Meten van leerprestaties in het (v)mbo:

assessment for learning
en
assessment of learning

Meten van leerprestaties in het (v)mbo:

assessment for learning
en
assessment of learning

© 2010 Uriël Schuurs & Ludo Verhoeven

Inhoudsopgave

Voorwoord

Voor u ligt een verslag van onderzoek op het gebied van meten van leerprestaties in het (v)mbo. Sinds het competentiegericht leren een plek heeft verworven in het (v)mbo, worden er andere eisen gesteld aan de manier waarop leeropbrengsten in beeld worden gebracht. Bij de nieuwe vormen van toetsing (competentietoetsing, performance testing, real-life taakstellingen, hands-on toetsprocedures, self-assessments) lijkt de inhoudsvaliditeit beter gewaarborgd en lijkt er ook een positief te waarderen backwash-effect op het leren aanwijsbaar (*assessment for learning* functie).
Naast deze voordelen zijn er ook nadelen. Allereerst lijken de objectiviteit en betrouwbaarheid van de toetsing minder gewaarborgd waardoor de validiteit van de *assessment of learning* functie in gevaar komt. Daarnaast blijken de metingen van opbrengsten van het onderwijs moeilijk vergelijkbaar en is vooral de predictieve validiteit van metingen met het oog op toekomstig succes op de arbeidsmarkt volstrekt onduidelijk.

In opdracht van NWO PROO d.d. 24 april 2009 is in voorliggende studie nagegaan wat er bekend is over nieuwe vormen van toetsing. Aan de hand van een computergestuurde zoektocht door literatuurbestanden en consultatie van leerkrachten en van een aantal experts op het terrein van toetsen, is antwoord gezocht op de volgende drie vragen:

1. Welke toetsvormen zijn geschikt om zowel de *assessment of learning* functie te vervullen als de *assessment for learning* functie?
2. Op welke manieren kunnen de opbrengsten van taalonderwijs in het (v)mbo op een gestandaardiseerde manier worden gemeten?
3. Wat is er bekend over predictieve validiteit van moderne toetsvormen?

Omdat de roep om nieuwe assessmentvormen is ontstaan tijdens, en ten gevolge van onderwijsveranderingen in het vmbo, schetsen we in hoofdstuk 1 eerst een beeld van het veranderend onderwijs. In hoofdstuk 2 gaan we in op doelen en functies van assessment; met name schetsen we de bruikbaarheid van assessment-instrumenten voor de *assessment OF learning*-functie tegen over de *assessment FOR learning*-functie. In hoofdstuk 3 gaan we nader in op de eisen die gesteld kunnen worden aan assessment-instrumenten; aan welke eisen meer belang moet worden gehecht, hangt af van de vraag of *assessment for learning* of *assessment for learning* wordt beoogd. In hoofdstuk 4 doen we verslag van een veldbevraging en van bevraging van experts op het gebied van assessment. In hoofdstuk 5 blikken we terug en geven we antwoord op de drie gestelde onderzoeksvragen. Bovendien doen we aanbevelingen voor de onderwijspraktijk en suggesties voor verder onderzoek.

Het onderzoek werd toegekend binnen het NWO Gebiedsbestuur Maatschappij- en gedragswetenschappen. Naast ondergetekenden hebben docenten en experts op het terrein van toetsing (zie bijlage 3 voor een overzicht) onmisbare inbreng gehad. Wij danken alle betrokkenen hiervoor van harte. Onze speciale dank gaat uit naar Marije Boer, die als project-assistent met het bijeenbrengen van de bestudeerde literatuur een niet te onderschatten functie heeft vervuld. Wij hopen dat het rapport zowel de praktische als de theoretische implicaties voldoende inzichtelijk maakt.

Nijmegen, februari 2010

dr. Uriël Schuurs
prof. dr. Ludo Verhoeven

Hoofdstuk 1 Veranderend onderwijs: vmbo en mbo

In dit hoofdstuk beschrijven het vmbo en het mbo en met name de veranderingen die in het onderwijs hebben plaatsgevonden. Deze veranderingen zijn van belang omdat in de dynamiek daarvan de vraag naar instrumenten voor *assessment for learning* is ontstaan. We leggen daarbij de nadruk op de consequenties die deze veranderingen hebben voor de taaltoetsing.

1.1.1 Beschrijving van het onderwijs

1.1.2 Het vmbo

Ongeveer 60% van de leerlingen in het voortgezet onderwijs zit in het vmbo (voorbereidend middelbaar beroepsonderwijs). De eerste twee jaar van het vmbo vormen de basisvorming: er wordt een uitgebreid vakkenpakket aangeboden dat voor alle leerlingen hetzelfde is. Aan het eind van het tweede jaar kiest de leerling één van de vier sectoren :

- Techniek
- Zorg en welzijn
- Economie
- Landbouw

Voor alle vmbo-leerlingen geldt een gemeenschappelijk deel van het curriculum; dit betreft de schoolvakken Nederlands, Engels, maatschappijleer, lichamelijke opvoeding en ten minste één van de creatieve vakken beeldende vorming, muziek, dans of drama. Voor elke sector geldt vanaf 1 augustus 2010 één verplicht sectorvak, bv. het vak economie voor de sector economie, of het vak wiskunde voor de sector techniek.

Het vmbo kent naast de vier sectoren ook vier leerwegen. Deze leerwegen verschillen van elkaar in de mate waarin de beroepspraktijk aandacht krijgt in het onderwijs. Van meest praktisch naar meest theoretisch gerangschikt zijn de volgende leerwegen te onderscheiden:

- Basisberoepsgerichte leerweg (BB) De basisberoepsgerichte leerweg is bestemd voor leerlingen die vooral praktisch ingesteld zijn. De leerlingen doen examen in vier algemene vakken en een beroepsgericht vak. Het examenprogramma voor deze leerweg is minder uitgebreid en meer praktisch dan dat van de andere leerwegen.
- Kaderberoepsgerichte leerweg (KB) De kaderberoepsgerichte leerweg is bedoeld voor leerlingen die theoretische kennis het liefst opdoen door praktisch bezig te zijn. De benaming verwijst naar het feit dat de leerlingen al bezig zijn met een opleiding die in zijn geheel gericht is op een functie op kaderniveau (niveau 3 of 4 in het mbo). De leerling doet examen in vier algemene vakken en een beroepsgericht vak of programma met een omvang van 960 uur.
- Gemengde leerweg (GL) De gemengde leerweg biedt een combinatie van theoretisch (algemeen) en praktisch (beroepsgericht) onderwijs en is bedoeld voor leerlingen die de theoretische vakken goed aankunnen, maar zich al gericht willen voorbereiden op bepaalde beroepen. De gemengde leerweg is qua niveau gelijk aan de theoretische leerweg. Het programma en het examen van de algemene vakken zijn precies gelijk

aan dat van de theoretische leerweg. Naast deze vijf algemene vakken kiezen leerlingen een beroepsgericht programma van 320 uur. Dit bestaat uit een beroepsgericht vak binnen de sector die de leerling heeft gekozen, zoals het vak elektrotechniek binnen de sector techniek, of het vak verzorging binnen de sector zorg en welzijn.

- Theoretische leerweg (TL) De theoretische leerweg is niet gericht op een bepaalde beroepskeuze, vandaar de benaming 'theoretische leerweg'. Samen met de gemengde leerweg heeft de theoretische leerweg het hoogste niveau waar het de cognitieve vakken betreft. De leerlingen doen examen in zes algemene vakken waarvan Nederlands en Engels verplicht zijn; daarnaast kiezen de leerlingen vier examenvakken uit de moderne vreemde talen inclusief Turks en Arabisch; daarnaast zijn natuurkunde, wiskunde, geschiedenis, aardrijkskunde enzovoorts mogelijke keuzevakken.

Het vmbo is niet bedoeld als eindopleiding, het leidt op tot het mbo. De leerroute van vmbo via mbo naar hbo wordt sinds een decennium wel aangeduid als de beroepskolom. Het ministerie van Onderwijs, Cultuur & Wetenschap en het beroepsonderwijs zelf werken sinds enkele jaren aan een beter functionerende beroepskolom.

Voor alle sectoren en alle richtingen is het schoolvak Nederlands voor alle leerlingen verplicht. Vaak zitten leerlingen van de theoretische leerweg en de gemengde leerweg samen in één klas en er zijn aparte klassen voor leerlingen uit de kaderberoepsgerichte leerweg en voor leerlingen uit de basisberoepsgerichte leerweg. Zo bezien zijn er voor Nederlands dus eigenlijk *drie* richtingen.

In het vmbo is het vak Nederlands een examenvak. De helft van het examencijfer wordt bepaald door het schoolexamen, de andere helft door een landelijk examen. De inhoud van het schoolexamen wordt jaarlijks vóór 1 oktober vastgelegd in een Programma van Toetsing en Afsluiting (PTA) . Het schoolexamencijfer wordt doorgaans bepaald op basis van enkele schriftelijke toetsen (schoolexamentoetsen), PO's (praktische opdrachten; meestal werkstukken) en handelingsdelen (verplichte opdrachten). Voor handelingsdelen wordt in het algemeen geen cijfer gegeven, maar een beoordeling in O(nvoldoende), V(oldoende) of G(oed). Een voorbeeld van een handelingsdeel bij het vak Nederlands is het schrijfdossier dat bijvoorbeeld een schrijfplan moet bevatten (een aanduiding van onderwerp, hoofdgedachte, schrijfdoel, beoogd publiek en per alinea een samenvattende zin) en de uitwerking van dat schrijfplan. Deze gang van zaken wijkt nauwelijks af van wat al decennia lang gebruikelijk is in het voortgezet onderwijs, zij het dat de leerkracht dankzij de handelingsdelen wat meer zicht krijgt op het proces dat de leerling doorloopt om bij een product te komen.

