddeld op A1-niveau zitten? Zijn we dan niet erg streng? Zijn we niet vergeten om verschillende relevante onderscheidingen aan te brengen? Een aardige anekdote. Ik heb onlangs de Dialangtoets gemaakt in het Nederlands. Ik kwam op niveau C1. Ik kan dus de meeste discussies in

het Nederlands volgen, tenzij ze erg abstract zijn. Dan moet ik zorgen dat ik de draad niet kwijt raak, en waar nodig hulp of uitleg vragen.

In dit verband is het aardig om een Amerikaanse pendant van het ERK, ACTFL te noemen. ACTFL en ERK zijn goeddeels vergelijkbare systemen, alleen is het hoogste niveau bij ACTFL lager, en wordt dus eerder bereikt, én zijn er meer lagere niveaus gedefinieerd. Hierdoor kan beter een onderscheid gemaakt worden in prestaties van tweedetaalleerders die pas begonnen zijn.

Kortom, het ERK is een stapje in de goede richting, maar er zijn nog zoveel problemen, dat ik vind dat we er nog niet massaal achteraan moeten lopen. En, zeker voor toetsing in de moedertaal is het ERK ontoereikend. Niet voor niets is de Raad van Europa hier een ander project voor gestart, wat helaas nog géén tastbare resultaten heeft opgeleverd. En, voor de vreemde talen wordt wel erg weinig onderscheid gemaakt tussen prestaties op lagere niveaus.

Praktische opdrachten en taalportfolio

Zoals al uit het voorafgaande blijkt ben ik niet zonder meer een voorstander van de zogenaamde praktische opdracht, voor summatieve toetsing. Laat ik eerst een voorbeeld geven van de soort waar ik moeite mee heb. Sommigen, zoals ik, hebben de spelvaardigheid van leerlingen geëvalueerd aan de hand van zelf geschreven teksten. Maar de vraag is wat meten we nu? Meten we werkelijk hoe goed leerlingen kunnen spellen? Of, meten we hoe goed leerlingen woorden die ze niet kunnen spellen kunnen omzeilen. Als we een uitspraak willen doen over de spellingsvaardigheid van leerlingen, dan is een dergelijke contaminatie natuurlijk een gruwel. Als we echter een uitspraak willen doen over spelfouten in zelfgeschreven teksten, wat wellicht een betere maat is, dan is het juist wenselijk om het zo te meten. Het hangt dus af van de doelstelling die we wensen te evalueren.

Ik ben in het algemeen geen voorstander van geïntegreerde opdrachten omdat ik dan niet meer weet wat ik meet, en wat ik zou moeten meten. Alles loopt op eens door elkaar heen: kennis, leesvaardigheid,

enzovoort. Een mooi voorbeeld is een onderzoek, waaraan ik mee heb mogen werken, naar de taalvaardigheid van oud-SDV-leerlingen, ten opzichte van andere leerlingen op het ROC. Voor deze leerlingen was een algemene opdrachtbeschrijvingen gemaakt. Bij één van de opdrachten, die aansluit bij de sector Sport, kregen de leerlingen de opdracht een jeu de boules-uitje te organiseren voor dertig bewoners van het verzorgingscentrum De Wartburg in Utrecht. Deze opdracht bestaat uit drie onderdelen. Allereerst dienen de studenten telefonisch een afspraak te maken met Sportpark Zuilenselaan in Utrecht. Daarbij moet met diverse zaken rekening gehouden worden. Zo zitten tien deelnemers in een rolstoel, waardoor zij de ballen niet zelf op kunnen rapen, bovendien beschikt het zorgcentrum zelf niet over jeu de boules-ballen. Een bijkomend probleem is dat het tijdstip van het uitje afgestemd moet worden op het bijgeleverde weerbericht én het weekprogramma van het bejaardencentrum, waarin ondermeer eet- en bezoektijden zijn vastgelegd. Als de afspraak telefonisch is vastgelegd moet per brief het benodigde vervoer geregeld worden. Bij het vaststellen van de route mochten de leerlingen gebruik maken van de routeplanner van de ANWB op het Internet. Tenslotte moest een begroting voor het uitje opgesteld worden. De totale opdracht bestaat dus vijf onderdelen: voorbereiding, telefoongesprek, brief, route en begroting. De toets als geheel is zeer betrouwbaar, maar wat meet deze? U heeft er geen idee van hoeveel moeite het gekost heeft om af te stappen van het idee dat de prestaties als geheel beoordeeld moesten worden.