Het voert te ver om in te gaan op de kwaliteit van de schoolexamens in het vmbo, zie daarvoor bv. *Het schoolexamen in het voortgezet onderwijs / Verslag van een onderzoek naar de kwaliteit van het schoolexamen bij de vakken Engels, Nederlands, biologie en wiskunde* (Cito, december 2008). Dit rapport doet verslag van onderzoek naar de vakinhoudelijke en toetstechnische kwaliteit van de schoolexamens (SE) in het voortgezet onderwijs. Het onderzoek is uitgevoerd bij de vakken Engels en biologie in het havo en Nederlands en wiskunde in het vmbo. Na een training in het beoordelen hebben drie

beoordelaars per vak de kwaliteit van het SE op negentien indicatoren beoordeeld. Voor Nederlands is de conclusie van dit rapport dat de kwaliteit van de schoolexamens grosso modo voldoet aan de te stellen eisen. Al werden er wel kwaliteitsverschillen geconstateerd, de beoordelaars hebben de kwaliteit van het SE Nederlands volgens de gehanteerde criteria van geen enkele school als onvoldoende aangemerkt.

Er is erg weinig onderzoek gedaan naar de feitelijke gang van zaken in het schoolvak Nederlands in het vmbo. Sporadisch vindt er onderzoek plaats op deelgebieden zoals:

1) de doorstroming van leerlingen van vmbo naar mbo
2) de kwaliteit van schoolboeken in het vmbo
3) de mate van integratie van het schoolvak Nederlands in de andere vakken.

Ad 1 Ogenschijnlijk is de doorstroming van vmbo naar mbo niet dramatisch slecht. Zo schatten Neuvel en Van Esch (2006) schatten het feitelijke aantal vmbo'ers dat niet (direct) doorstroomt naar het mbo tussen de 3 en 9 procent, waarbij het overgrote deel afkomstig is van de basisberoepsgerichte leerweg. Daar staat tegenover dat 6 procent van de vmbo-leerlingen uit het derde en vierde leerjaar voortijdig de school verlaat en in 2005 verliet 34 procent van de deelnemers het mbo op niveau 1 zonder diploma. De Bruijn (2007) signaleert dat er in de doorstroming van vmbo naar mbo de laatste jaren met wisselend succes de zogenaamde harde blokkades zoals intree-eisen, beroepsdomein, timing en fasering zijn aangepakt, maar dat vmbo en mbo niet op elkaar aansluiten als gevolg van de zogenaamd zachte blokkades: onderwijscultuur en pedagogische benadering, de gehanteerde didactiek en de benadering van inhouden (abstractieniveau).

Ad 2 Schoolboekteksten bedoeld voor het vmbo zijn van slechte kwaliteit en daardoor lastig te begrijpen voor de leerlingen. Met name de verbanden tussen de zinnen wordt ten onrechte vaak niet expliciet weergegeven door de auteurs (Land et al. 2006). Ook ondersteuning van de docent Nederlands leidt niet tot een beter begrip en beter onthouden van theoretische teksten door vmbo-leerlingen binnen sector Zorg en Welzijn (Toorenaar & Rijlaarsdam, 2007). Deze en soortgelijke studies onderstrepen de noodzaak om school-boekteksten zorgvuldiger op de doelgroep af te stemmen.

Ad 3 Onder deze noemer verschijnen met regelmaat studies naar de stand van zaken met het taalbeleid in het vmbo. Met taalbeleid wordt doorgaans de integrale benadering van het taalonderwijs op schoolniveau bedoeld. Herder & Berenst (2008) laten zien dat betrokkenen onder de term taalbeleid verschillende dingen verstaan:

- Taalbeleid als een (didactisch) concept van de hele school waarin vanuit taal het gehele onderwijs wordt ingevuld, van belang voor alle leerlingen en onder de verantwoordelijkheid van het gehele team. Bijvoorbeeld kan de school vastleggen op welk moment leerlingen een bepaald niveau moeten hebben bereikt en hoe dit wordt getoetst, of welke maatregelen er worden genomen bij leerlingen die een bepaald niveau niet tijdig gehaald hebben.
- Taalbeleid als een verzamelnaam voor concrete activiteiten en materialen op het gebied van taalondersteuning ten behoeve van specifieke (groepen) leerlingen onder verantwoordelijkheid van de directie en de taalcoördinator.
- Taalbeleid als een vorm van zorg (screening en remediëring) van de school bestemd voor een specifieke en kleine groep leerlingen die taalproblemen hebben, uit te

voeren door speciaal aangestelden, en zonder betrokkenheid van de directie en de overige docenten.

In het vmbo is, als onderdeel van taalbeleid als concept voor de hele school, vaak de vraag aan de orde of het schoolvak Nederlands niet geheel of gedeeltelijk moet worden geïntegreerd in de andere schoolvakken. Daarmee zou de inrichting van het vmbo-onderwijs meer gaan lijken op de praktijk in het mbo (zie verder paragraaf 1.2). De studies van Bonset et al. (2006) en Bonset & Ebbers (2007) hebben betrekking op de mate van integratie van het vak Nederlands in de andere schoolvakken. Achterliggende assumptie is, dat Nederlands centraal voor alle leerlingen verplicht vak is dat een dienende functie kan vervullen bij de andere vakken waar het algemeen inzetbare competenties betreft, zoals het begrijpen van school- en instructietaal, studievaardigheden, luistervaardigheid, gespreksvaardigheid en informatieverwerking. De resultaten laten zien dat Nederlands in de bovenbouw van het vmbo nog steeds overwegend gegeven wordt als zelfstandig vak met aparte uren op het rooster. De inhoud van de lessen Nederlands wordt volgens de bevraagde leerkrachten wel afgestemd op wat de leerling bij andere vakken nodig heeft voor de onderdelen schrijfvaardigheid (71% van de docenten), spreek- en gespreksvaardigheid (61%) en leesvaardigheid (59%). De meerderheid van de docenten (56%) ziet Nederlands liefst deels als apart vak en deels geïntegreerd met andere vakken; 39% houdt Nederlands liefst als apart vak op het rooster. Slechts een zeer kleine minderheid is voor een volledige integratie van het vak Nederlands met beroepsgerichte vakken of zaakvakken.

De geringe aanhang voor een volledige integratie van het vak Nederlands met de zaakvakken is opmerkelijk als we vmbo en mbo in één adem bespreken. Op het mbo bestaat het schoolvak Nederlands niet meer: dit vak is op vrijwel alle mbo-opleidingen – voorzover er al wat van over is - volledig geïntegreerd in de beroepsvakken.

De Onderwijsinspectie heeft in 2007 onderzoek gedaan naar de taalprestaties van vmbo-leerlingen. Het onderzoek concentreerde zich op klas 1 en 2 van afdelingen van vmbo-scholen voor basisberoepsgerichte en kaderberoepsgerichte leerwegen en scholen voor praktijkonderwijs . Met behulp van de toets Diataal zijn de prestaties van de leerlingen gemeten op tekstbegrip, woordenschat en luistervaardigheid. Metingen vonden plaats in twee cohorten: de toets Diataal is bij de leerlingen van het tweede leerjaar afgenomen aan het einde van het leerjaar, in de maand juni 2007 en bij de leerlingen van het eerste leerjaar aan het begin van het leerjaar, in de maand september 2007. De belangrijkste bevindingen staan samengevat in tabel 1.1 (vgl. Inspectie van het Onderwijs 2007):

Tabel 1.1: Percentage vmbo-leerlingen met achterstanden op drie deeltaalvaardigheden

	Luistervaardigheid	Tekstbegrip	Woordenschat
Begin leerjaar 1	83%	60%	51%
Eind leerjaar 2	66%	72%	25%

De Inspectie noemt voor deze resultaten zes mogelijke verklaringen die we hier integraal overnemen omdat ze bijdragen aan het beeld van het onderwijs Nederlands op het vmbo:

1 Op de helft van de scholen met vmbo-bbl/kbl beschikken de leraren Nederlands bij de start van het eerste leerjaar niet over een overzicht van de eindniveaus voor taal die elke leerling afzonderlijk in het basisonderwijs heeft bereikt.
2 De onderwijstijd voor het vak Nederlands wordt in omvang doorgaans niet afgestemd op groepen leerlingen met meer of minder taalachterstanden.
3 Op vier van de tien onderzochte scholen geven de leraren Nederlands geen huiswerk op om bepaalde taalvaardigheden verder te oefenen.
4 Slechts 30 procent van de scholen heeft vastgelegd welk niveau de leerlingen aan het eind van het tweede leerjaar zouden moeten bezitten. Driekwart van de onderzochte scholen toetst niet of de leerlingen op het eind van het tweede leerjaar het gewenste niveau heeft bereikt.
5 Op de meeste scholen besteden de leraren van andere vakken dan Nederlands geen (structurele) aandacht aan taalachterstanden.
6 Op driekwart van de scholen is geen structurele en herkenbare aanpak van de taalachterstanden vastgelegd.

1.1.3 Het mbo

Meer dan eens is geschetst hoe het beroepsonderwijs, en dan met name het mbo, in Nederland ingrijpend is veranderd (Kuhlemeijer 2005; Nijhof 2006; Neuvel & Van Esch 2006; De Bruijn 2007). Kuhlemeijer (2005) schetst drie krachtige tendenzen die het mbo ingrijpend veranderen:

- Aanbodgericht onderwijs wordt vervangen door vraaggestuurd leren: cursisten bepalen binnen randvoorwaarden zelf wat zij leren, hoe en wanneer dat gebeurt en waar zij dat doen.
- Vakmatig beroepsonderwijs wordt getransformeerd in competentiegericht leren: daarbij worden zowel algemene schoolvakken en beroepsgerichte vakken als theorie en praktijk geïntegreerd tot een vorm van onderwijs waarbij de cursist wordt geplaatst in min of meer stereotype situaties uit de beroepspraktijk (vgl. Van der Sanden et al. 2003a).
- Tot slot komen niet alleen beroepsgebonden vaardigheden aan de orde in het onderwijs, maar ligt er steeds meer nadruk op de zogenaamde kerncompetenties: dit zijn algemene competenties zoals geletterdheid, gecijferdheid, communiceren, ICT, samenwerken, problemen oplossen en reflecteren en evalueren. Van dit soort competenties mag worden verondersteld dat iedereen deze in bepaalde mate moet beheersen om zich in maatschappelijk en beroepsmatig opzicht te kunnen redden (vgl. Inspectie van het Onderwijs 2004).