Het gaat natuurlijk om de onderdelen. Pas als die ok zijn, als die betrouwbaar zijn weten we dat we aan de minimumvoorwaarde voor een valide toets voldoen. Maar wat nu te doen met een leerling die de opdracht nauwelijks leest en als een speer uit de startblokken vliegt, zonder zich ook maar één moment af te vragen waar hij heen moet gaan. Zo'n leerling doet het slecht op alle opdrachten. Een dergelijke afhankelijkheid is funest voor de meting; weliswaar wordt de betrouwbaarheid van de meting verhoogd, maar dat is slechts een doekje voor het bloeden, want de validiteit is dan zonder meer laag. In dit project hebben we er dan ook voor gekozen om de

opdracht niet alleen schriftelijk aan te bieden, maar ook mondeling; de leerlingen hoorden de opdracht, en zij konden op elk moment de opdracht nog eens terug lezen of terug horen. Bij de beoordeling van de prestaties is zoveel mogelijk getracht deze onafhankelijk te laten zijn van de andere opdrachten; aparte oordelen over elk onderdeel. Je krijgt dan een staatje als:

Leerling: Laurens			
Onderdeel	Maximale score	Score	Cijfer
Voorbereiding	2	1	6
Telefoongesprek	12	8	7
Route	8	5	7
Begroting	9	8	9
Brief	15	9	6
Gemiddeld			7

Niet één algemeen oordeel over de gehele prestatie maar oordelen/cijfers per onderdeel. Anders telt een onderdeel met veel items, en/of veel variantie extra zwaar voor het totaaloordeel. Als we dat niet doen dan tellen de voorbereiding, route en begroting veel minder zwaar mee. Kortom, geef oordelen per onderdeel.

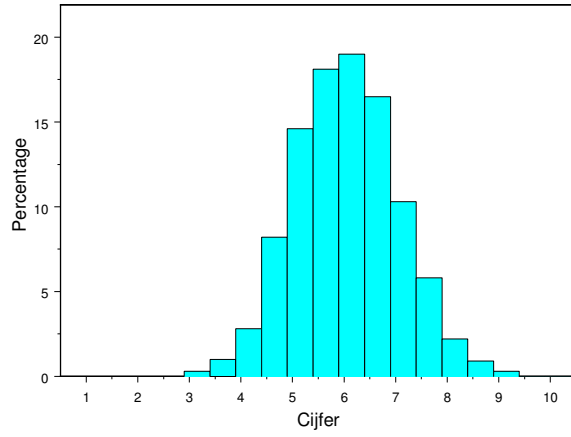
Het portfolio wordt tegenwoordig steeds vaker ingezet als beoordelingsinstrument, in plaats van als rapportagemiddel. Dat geldt niet alleen op het MBO, maar ook op de universiteit. In verschillende sectoren van de universiteit worden portfolio's als beoordelingsinstrument ingezet in het kader van een summatieve toetsing. Hierbij krijg ik soms het idee dat we terug zijn in de jaren '70. Als studenten zelf maar het idee hebben dat ze genoeg geleerd hebben én daar een voorbeeld van kunnen geven, dan is het wel goed; dan is een verdere evaluatie overbodig. U begrijpt dat ik dit niet zo handig vind. Als we iets weten van zelfbeoordeling is dat deze negatief correleert met 'echte' cijfers; de zwakkere broeders en zusters vinden zichzelf beter dan de goede. En, dat kan natuurlijk nooit een gezonde basis voor beoordeling zijn. Terug naar het portfolio. Wat ik essentieel vind is dat

alle leerlingen langs *dezelfde* meetlat gelegd worden, hoe lastig het ook is om zo'n meetlat te construeren, of hoe problematisch zo'n meetlat ook mag zijn. Als we dat principe loslaten, dan hebben we feitelijk verschillende meetlatten voor verschillende personen; de ene leerling wordt geëvalueerd met eenvoudige opgaven, en de andere op basis van veel moeilijkere zaken. Als we het principe van één meetlat voor allen loslaten is een eerlijke vergelijking onmogelijk. Dit wil natuurlijk niet zeggen dat portfolio's niet gebruikt kunnen of moeten worden. Maar wel dat portfolio's voor summatieve toetsing, in mijn ogen, niet geschikt zijn!