Ogenschijnlijk lijkt het mbo op het vmbo: net als bij het vmbo worden er mbo-opleidingen gegeven op vier verschillende niveaus:

- niveau 1: assistent beroepsbeoefenaar (hier is geen startkwalificatie nodig)
- niveau 2: medewerker / basisberoepsbeoefenaar (startkwalificatie is BB van vmbo)
- niveau 3: zelfstandig medewerker / zelfstandig beroepsbeoefenaar / vakopleiding
- niveau 4: middenkaderfunctionaris / gespecialiseerd beroepsbeoefenaar (geeft toegang tot hbo)

Voor de hoogste twee niveaus geldt een vmbo-diploma KB, GL of TL als startkwalificatie. Vanaf 1996 vormen de mbo-opleidingen onderdeel van de ROC's. De mbo's hebben inhoudelijk vier sectoren: landbouw, techniek, economie en zorg en welzijn. In totaal leiden de mbo's op tot ruim 700 soorten beroepen. Per beroep en per niveau is vastgelegd welke competenties de deelnemer moet bezitten.

Structureel bezien lijken vmbo en mbo op elkaar, maar het onderscheiden van competenties per beroep en per niveau maakt al duidelijk dat het mbo erg afwijkt van het vmbo. Het mbo kent bovendien veel verschijningsvormen: het wordt aangeboden in tal van leeromgevingen en er is niet één bepaalde dominante beroepsdidactiek aanwijsbaar. In de woorden van Nijhoff (2006): "Sinds het begin van de jaren negentig is er sprake van bijvoorbeeld modulair onderwijs al dan niet in klasseverband, projectonderwijs, probleemgestuurd leren, mastery learning, klassikaal onderwijs, Leittextsystemen, simulatiesystemen, het individuele leergroepen systeem, individueel gestuurde loopbaansystemen (gebaseerd op de commissie Boekhoud), en de afwisseling tussen internships (stages) en apprenticeship (leerlingwezen) in allerlei varianten. Wat de effectiviteit van al deze varianten is, is onbekend."

Een derde opmerkelijk verschil tussen vmbo en mbo is, dat het schoolvak Nederlands op het mbo is verdwenen. Verondersteld wordt dat leerlingen de benodigde taalvaardigheid al bezitten uit eerdere opleidingstrajecten en dat ze voor het beroep noodzakelijke taalbeheersing kunnen aanleren tijdens beroepsgebonden activiteiten en projecten. In het mbo is Nederlands geen zelfstandig schoolvak meer. Uit onderzoek van Visscher (2008) is gebleken dat er scholen zijn met docenten Nederlands en een talencentrum, terwijl andere scholen geen voorzieningen op dit terrein hebben. Uit hetzelfde onderzoek bleek, dat er mbo-scholen zijn waar geen taalbeleid is, of waar nog maar net een begin gemaakt wordt met taalbeleid.

Dat betekent niet dat er geen onderwijs Nederlands plaatsvindt: in 2004 werd gemiddeld in de eerste drie leerjaren van het mbo 2,5 lesuur besteed aan activiteiten die verbetering van de Nederlandse taalvaardigheid tot doel hebben (vgl. Neuvel et al. 2004). Deze activiteiten vinden vooral plaats in de lagere leerjaren en in de lagere mbo-niveaus. Een groot deel van deze activiteiten voor Nederlands vindt plaats in projectvorm waarbij de deelnemers vaardigheden zoals presenteren, rapporteren en vergaderen oefenen (vgl. Onderwijsinspectie 2006). De talige eisen die daarbij worden gesteld, zijn volledig geïntegreerd met vakinhoudelijke eisen. Per beroep zijn deze eisen vastgelegd in zogenaamde kwalificatiedossiers. Een kwalificatiedossier beschrijft voor een beroep of beroepsgroep:

- de inhoud van het beroep;
- de benodigde competenties voor een beginnende beroepsbeoefenaar;
- de benodigde kennis en vaardigheden voor een beginnende beroepsbeoefenaar.

Vanaf augustus 2010 moeten alle mbo-opleidingen al hun opleidingen in lijn met de kwalificatiedossiers aanbieden. Tot die tijd is er nog sprake van een experimenteerfase.

Deze kwalificatiedossiers beschrijven minutieus welke taken de beroepsbeoefenaar moet kunnen uitvoeren en welke combinaties van deelvaardigheden daarom beheerst moeten worden. Het kwalificatiedossier bezit een taalcompetentieprofiel: een transparante

beschrijving van niveaus van taalvaardigheid waarin verwezen wordt naar de Europese standaard, het Common European Framework of Reference (Council of Europe 2001; zie ook http://www.erk.nl/) . Bovendien is het profiel competentiegericht en gedifferentieerd naar vijf taalvaardigheden: Spreken, Luisteren, Gesprekken voeren, Lezen en Schrijven.

Bijvoorbeeld wordt van de "medewerker sociale zekerheid" gesteld dat die moet beschikken over "*specifieke vaardigheden in de omgang met klanten, zoals verschillende gesprekstechnieken en vaardigheden op het gebied van conflicthantering. De medewerker sociale zekerheid beheert een eigen klantenbestand en heeft te maken met procedures in de verschillende stadia van afhandeling. Hij moet in staat zijn prioriteiten te stellen en processen te bewaken. De medewerker sociale zekerheid bij een afdeling sociale zaken van een gemeentelijke organisatie heeft te maken met klanten met veelal complexe problemen van psychische, sociaal medische en/of financiële aard. Dit stelt hoge eisen aan zijn incasseringsvermogen en communicatieve vaardigheden*." (uit *Landelijke Kwalificaties mbo / Medewerker sociale zekerheid*. Bron: http://pdf.kwalificatiesmbo.nl/smartsite.shtml?id=295671)

Op basis van bovenstaande beschrijving wordt van kandidaten "medewerker sociale zekerheid" op mbo-niveau 4 het volgende verlangd:

	Luisteren	Lezen	Gesprekken voeren	Spreken	Schrijven
C1		x	x		
B2	x	x	x	x	x
B1	x	x	x	x	x
A2	x	x	x	x	x
A1	x	x	x	x	x

Ter contrast: in de Landelijke Kwalificaties mbo / Metselaar worden voor de "Allround Metselaar inclusief nieuwe metseltechnieken" de volgende talige eisen gesteld:

	Luisteren	Lezen	Gesprekken voeren	Spreken	Schrijven
C1					
B2					
B1	X	X	X		X
A2	X	X	X		X
A1	X	X	X		X

De toetsing van de vereiste taalvaardigheden vindt geïntegreerd plaats tijdens de toetsing van de beroepsgebonden taken.

Samenvattend kunnen we stellen dat het vmbo grotendeels traditioneel gebleven is, terwijl het mbo zich daarvan onderscheidt op drie manieren: het hedendaagse mbo

1 is competentiegericht
2 is zeer heterogeen in didactiek
3 heeft de aandacht voor taalonderwijs geïntegreerd in andere vakken

Deze veranderingen beperken zich goeddeels tot het mbo. In de literatuur worden voor het vmbo wel onderwijsvernieuwingen beschreven, maar deze vernieuwingen hebben in de praktijk goeddeels betrekking op het mbo. Afgezien van experimenten met buitenschoolse opdrachten is het vmbo goeddeels hetzelfde gebleven en binnen het schoolvak Nederlands doen zich nauwelijks veranderingen voor (bron: pers. communicatie Piet Litjens). Deze bevindingen stroken met de manier waarop vmbo en mbo worden afgesloten.

Het wekt geen verwondering dat binnen de geschetste verandering van het mbo een roep is ontstaan naar andersoortige vormen van assessment. Daarop gaan we in paragraaf 1.3 nader in, eerst bespreken we het krachtenveld waarbinnen de veranderingen hebben plaatsgehad.

1.2 Verschillende opvattingen over onderwijs

Aan de geschetste veranderingen in het onderwijs lijkt een fundamentele verschuiving in opvatting over leren ten grondslag te liggen die zich van de internationale grenzen weinig aangetrokken heeft. In de literatuur wordt met name verwezen naar een paradigmatische verschuiving vanaf behavioristische theorieën naar het sociaal-constructivisme (vgl. bv. Biggs et al. 1996; Biggs 1996; Mabry 1999, Bachman 2000; McMillan 2001).

Onder invloed van het sociaal-constructivisme is de nadruk meer komen te liggen op de individuele leerling en is er meer aandacht voor de actieve en zelfstandige rol van de individuele leerling binnen het onderwijsproces. Het sociaal-constructivistische perspectief vertrekt vanuit de idee dat kennis een mentale constructie is. Ieder mens creëert, in sociale interactie, zijn eigen waarheid. Kennis wordt daarmee flexibel en veranderlijk in tijd. Leren vraagt activiteiten van een lerende: het opbouwen van eigen mentale modellen, interactie aangaan met anderen en ervaring opdoen. Leren hangt daardoor nauw samen met de context waarin een lerende zich bevindt. Leren leidt dus niet tot een eenduidige voorspelbare uitkomst en is niet direct van buitenaf te sturen.

Tabel 1.2: Kenmerken van positivistisch en sociaal-constructivistisch denken (naar Koopmans 2006, 33)

Positivistisch perspectief	Sociaalconstructivistisch perspectief
Kennis is hard, overdraagbaar, feitelijk, wetmatig en onafhankelijk van mensen. Er is één waarheid.	Kennis is een persoonlijk mentaal construct dat ontstaat in interactie en door ervaring en is dus flexibel. Er bestaan meerdere waarheden.
Leren is voorspelbaar. Op basis van een goed curriculum kan ieder individu een gewenst leerresultaat behalen.	Leren is een actief proces van interactie en ervaren. Het resultaat is onvoorspelbaar, individueel en hangt nauw samen met de context.
Leren kan van buitenaf door deskundigen worden gestuurd.	Leren is een actief proces van de lerende zelf.