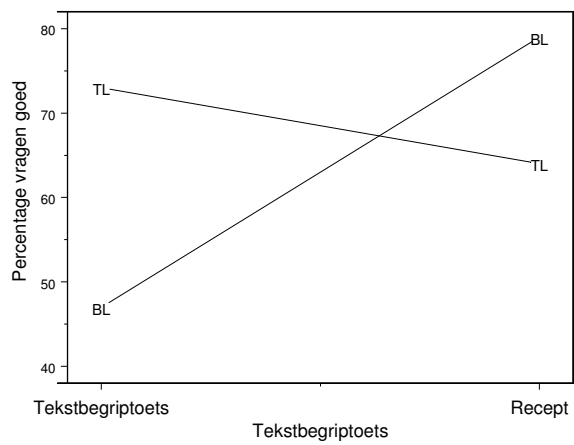
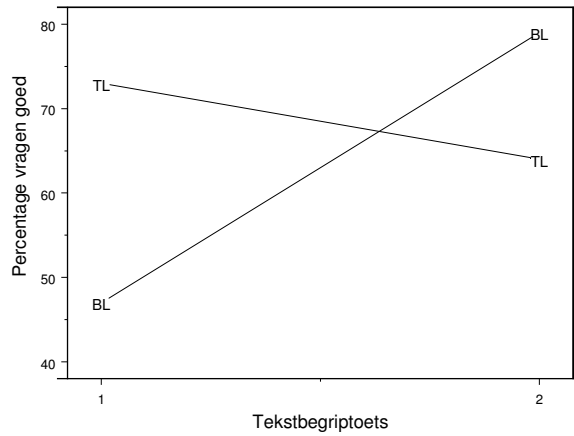
Als we toch portfolio's willen gebruiken, dan moeten zij met dezelfde informatie gevuld moeten worden; als de ene leerling er een briefje in kan doen, en de andere een betoog, is feitelijk niet aan deze aanname voldaan. Bovendien moeten de leerlingen de opdrachten onder dezelfde omstandigheden gemaakt hebben. Maar, ja, dat is eigenlijk allemaal tegen het wezen van portfolio's. Is één van beide niet het geval, dus of verschillende bewijzen van bekwaamheid, of andere omstandigheden, dan is een eerlijke vergelijking eigenlijk niet mogelijk. Dat betekent dat het portfolio, zoals dat vaak gebruikt wordt, eigenlijk niet gebruikt zou mogen worden als meetlat, om een summatieve beslissing op te baseren.

Samenvattend: probeer als eerste zo goed mogelijk te definiëren wat geëvalueerd moet worden. Dit kan nauw aansluiten bij de doelen van het onderwijs en/of bij hetgeen feitelijk onderwezen is. Maak op grond daarvan instrumenten. Probeer deze instrumenten, als het ook maar half mogelijk is, uit, voordat ze echt gebruikt gaan worden. Maak per dimensie (lezen, schrijven, luisteren, spreken, of een combinatie hiervan) één (of meer) instrumenten, maar laat deze zo min mogelijk beïnvloed worden door allerlei andere zaken, ook al zijn die belangrijk. Maak daarvoor als het nodig is andere instrumenten. De instrumenten kunnen best gecombineerd worden tot een thematisch geheel, maar elk instrument moet als het ware onafhankelijk kunnen functioneren. En, tot slot geef oordelen per dimensie, in plaats van geglobaliseerde oordelen over het geheel.

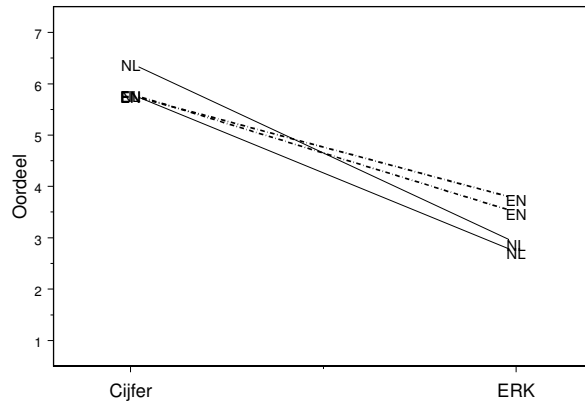
Figuur 1. Variatie in cijfers ten gevolge van de factor tekst.



Figuur 2. Het gemiddelde percentage goed beantwoorde vragen (y-as) op twee type tekstbegriptoetsen van leerlingen in de Theoretische Leerweg (TL) en leerlingen in de Basisberoepsgerichte-leerweg (BL).



Figuur 3. Gemiddelden voor twee beoordelingsmethodes van twee teksten door Nederlandse docenten Engels (NL) en Engelse docenten Engels (EN).



Authentieke toetsen