Met de opkomst van de constructivistische optiek zijn er andere onderwijskundige veranderingen te signaleren:

- Veranderingen in rollen in de klas: leerlingen zijn actief, nemen zelf initiatief en bepalen hun eigen leerroute.
- Het curriculum moedigt onderzoek door leerlingen aan en omvat een combinatie van kennis, vaardigheden en samenwerkend leren.
- Schoolorganisatie is veranderd van klassieke klaslokalen voor directe instructie naar multifunctionele ruimtes waar leerlingen samenwerkend en exploratief kunnen leren.
- Leren wordt steeds minder tijd- of plaatsgebonden. De periode dat men leert is niet beperkt tot de periode dat men een fulltime opleiding volgt, men leert het hele leven. De *éducation permanente* is niet alleen aanwijsbaar in de vele scholings-trajecten in bedrijven, maar blijkt ook uit de intensievere samenwerking tussen opleidingen en bedrijven.

In dit veranderend opleidingslandschap signaleert Bachman (2000) binnen de grenzen van de onderwijstoetsing een belangrijke paradigmatische verandering: was er voorheen Assessment OF learning, vanaf 1980 is er een groeiende aandacht voor de assessment FOR learning. Dit betekent dat toetsing niet langer achteraf plaatsvindt met het doel om de onderwijsrendementen op individueel en groepsniveau in beeld te brengen, maar dat er vaak ook tussentijds getoetst wordt om individueel richting te geven aan de inhoud van het verdere onderwijs (zie verder hoofdstuk 2.1). Deze verandering is volgens Bachman (2000) terugvoerbaar op drie bewegingen binnen de taal- en testwetenschap:

1 Veranderend inzicht in taalontwikkeling, met name door de *communicative approach* in de jaren 80 van de vorige eeuw. Daarbij lag de focus niet langer op de beheersing van het formele taalsysteem, maar op het effectief overbrengen van intenties (vgl. Canale & Swain 1981; Alderson 1981; Spolsky 1985).
2 Veranderend methodologisch inzicht: assessment werd criterionreferenced (dus: aan de taak gerelateerd) in plaats van normreferenced (aan het gemiddelde van een groep gerelateerd; vgl. ook Orr 2008); dankzij de ontwikkeling van nieuwe statistische benaderingen zoals de IRT (vgl. bv. De Jong 1991; Verhelst et al 1995) werd *tailormade assessment* mogelijk alsmede het op één meetschaal plaatsen van toetsen (vgl. bv. Schuurs 2001).
3 Veranderende technische mogelijkheden: met name de technische mogelijkheid om met behulp van de computer, dus geautomatiseerd adaptief te toetsen zodat het niveau van assessment optimaal in overeenstemming is met het bereikte niveau (vgl. echter Canale 1986).

Dankzij deze verbreding van perspectief en van mogelijkheden kan er, afhankelijk van waar men de nadruk wil leggen, bij het testen van taalvaardigheid gekozen worden uit drie fundamenteel verschillende benaderingen: een psychometrische benadering, een contextuele benadering of een gepersonaliseerde benadering. De kenmerkende verschillen tussen deze drie benaderingen zijn samengevat in onderstaande tabel:

Tabel 1.3: Kenmerken van drie paradigma's bij assessment (op basis van Mabry 1999, McMillan 2001, Straetmans 2005.

	Psychometric	Contextual	Personalized
Inhoud	Meerkeuzevragen	Curriculum-afhankelijk	Individueel afgestemd
Afname	Gestandaardiseerd	Classroom settings	Individueel in tijd en ruimte
Opgaven	Voor alle leerlingen gelijk	Vooraf bepaald en individueel afgestemd	Afhankelijk van de leerling, deels zelf gekozen
Scoring	Machinaal scoorbaar	Door docent beoordeeld	Door docent en anderen
Rol van de leerling	Geen inbreng van de leerling	Zelfevaluatie is van belang	Zelfevaluatie is essentieel
Toetsdoelstelling	Voornamelijk summatief en certificerend	Summatief en formatief	Voornamelijk formatief
Feedback	Cijfer	Cijfer en klassikale bespreking	Uitgebreide feedback

Uit tabel 1.3 spreken de tegenstellingen waarmee we in het onderwijs te maken hebben: sommige toetsen, duidelijk een resultaat van het psychometrische paradigma, meten weliswaar op een betrouwbare manier maar aan de zinnigheid van de toetsresultaten wordt getwijfeld omdat de toets bepaalde relevante zaken niet zou meten – volgens sommigen niet kan meten. Andere assessmentvormen, met name de gepersonaliseerde vormen, doen misschien meer recht aan de individuele leerling maar zijn arbeidsintensief en geven geen betrouwbaar beeld. Een aanzet tot de bepleite procesbenadering is te vinden in Prodromou (1995, p. 23) waar hij adviezen voor de leerkrachten formuleert: *When you get a 'right' answer do not just move on to the next item: ask other students 'Do you agree?', 'What have you got?' Do not reveal the right answer too soon. The process is as important as the product.* Black & Wiliam (1998b, 12) signaleren hetzelfde: "*There are several ways to break this particular cycle. They involve giving pupils time to respond, asking them to discuss their thinking in pairs or in small groups so that a respondent is speaking on behalf of others (…)"*

In de volgende paragraaf wordt beschreven hoe de geschetste veranderingen doorwerken in het beroepsonderwijs.

1.3 Assessment in dynamisch perspectief

In de dynamiek van de geschetste veranderingen krijgt assessment een andere functie. Dit is te illustreren aan twee kernwoorden die telkens opduiken als het gaat om het mbo-onderwijs en de plaats van assessment daarbinnen. Het mbo wordt wel gekarakteriseerd als

“competentiegericht werken met behulp van authentieke opdrachten”. De revolutionaire veranderingen in het mbo zijn duidelijk te maken aan de hand van een toelichting op de termen competentiegericht en authentieke opdrachten.

1.3.1 Authentieke opdrachten

In het mbo vindt het leren plaats met behulp van min of meer authentieke opdrachten in gesimuleerde praktijksituaties (vgl. bv. Van der Sanden et al. 2003a, 2003b): tijdens deze praktijksimulaties ontwikkelen leerlingen competenties die zijn gekoppeld aan rollen uit de beroepspraktijk, zoals kapper, kok of kantoormedewerker. Vaktheorie en vakpraktijk worden geïntegreerd via voorgestructureerde opdrachten van het type ‘eerst-lezen-dan-voorbereiden-dan-uitvoeren’. De kennisoverdracht door de docent is grotendeels vervangen door een theorieopdracht die de leerling individueel bestudeert voordat de praktijkopdracht wordt uitgevoerd.

Een ultieme consequentie van leren met authentieke opdrachten is dat het leren ook plaats kan vinden buiten de school. In toenemende mate wordt er dan ook gebruik gemaakt van buitenschoolse opdrachten en is er naast het formele, gestuurde leren aandacht voor ongestuurd, informeel leren. In de optiek van sommigen moet leren liefst plaatsvinden op de (latere) werkplek (zie bv. Ruijters 2007).

In allerlei soorten beroepsopleidingen op mbo-niveau zijn de integratie van theorie en praktijk en de zelfwerkzaamheid van de leerling daarbij goed aanwijsbaar. Zo hanteert het Ontwikkelcentrum in Ede in aldaar ontwikkelde lesmaterialen vanaf 2000 een vast stramien waarbij de leerling bij taakuitvoering de volgende stappen zichtbaar maakt (zie ook Van Driel 2004):

- oriënteren op de uit te voeren taak
- voorkennis activeren
- de taakuitvoering plannen
- de noodzakelijke informatie verzamelen en structureren
- de taak uitvoeren en
- tot slot het hele proces evalueren.

De verwerving van theoretische kennis staat daarbij nadrukkelijk in het teken van praktijkopdrachten en de theoretische kenniscomponent wordt door de praktijk in gang gezet.

Een soortgelijke aanpak met integratie van theorie en praktijk zien we terug in de opleidingen voor bijvoorbeeld de metaalsector. Project Het Metalen Scharnierpunt baseert zich op een didactiek waarbij steevast zeven stappen aan de orde komen (vgl. Huisman & Pijnenburg 2009, p. 11 e.v.):

1 Oriëntatie: uitzoeken wat je moet doen, wat ga je maken? Je zorgt ervoor dat je in grote lijnen weet wat er gedaan moet worden.
2 Definitie: afspreken aan welke eisen het product moet voldoen, hoe gaat het er uitzien?
3 Ontwerp: een technische tekening ontwerpen van het product.
4 Werkvoorbereiding: het werk organiseren en plannen, zodat je kunt beginnen met de uitvoering.
5 Uitvoering: maken van het product.
6 Oplevering: het product aan de klant of de docent laten zien en bespreken of het goed is.
7 Nazorg: terugkijken naar hoe je gewerkt hebt; wat kun je een volgende keer beter doen.

In deze benadering is de beoordelingscomponent ook afwijkend van wat doorgaans gebeurt in het mbo (vgl. Huismans en Pijnenburg 2009, p. 20): in deze geïntegreerde benadering krijgt de leerling bij elke stap vragen over de theorie die nodig is om de praktijkopdracht correct te kunnen uitvoeren. Op deze manier wordt aan de theorie een functionele rol gegeven binnen het praktijkonderwijs. De beoordeling vindt per stap plaats door de docent die de prestaties cijfermatig waardeert: alleen bij een voldoende mag de leerling verder naar de volgende stap. Per taakuitvoering van een praktijkopdracht worden de relevante competenties van de leerlingen beoordeeld (goed, voldoende of onvoldoende). De beoordeling vindt op deze manier *criteriumgestuurd* plaats: in plaats van een *normreferenced* beoordeling waarbij bv. het groepsgemiddelde bepaalt waar de zak/slaaggrens ligt, is er nu een inhoudelijk te beargumenteren keuze gemaakt over wat er door de leerling gekend en beheerst moet worden om een voldoende te kunnen halen.

1.3.2 Competentiegericht evalueren

Met name het competentiegerichte karakter van het onderwijs zorgt voor andersoortige eisen aan assessment. Anders dan bij traditioneel onderwijs wordt bij het competentiegericht leren niet alleen het resultaat van het leren getoetst, maar wordt nadrukkelijk ook aandacht geschonken aan de bewandelde leerwegen, de didactische aanpak van de docent en de methoden om resultaten ervan in beeld te brengen. Maar wat competenties precies zijn, blijft onduidelijk. Doorgaans wordt onder competentie zoiets verstaan als een samenvoeging van kennis, vaardigheid en houding, plus het vermogen om deze combinatie optimaal in te zetten in nieuwe situaties. Het is een kwestie van kennen, kunnen, willen en daadwerkelijk toepassen.

Straetmans (2005, 19-29) bespreekt vijf verschillende opvattingen van het begrip competentie, gedestilleerd uit achtereenvolgens de Amerikaanse literatuur, Engelse vakliteratuur, de opvattingen van de Onderwijsraad, die van het COLO en van het Cito. Op basis van de uiteenlopende opvattingen komt Straetmans tot de volgende definitie: Competentie is "de bekwaamheid om op creatieve, bewuste en verantwoorde wijze geleerde kennis en vaardigheden in te (willen) zetten in slecht-gestructureerde taaksituaties uit een bepaald criteriumdomein, leidend tot een resultaat (proces en product) dat voldoet aan de geldende kwaliteitsnormen gelet op de te vervullen functie of rol van de beginnend beroepsbeoefenaar." Handzamer is de definitie van Kuhlemeijer (2005, p. 19): een competentie is '*de bekwaamheid om kennis, vaardigheden en houdingen geïntegreerd en bewust in te zetten in verschillende praktijksituaties volgens een erkende kwaliteitsstandaard*'.

Uit deze definitie is af te leiden dat het gaat om een bewuste inzet van inzichten, het gaat dus niet om aangeleerde trucjes. Het gaat bovendien om taaksituaties uit een bepaald beroepsgebonden criteriumdomein, dus het betreft geen algemeen inzetbare vaardigheden. Tot slot moet de competentie voldoen aan bepaalde kwaliteitsnormen die gerelateerd zijn aan wat een beginnend beroepsbeoefenaar moet kunnen. Elders, bijvoorbeeld in het Raamwerk Taal (Bohnenn et al. 2007, 18 e.v.) gaat het niet alleen om beroepsgerelateerd taalgebruik maar ook om taal

- om te participeren in het onderwijs,
- om te participeren in de wereld van het beroep, en
- om te participeren in de samenleving.

In hoofdstuk 2 schetsen we de consequenties van de veranderingen in het onderwijs voor de toetsinstrumenten die tegenwoordig in gebruik zijn.

1.4 Conclusies

In dit hoofdstuk is geschetst hoe het mbo zich onafhankelijk van het vmbo heeft ontwikkeld tot een nieuwe opleidingssoort met specifieke karakteristieken: het hedendaagse mbo is competentiegericht. Om die reden wordt er veelal gewerkt met authentieke opdrachten. Daarbij is er voor een onafhankelijk en autonoom schoolvak Nederlands geen plaats meer: de aandacht voor taalonderwijs komt geïntegreerd in andere vakken aan de orde
Verder is het mbo zeer heterogeen in didactiek. Tot slot wordt er zoveel mogelijk competentiegericht geëvalueerd.

Binnen deze dynamiek is er een steeds sterker klinkende roep ontstaan om andersoortige assessment: niet alleen toetsing achteraf – *assessment of learning* - maar ook assessment tijdens, zelfs geïntegreerd binnen het beroepsonderwijs. Hierbij gaat het om *assessment FOR learning*. De verhouding tussen *assessment of learning* en *assessment for learning* staat centraal in hoofdstuk 2.

Hoofdstuk 2 Functies, doelen en vormen van assessment

In het veranderend onderwijslandschap doet zich een grote diversiteit aan soorten assessment voor (in kaart gebracht door onder andere Brown & Hudson 1998, Straetmans 2006, Hendriks & Schoonman 2006). Vaak is geprobeerd om de veelheid aan soorten assessment in te delen. Zo merken Brown & Hudson (1998) op dat er bij taaltoetsing veel meer keuzen te maken zijn dan bij toetsing in andere schoolvakken; om het grote aantal mogelijkheden overzichtelijk te maken voor taaldocenten, rangschikken ze toetssoorten in drie hoofdcategorieën op basis van *het soort antwoord* dat de leerling moet geven. Ze onderscheiden daarbij:
(a) selected-response assessments (including true-false, matching, and multiple-choice assessments);
(b) constructed- response assessments (including fill-in, short-answer, and performance assessments); and
(c) personal-response assessments (including conference, portfolio, and self- or peer assessments).

Anderen proberen vormen van assessment te ordenen naar bijvoorbeeld *de functie van de assessment* of naar *de inhoud* die in de assessment aan de orde komt. Zo hanteert Nijhof 2006 onderstaande indeling.

Voorwaardelijke kennistoets			Performance assessment	
Kennis	Vaardigheid	Attitude	Proces	product

Voor ons doel lijkt deze indeling op het eerste gezicht niet erg bruikbaar: zo lijkt het weinig zinnig om vaardigheid en attitude onder de noemer kennistoets te scharen en in het taaldomein is de operationalisering van het onderscheid tussen kennis, vaardigheden en attitude problematisch (zie echter par. 5.3)

In dit hoofdstuk bespreken we respectievelijk
1 Functies en doelen van assessment
2 De rol van authenticiteit en autonomie bij assessment
3 Veelgebruikte vormen van assessment

2.1 Functies en doelen van assessment

Doorgaans worden er in de literatuur zes verschillende functies toegekend aan assessment (zie onder andere Johnson & Wentling 1996; Craigh 2001; Greenleaf et al. 1997, Laurier 2004). Een assessment kan de functie hebben van

- Intake- en selectietoets, bedoeld om te bepalen welke leerlingen in een bepaalde opleiding kunnen instromen;
- Plaatsingstoets, bedoeld om leerlingen te delen in niveaugroepen;
- Vorderingentoets, voor regelmatige rapportage van de vorderingen aan leerlingen en externe partijen, bijvoorbeeld de ouders van leerlingen;
- Diagnostische toets, bedoeld om na te gaan welke onderdelen door een leerling nog onvoldoende worden beheerst zodat daar het onderwijs op kan worden gericht;

- Certificering, bedoeld om extern te legitimeren dat de leerling een bepaald beheersingsniveau heeft dat nodig is voor de beroepsuitoefening;
- Programma-evaluatie, een niet individueel-gerichte vorm van toetsing waarmee het niveau van het geboden onderwijs in beeld wordt gebracht.

In dit rijtje functies kan opvallen dat sommige van de assessments voorafgaande aan een onderwijsperiode worden afgenomen, andere ergens middenin en weer andere na afloop van de onderwijsperiode. Verder valt op dat maar één soort assessment directe invloed heeft op inhoud of vormgeving van het onderwijs dat volgt op de afname, namelijk de diagnostische toets. Daarmee is de diagnostische toets een voorbeeld van een zogenaamde formatieve toets. Een diagnostische toets geeft door middel van feedback de leerling inzicht in wat hij of zij nog niet voldoende beheerst in relatie tot wat in het curriculum aan de orde komt (vgl. onder andere Brown & Hudson 1998, Black & Wiliam 1998a; Laurier 2004; Dunn & Mulvenon 2009); aan de deelscores kan worden afgelezen aan welke onderdelen de leerling met name moet gaan werken.

Afhankelijk van de vraag of het resultaat van assessment effect heeft op het daarop volgende onderwijs, is er sprake van formatieve assessment. Een formatieve assessment is - globaal gesproken - een evaluatie die ervoor bedoeld is om de leerlingen individueel te laten weten wat ze beheersen en wat nog niet. In navolging van Biggs (1996) wordt wel gesproken over *consequential validity*: daarmee wordt gedoeld op de "the intended and unintended effects of assessment on instruction or teaching and student learning" (vgl. Gielen et al 2003; Gulikers et al 2004; Gulikers et al 2006b).

Bij summatieve assessment, waartoe eindexamens per definitie behoren, is het doel niet om het onderwijs aan te passen naar aanleiding van de resultaten. Bij een summatieve vorderingentoets is de aard van de feedback anders: daar wordt immers vastgesteld wat de leerling nog niet beheerst, zonder dat de toets dicteert wat de consequentie van het niet beheersen is: dat kan een herexamen zijn, of het besluit om (een deel van) het jaar over te doen. Echter, er is pas sprake van formatief toetsen wanneer de resultaten op de toetsing (beter: taakstelling) leiden tot feedback die op individuele leerlingen gericht is.

Het cruciale verschil tussen formatieve en summatieve assessment ligt in de feedback naar aanleiding van een toetsafname: wanneer een assessment wordt gebruikt als een middel om de taalvaardigheid van een leerling vast te stellen en vervolgens nauwgezet aan te geven welke aspecten van de taalvaardigheid verdere ontwikkeling behoeven, dan is er sprake van formatieve assessment. Dit nauwgezet gebruik maken van toetsresultaten om de leerling aanwijzingen te geven voor het verdere leerproces is een minimale eerste vorm van feedback. Onder andere Laurier (2004) omschrijft de formatieve assessment als de monitorfunctie van toetsen. Doel van deze monitoring is volgens Laurier om het niveau van klasactiviteiten aan te passen aan het taalniveau van de leerlingen en om voortdurend feedback te geven aan de leerling over de individuele vorderingen. Ook elders in de vakliteratuur wordt de mogelijkheid benadrukt om klasactiviteiten aan te passen naar aanleiding van toetsresultaten (bv. Whitehead 2007; Pellegrino 2008). Wiliam & Black (1996) willen pas spreken over formative assessment als het niet alleen informatie oplevert voor verdere instructie maar die informatie ook daadwerkelijk effectief gebleken is: "To sum up, in order to serve a formative function, an assessment must yield evidence that, with

appropriate construct-referenced interpretations, indicates the existence of a gap between actual and desired levels of performance, *and suggests action that are in fact successful in closing the gap*." (cursivering US).

Als feedback – of werkzame feedback - inderdaad het onderscheid is tussen formatieve en summatieve assessment, is het in deze optiek strikt genomen aan de docent om te bepalen welke vorm van assessment gehanteerd worden als instrument om er feedback aan de leerlingen op te baseren. De leerkracht kan immers beslissen om op basis van enigerlei toets feedback te geven aan de leerlingen en het onderwijs naar inhoud, niveau of didactiek aan te passen. Afhankelijk van de consequenties voor het onderwijs is een toets als formatief of als summatief te beschouwen. Maar doorgaans wordt met name gekeken naar het beoogde doel volgens de makers van de toets en niet zozeer naar wat de individuele docent daarmee doet.

In ieder geval is feedback alleen in de vorm van een cijfer niet erg zinnig. Minstens zou de leerling - al dan niet via tussenkomst van de leerkracht - aan een assessment moeten kunnen ontlenen welke onderwijsdoelen nog niet voldoende worden beheerst. Feedback wordt maximaal bruikbaar wanneer de toetsdoelen en de gebruikte procedure zoveel mogelijk zijn afgestemd op onderwijsdoelen en de daar gehanteerde methode (Brown & Hudson 1998, 699).

Overigens proberen leerkrachten vaak aan een en dezelfde toets zowel formatieve als summatieve functies toe te kennen (vgl. Greenleaf 1997; Harlen 2006). Op een Nederlandse mailinglist voor docenten in het v.o. , de list-nederlands@digischool.nl kon op wo 20-1-2010 de volgende vraag worden aangetroffen (hier geanonimiseerd weergegeven):

> "In het kader van een plan om tot verbetering van onze eindexamenresultaten te komen heb ik een vraag namens mijn sectie. Weet iemand hoe we aan de precieze eindexamengegevens van oude examens - zowel Havo als Vwo - kunnen komen? We willen graag weten
> 1. het landelijk gemiddelde voor de tekst
> 2. het landelijk gemiddelde voor de samenvatting
> 3. het landelijk gemiddelde van de aftrek voor spel- en taalfouten.
> We willen deze vergelijken met onze schoolgegevens om zo te weten te komen welk probleem we als eerste moeten aanpakken."

Hier wordt geprobeerd om feedback te ontlenen aan een strikt summatief bedoelde toetsvorm teneinde toekomstig onderwijs te verbeteren. Omdat het onderwijs bedoeld is voor een andere groep leerlingen dan de groep bij wie de toets is afgenomen, is er strikt genomen geen sprake van formatief assessment-gebruik: het gaat niet om feedback aan individuele leerlingen die de assessment hebben gemaakt.

In de internationale literatuur loopt het onderscheid tussen summatief en formatief assessment parallel aan het onderscheid tussen *assessment for learning* en *assessment of learning*. Doorgaans wordt *assessment for learning* gelijkgesteld aan formatieve toetsing en *assessment of learning* aan summatieve toetsing (vgl. bv. Arter 2003, 264; Leahy et al. 2005; Harlen 2005). *Assessment for learning* wordt dan gedefineerd als toetsing met de bedoeling om de student optimaal te helpen in het verdere leerproces, terwijl *assessment of learning* bedoeld is om op bepaalde tijdstippen de vorderingen van de cursist in beeld te brengen. In het missionaire pamflet van de Assessment Reform Group (2002) wordt *assessment of learning* gedefineerd als: *Assessment for Learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.* Stiggins (2005b) maakt

onderscheid tussen de traditionele formatieve assessment en *assessment for learning:* terwijl formatieve assessments met enige regelmaat worden afgenomen zou de assessment for learning bedoeld zijn "to provide students, teachers, and parents with a continuous stream of evidence of student progress".

Met name de summatieve assessment heeft in de tweede helft van de 20e eeuw veel aandacht gehad, omdat dat highstake testing betreft: toetsen waarmee bepaald wordt of een leerling over kan naar het volgende leerjaar, een eindexamen om te bepalen of een leerling gecertificeerd mag worden of een toelatingsexamen om te bepalen of een student aan een studie mag beginnen. Het spreekt voor zich dat toetsen met een zo groot belang een hoge mate van betrouwbaarheid moeten bezitten. Met name in de Verenigde Staten heeft het gebruik van gestandaardiseerde toetsen een hoge vlucht genomen: vanaf de jaren '50 waren er op commerciële basis geconstrueerde toetsen verkrijgbaar die normreferenced zijn, dat wil zeggen dat een groepsgemiddelde bepalend is voor de zak/slaaggrens; in de jaren '70 werd er meer en meer gebruik gemaakt van "statewide testing programs" en vanaf de jaren '80 toetsen die op nationaal niveau waren gestandaardiseerd en genormeerd (zie Brennan 2006 voor een overzicht).

Johnson & Wentling (1996) schetsen de geschiedenis in de Verenigde Staten als volgt. In 1969 werd de National Assessment of Educational Progress gesubsideerd door het Congress. Het doel ervan was om alle leerlingen van 9, 13 en 17 jaar regelmatig te toetsen op leerprogressie; dit was een zo omvattende taakstelling dat er om redenen van efficiëntie vooral gebruik gemaakt werd van gestandaardiseerde toetsen met meerkeuzevragen. Zowel de vorm als de inhoud van deze toetsen stond in de ervaring van docenten ver af van wat er in het onderwijs aan de orde kwam. Dit werkte de onvrede met traditionele toetsen in de hand. De onvrede betreft met name:

- de nadruk op memorisatie in plaats van op begrip en toepassing;
- de gevreesde backwash-effecten op het onderwijs waardoor *drill and practice* oefening de boventoon ging voeren;
- de gebrekkige diagnostische waarde van de toetsen;
- de normering die *normreferenced* is (de zak/slaaggrens wordt afgeleid uit de groepsprestaties en niet vooraf op inhoudelijke gronden bepaald);
- in beroepsopleidingen was er geen relatie tussen toetsinhoud en benodigde beroepsvaardigheden;
- de toetsing was gericht op geïsoleerde deelvaardigheden in plaats van op complexe, samengestelde vaardigheden.

Op basis van bovenstaande kritiekpunten ontstond vanaf ongeveer 1985 bij docenten in de Amerikaanse *vocational training programs* en *prevocational training programs* een brede beweging die authentiek toetsen bepleitte: assessment die inhoudelijk herkenbaarder was voor de docenten en leerlingen en die bovendien direct bruikbare suggesties voor onderwijs-verbetering opleverde. Popham (2006) spreekt nog cynisch over een "federally installed school-evaluation scoreboard in the sky, and it reports whether a school's students have made adequate yearly progress (AYP) on the standardized achievement tests".

Het belangrijkste argument voor gebruik van formatieve assessment in de jaren '90 is, dat formatief toetsen een krachtige impuls geeft aan het leerproces. Met name maakt

formatieve assessment het mogelijk om effectieve feedback te geven aan de leerlingen (vgl. Black & William 1998a; Arter 2003; Stiggins 2005b; Stobart 2006; Marzano & Miedema 2008). In navolging daarvan merkt Shepard (2009) op: "*For example, feedback is more likely to lead to improved student learning if it is directed toward successful completion of the learning task, with clear guidance about how to improve*". Met name de invloedrijke metastudie van Black & William (1998a) over ruim 250 studies laat zien dat nauwkeurig gedoseerde formatieve assessment vrijwel altijd leidt tot een sterke vooruitgang in leerrendement (vgl. o.a. Harlen & Winter 2004, 392; Shepard 2009). Dit geldt met name wanneer

- learners receive feedback that enables them to know how to improve their work and take forward their learning;
- teachers and learners share an understanding of the goals of particular pieces of work;
- learners are involved in assessing their work (both self- and peer-assessment).

Bij *assessment for learning* ofwel formatieve assessment wordt vaak gebruik gemaakt van andersoortige, niet-traditionele toetsvormen. Om die reden worden *assessment for learning* en *alternative assessment* vaak in één adem genoemd. Waar *assessment for learning* betrekking heeft op de inhoudelijke invloed van toetsen op onderwijs, wordt met de term alternative assessment gedoeld op het gebruik van toetsvormen die voorheen ongebruikelijk waren. Op basis van een literatuurreview constateren Brown & Hudson (1998) dat alternative assessment-vormen aan de volgende kenmerken voldoen:

1 require students to perform, create, produce, or do something;
2. use real-world contexts or simulations;
3. are nonintrusive in that they extend the day-to-day classroom activities;
4. allow students to be assessed on what they normally do in class every day;
5. use tasks that represent meaningful instructional activities;
6. focus on processes as well as products;
7. tap into higher level thinking and problem-solving skills;
8. provide information about both the strengths and weaknesses of students;
9. are multiculturally sensitive when properly administered;
10. ensure that people, not machines, do the scoring, using human judgment;
11. encourage open disclosure of standards and rating criteria; and
12. call upon teachers to perform new instructional and assessment roles.

Bovenstaande lijst is erg heterogeen: het is een combinatie van eisen aan en kenmerken van alternative assessment en genoemde punten hebben nu eens betrekking op de inhoud en dan weer op de vorm, nu eens op het soort opgaven en denkprocessen dat ermee getriggerd wordt en dan weer op het soort onderwijs en de rolverdeling die erbij hoort.

Consistenter, want duidelijk alleen gericht op het leerproces en de rol van de leerkracht en de leerling daarin, is de lijst van kenmerken die de Assessment Reform Group (2002) heeft gepubliceerd over de eisen die men kan stellen aan *assessment for learning.* Omdat de 10 statements verhelderend zijn over de doelen van *assessment of learning* worden ze hier geciteerd en een voor een kort toegelicht.

Assessment of learning:

1) *is part of effective planning:* deze vorm van assessment maakt leerlingen bewust van de leerdoelen en van hoe ver ze nu zijn in hun leerproces op weg naar die leerdoelen
2) *focuses on how students learn:* maakt leerlingen bewust van hoe ze leren en van wat ze leren
3) *is central to classroom practice:* in de dagelijkse interactie lokken vragen en opdrachten leerlingreacties uit waarmee ze hun kennis, vaardigheden en competenties laten zien. Hierdoor krijgen leerkracht en leerlingen elke dag feedback op het leerproces.
4) *is a key professional skill:* leerkrachten moeten in staat zijn om te observeren, te interpreteren en feedback te geven op zo'n manier dat leerlingen zich van hun eigen kunnen bewust worden.
5) *is sensitive and constructive*: leerkrachten moeten zich ervan bewust zijn dat alleen positieve en juist gerichte feedback motiverend werkt.
6) *fosters motivation:* feedback moet motiveren en derhalve constructief zijn.
7) *promotes understanding of goals and criteria:* als leerlingen een actieve rol hebben in het bepalen van leerdoelen en assessment, hebben ze een beter begrip van de leerdoelen en de beoordelingscriteria.
8) *helps learners know how to improve:* leerkrachten kunnen de individuele leerling duidelijk maken hoe die de sterke punten verder kan uitbouwen.
9) *develops the capacity for self-assessment:* leerkrachten kunnen leerlingen zelfstandig en zelfbewust maken door ze te helpen realistische self-assessments te maken.
10) *recognises all educational achievement:* leerkrachten kunnen de leerlingen helpen door alle mogelijkheden om te leren ten volle te laten benutten en ze daarvan ook bewust te maken.

Met name door wetenschappers verbonden aan de Assessment Reform Group wordt het begrip assessment zodanig verbreed dat er ook allerlei mondelinge activiteiten op microniveau bij kunnen horen, wanneer die maar geïnterpreteerd kunnen worden als een bijdrage aan het leerproces en de bewustwording van eigen kunnen. Binnen deze gedachtewereld bespreekt bv. Stiggins (2005a) de "personal communication as assessment": naast de schriftelijke essay assessment en de peformance assessment onderscheidt Stiggins (2005, 67) de personal communication: "on reflection, it will become clear to you that certain forms of personal communication definitely do privide evidence of the level of student achievement". Overigens lijkt de dynamic assessment die sinds 2000 in de vakliteratuur aandacht krijgt, direct op dit gedachtegoed aan te sluiten (zie bv. Craig 2001; Stiggins 2005b; Poehner & Lantolf; Lantolf & Poehner 2008).

In deze reviewstudie hanteren we een strakkere definitie van assessment. We sluiten daarmee aan bij Straetmans (2006): assessment is "het doelbewust verzamelen en bewerken van informatie over de prestaties van personen in een bepaald domein, met het oog op het nemen van beslissingen over die personen". Met doelbewust verzamelen van informatie bedoelen we dan niet het voeren van korte dialoogjes in de klas waarbij het doel ligt in vergroting van zelfbewustzijn van de leerling, maar alleen activiteiten die voor leerkracht en leerling duidelijk herkenbaar zijn als middel om te achterhalen en administratief vast te leggen welk kennis- en vaardigheidsniveau de individuele leerling op dat moment heeft. Assessment veronderstelt in deze optiek dus een administratieve

verwerking. Bij de *assessment of learning* heeft deze administratieve verwerking geen directe consequenties voor het verdere leerproces van de betreffende leerling, bij *assessment for learning* heeft het wel directe consequenties voor het verdere leerproces voor de leerlingen die de assessment hebben ondergaan.

Als we ons beperken tot toetsen met administratieve verwerking, is de bespiegeling van Harlen (2006) nog van belang. Harlen 2006 beschrijft de manier waarop in de praktijk formative and summative assessment op elkaar lijken:

- Summatieve assessment wordt vaak gebruikt als manier om het leren te bevorderen (en dus als een vorm van formatieve assessment); dit illustreert Harlen aan het portfolio waar eerdere bijdragen van een student vaak worden vervangen door latere op "the common dimension of learning". De eerdere bijdragen hebben daarbij aantoonbaar een formatieve functie gehad, met aanwijzingen voor verbetering aan de leerkracht …
- Aan formatieve toetsen wordt een summatieve functie toegekend: wanneer een onderscheid wordt gemaakt tussen de formatieve gegevens en de interpretatie ervan, vindt Harlen het verdedigbaar dat gegevens uit formatieve evaluaties ook dienen als input voor summatieve interpretatie. Voorwaarde daarbij is, dat de data worden geïnterpreteerd met behulp van de criteria die gelden voor de summatieve evaluatie.

Op grond van haar observaties besluit Harlen dat er geen dichotomie bestaat tussen formatief en summatief toetsen, er is eerder sprake van een continuüm.

Gevreesd moet worden dat Harlen een veel voorkomend praktijkverschijnsel benoemt. Wat de leerling als een oefening beschouwt, blijkt achteraf mee te tellen als beoordelings-moment. Waarschijnlijk worden formatieve evaluaties vaak zelfs zonder herinterpretatie gebruikt binnen de summatieve evaluatie: de niet-aangekondigde klassikale overhoring in het vmbo levert dan een cijfer op dat meetelt voor het rapportcijfer. Slechts een enkele keer lijkt een toets ervoor bedoeld om richting te geven aan het onderwijs, volgens de leerkracht die het instrument inzet dan wel beleidsmakers op scholen, dan wel door de maker van de toets.

Het lijkt erop dat aan elke vorm van assessment elke toetsfunctie kan worden toegedicht. Hoe algemener de toetsdoelen echter geformuleerd zijn, hoe lastiger het is om op basis van de resultaten van de assessment zinnige en gerichte feedback te formuleren om het verdere onderwijs te richten. Wiliam & Black (1996, 544) merken hierover op: *(…) all assessments can be summative (i.e. have the potential to serve a summative function), but only some have the additional capability of serving formative functions. The question is not, therefore, can an assessment serve both functions, but the extent to which serving one has an adverse effect on its ability to serve the other*.

Dunn & Mulvenon (2009) maken onderscheid tussen formatief en summatief assessment aan de ene kant en evaluation aan de andere kant. De assessment betreft de gebruikte instrumenten, de evaluatie betreft het gebruik van de gegevens die voortvloeien uit assessment. Dit onderscheid sluit aan bij de door Hendriks & Schoonman (2006) en Straetmans (2006) bepleite tendens in het Nederlandse mbo en hbo om assessoren aan te stellen die geen andere betrokkenheid bij het onderwijs hebben dan toetsen afnemen en

assessments beoordelen (dat wil zeggen: scores toekennen); anderen cq. de reguliere leerkrachten nemen vervolgens beslissingen op basis van de verzamelde gegevens over beheersing. De drie kerntaken van een assessor bestaan uit het afnemen van de proeve van bekwaamheid, het beoordelen van het portfolio en het afnemen van een beoordelings-interview (vgl. Ehren 2009).

2.2 De rol van authenticiteit en autonomie bij assessment

In de *assessment of learning*-benadering wordt aan de leerder een grote mate van zelfstandigheid toegekend. In competentiegericht onderwijs is het doel de deelnemers zo ver te brengen dat ze beroepsmatige taken in natuurlijke omstandigheden kunnen vervullen. Als we de twee noties “zelfstandige leerder” en “natuurlijke omstandigheden” met elkaar combineren, kunnen we assessmentvormen onderscheiden naar de mate van authenticiteit en naar de aanwezigheid of afwezigheid van de assesor. Assessmentvormen kunnen op die manier worden ingedeeld naar deze twee dimensies (persoonlijke communicatie G. Straetmans):
a is de assessor expliciet en lijfelijk aanwezig?
b vindt de toetsactiviteit plaats onder natuurlijke omstandigheden dus zonder ingrijpen?

Door deze twee dimensies te combineren, ontstaat een overzicht met 4 kwadranten:

Tabel 2.1: Vier soorten toetssituaties

	Natuurlijke omstandigheden	**Geen natuurlijke omstandigheden**
Assessor afwezig	1 Retrospectief, bv. 360 graden feedback	2 Mystery guest methode (bv bij stages
Assessor aanwezig	3 Proeve van bekwaamheid, als LIO of als leider proefboerderij	4 Performance assessment: Klassieke tests Handsoff Simulaties Handson Alternative assessment

Ad 1: Assessment die onder volledig natuurlijke omstandigheden plaatsvindt en waarbij bovendien de assessor afwezig is, is zeldzaam. Het gaat hier om een proeve van bekwaamheid in de authentieke zin des woords, vergelijkbaar met de manier waarop een gezel van het schildersgilde eeuwen geleden het meesterstuk maakte: in het eigen atelier, op zichzelf teruggeworpen, zonder dat een meester hem of haar op de vingers keek en zonder tussentijds ingrijpen of bijsturing. Pas achteraf, als het werkstuk wordt gepresenteerd, vindt er een beoordeling plaats. Een moderne variant, die voornamelijk in het bedrijfsleven aan te treffen is, betreft de 360 graden feedback methode: daarbij geven alle betrokkenen een oordeel over het functioneren van betrokkene.

De Proeve van Bekwaamheid (PvB) zoals die tegenwoordig in het mbo wordt gebruikt, valt hier in sommige gevallen ook onder: deze PvB is een afsluitende toets in een realistische (authentieke) context, waarmee wordt vastgesteld of de kandidaat de beroepscompetenties in voldoende mate beheerst en geïntegreerd weet toe te passen. De PvB wordt per kerntaak of cluster van kerntaken en kernopgaven uitgevoerd en alleen al om die reden is er geen sprake van natuurlijke omstandigheden. Vaak - maar niet altijd - is de assessor afwezig en wordt het eindresultaat achteraf beoordeeld (zie verder par. 2.3).

Ad 2: Assessment die niet de natuurlijke omstandigheden nabootst en waarbij de assessor afwezig is, komt voor bij onder andere de stage: de stagiaire draait wel mee in het bedrijf maar beseft dat hij of zij nog geen volwaardige status heeft en dat er beoordeling vanuit de opleiding plaats zal vinden. Door aangewezen personen - soms een mystery guest, soms alle betrokkenen (tijdens de 360 gaden feedback) - worden aspecten van het handelen van de stagiaire beoordeeld.
Ook simulaties kunnen onder deze noemer vallen: daarbij moet de kandidaat een aantal zo realistisch mogelijk nagebootste taken uitvoeren zonder dat de assessor lijfelijk aanwezig is of kan ingrijpen. Een voorbeeld hiervan is de virtual realilty beroepsbekwaamheidstoets voor weginspecteurs (zie Van Gelooven & Veldkamp 2006). In deze toets wordt de kandidaat via een beeldscherm geconfronteerd met een aantal virtuele taaksituaties, zoals een onverwachte file of een pechgeval op een snelweg. De kandidaat geeft aan welke handelingen verricht moeten worden en ervaart direct de gevolgen ervan in de veranderende verkeerssituatie.

Ad 3: Assessment die onder volledig natuurlijke omstandigheden plaatsvindt onder het toeziend oog van de assessor, komt niet veel voor in het Nederlandse onderwijsstelsel: te denken valt aan bv. de LIO (Leraar in Opleiding) die voor de klas staat en moet laten zien een goede docent te zijn onder wiens leiding de leerlingen op een ordelijke en rustige manier de leerdoelen kunnen bereiken. De assessor zit achterin de klas om te kijken hoe het gaat maar grijpt principieel niet in. In het groenonderwijs krijgt de student de dagelijkse leiding over een proefboerderij die een week lang gemanaged moet worden (bron: pers. communicatie G. Straetmans).

Ad 4: Assessment die niet volledig de natuurlijke omstandigheden nabootst en waarbij de assessor bovendien aanwezig is om eventueel in te grijpen, komt verreweg het meeste voor in het onderwijs. In dit kwadrant bevinden zich 90% van de toetsvormen die in omloop zijn, variërend van een gestandaardiseerde meerkeuze-luistertoets tot een competence test waarbij gebruik gemaakt wordt van simulatie. In dit kwadrant maken we een nader onderscheid naar vier toetsvormen:

A Klassieke (gestandaardiseerde) tests, zoals de Citotoets eind basisonderwijs of de TAK woordenschattoets. Taaltoetsen van deze soort kunnen indirect meten, bv. met meerkeuzetoetsen waarbij de kandidaat de beste reactie moet kiezen in een problematische communicatieve situatie, zoals in toetsbatterij STAAL (vgl. Schuurs 1993). Er kan ook direct worden gemeten, bv. bij een spreektoets waarbij een situatie wordt geschetst en de kandidaat een mondelinge respons moet geven op een auditieve stimulus (vgl. het NT2-examen Spreken).

B Hands-off instrumenten: een of meer aan de praktijk gerelateerde probleemsituaties worden nagebootst, de kandidaat moet de best mogelijke oplossing kiezen of aandragen.

C Simulaties ofwel gedragsproeven: de werksituatie is hierbij nagebootst, al dan niet met behulp van de pc en andere apparatuur. De assessor is aanwezig en kan ingrijpen als dat nodig is, bv. om ervoor te zorgen dat een minimum aantal cruciale beoordelingssituaties aan de orde komt.

D Hands-on: de werksituatie wordt zoveel mogelijk intact gehouden of nagebootst terwijl de assessor aanwezig is. Deze variant kan allerlei vormen aannemen, bv. een rollenspel op school of een situatie op de werkplek waarbij de te beoordelen leerling de rol overneemt van de werknemer in de betreffende functie.

Deze categorieën kunnen we illustreren aan wat er gebeurt bij het (leren) autorijden. Een toets van type A, de klassieke (gestandaardiseerde) test is het theorie-examen. Dat examen bestaat uit een groot aantal gestandaardiseerde tweekeuzevragen waarbij soms kennis van de verkeersreglementen wordt getoetst, soms gevraagd wordt naar de beste manier van handelen enz.
Toetsen van type B, de hands-off instrumenten, komen in de praktijk van het leren autorijden wel voor maar hebben daar geen toets-status: in rijscholen vinden groepslessen plaats waarbij de instructeur met behulp van dia's een verkeerssituatie voorschotelt; de cursisten moeten dan individueel of in de vorm van een groepsgesprek beoordelen welk rijgedrag in de betreffende situatie het meest geschikt is.
Toets type C, de simulatie, vindt plaats met behulp van simulatie-apparatuur waarbij het lijkt alsof de kandidaat een auto bestuurt; feitelijk gaat het om virtual reality. De assessor kan aanwijzingen geven en bepaalt welke onderdelen aan bod komen.
Toets type D betreft het praktische deel van het rij-examen: de kandidaat bestuurt hands-on een auto terwijl de examinator ernaast zit en indien nodig kan ingrijpen.

Bij assessment uit het derde kwadrant horen dus allerlei traditionele en modernere toetsvormen die in verschillende mate voldoen aan de uiteenlopende eisen die aan toetsen worden gesteld. De modernere toetsbenaderingen worden aangeduid als "alternative assessment" en kent drie hoofdgroepen (vgl. Straetmans 2006, 17 e.v.):

Direct assessment: deze benadering is in de USA populair geworden als reactie op de veelheid aan gestandaardiseerde toetsen in het onderwijs en de aanpassingen in het onderwijs om met dit soort toetsen te oefenen in een ultieme poging om het resultaat te laten stijgen (vgl. Brennan 2006). Bij direct assessment krijgt de kandidaat taken die de te meten vaardigheid direct uitlokken. Bij het toetsen van schrijfvaardigheid krijgt de kandidaat dus een schrijfopdracht en niet een aantal meerkeuzevragen over het schrijfproces.

Authentic assessment: de term is populair geworden dankzij Wiggins (1989) die directe toetsing van de beoogde vaardigheden bepleitte. In deze benadering moeten de taken in de test authentiek zijn, dwz gelijk aan de taken waarmee de beroepsbeoefenaar geconfronteerd wordt; volgens Wiggins (1989, 703) heeft een kandidaat bij een "true test" de mogelijkheid "to ask for clarificaton of questions and explain his or her answers".

Performance assessment: bij deze vorm van assessment voert de kandidaat een zo realistisch mogelijke opdracht uit onder omstandigheden die zo goed mogelijk de werkelijkheid van de beroepspraktijk nabootsen (vgl. bv. Roelofs & Straetmans 2006, 18). Als gevolg daarvan worden niet alleen productkenmerken beoordeeld – zoals bij de direct assessment nog wel het geval is – maar ook proceskenmerken.

Met name direct assessment, authentic assessment en performance assessment hebben de laatste 15 jaar veel aandacht gekregen in de vakliteratuur. Om te voorkomen dat allerlei vormen van alternative assessment die onder deze noemers vallen - te denken valt aan portfolio's, conferences, diaries, self-assessments and peer assessments - als nieuwlichterij en om die reden als weinig degelijk worden afgedaan, stellen Brown & Hudson (1998) voor om te spreken over 'alternatives in assessment'. Liever dan dat deze nieuwe vormen van assessment als alternatief worden gezien, hebben ze dat docenten alle toetsvormen, klassiek of nieuw, als gelijkwaardige alternatieven van elkaar zien. In hoofdstuk 3 brengen we kenmerken van beide soorten met meer detail in beeld.

Op grond van bovenstaande kan de indruk ontstaan dat er in mbo en vmbo een veelheid aan toetsvormen in gebruik is; het merendeel van de genoemde soorten is met name veelbesproken en aanbevolen in de literatuur, maar de meeste soorten worden niet of slechts op kleine schaal beproefd (zie verder par. 2.3).

De scope die in het bedrijfsleven wordt gebruikt, is juist veel breder (vgl. bv. Hendriks & Schoonman). Een voorbeeld daarvan is te vinden in het zogenaamde *5-P model*, gebaseerd op het model voor portfolio rationalisatie (zie Stoel 2006). Dit model wordt door opleiders in het bedrijfsleven wel gehanteerd om effect en rendement van bedrijfsopleidingen te meten. Het model pretendeert de effecten van een bedrijfsopleiding op vijf parameters te meten:

- *Effect op pleasure: plezier en 'good feeling', voornamelijk bij de deelnemers*
- *Effect op potential: aantoonbaar verworven kennis, inzicht en vaardigheden*
- *Effect op performance: toepassing van het geleerde in de praktijk*
- *Effect op productivity: meer productie, hogere klanttevredenheid, minder klachten, etc.*
- *Effect op profitability: winst en/of (financieel) toegevoegde waarde*

Al met al is er inmiddels een bonte verzameling van assessment-instrumenten bekend. In paragraaf 2.3 bespreken we kort een aantal veelgebruikte instrumenten, in hoofdstuk 3 staat de vraag centraal aan welke kenmerken deze instrumenten voldoen resp. zouden moeten voldoen.

2.3 Veelgebruikte vormen van assessment

De geschetste onvrede met de gestandaardiseerde toetsen heeft geleid tot een bonte verzameling van assessmentvormen die – al dan niet terecht –in het onderwijs worden ingezet. Met assessmentvorm bedoelen we hier: een concreet pakket van regels en procedures dat voorschrijft hoe gedrag wordt uitgelokt, gescoord en geëvalueerd (vgl. Straetmans 2006). Sommige van deze assessmentvormen zijn veelbesproken maar (nog) niet vaak in de praktijk ingezet; andere worden al sinds jaar en dag gebruikt, terwijl uit onderzoek blijkt dat daar de nodige nadelen aan kleven. Een aantal prominente assessmentvormen wordt in deze paragraaf besproken.
De selectie van op te nemen assessmentvormen is als volgt tot stand gekomen: eerst is in de literatuur een aantal veelgenoemde assessmentvormen geregistreerd; deze bleken bijna alle ook te zijn genoemd door Hendriks & Schoonman (2006). Om te borgen dat er geen belangrijke assessmentvormen over het hoofd werden gezien, is op diverse websites (VET, Kennisnet, mbo-instellingen) gezocht naar aanvullende vormen van assessment. Tot slot is in een aantal halfgestructureerde interviews met mbo-docenten geïnformeerd naar gehanteerde vormen van assessment. Wanneer een assessmentvorm voorkomt in twee of meer van genoemde bronnen, is deze opgenomen voor verdere studie.

Tabel 2.2: Veelgenoemde vormen van assessment

		Hendriks & Schoonman 2006	Websites (VET Kennisnet, ROC's)	Genoemd tijdens veld-raadpleging	Opgenomen in onderzoek
1	afstudeeropdracht		*	*	*
2	beroepsproduct	*			
3	casustoets	*	*	*	*
4	criteriumgericht interview		*	*	*
5	essaytoets		*	*	*
6	gedragsassessment		*	*	*
7	intaketoets		*		
8	kennistoets		*	*	*
9	peer assessment	*	*	*	*
10	portfolio assessment	*	*	*	*
11	presentatie	*			
12	procesverslag	*			
13	projectopdracht		*	*	*
14	reflectie-opdracht	*	*	*	*
15	self-assessment	*		*	*
16	simulatie	*			
17	stage-opdracht		*	*	*
18	vaardigheidstoets		*	*	*
19	voortgangstoets	*	*		*

Naast de in tabel 2.2 genoemde toetsvormen is de computergestuurde toets toegevoegd als aparte toetsvorm: op grond van het toenemende lerarentekort en het toenemende gebruik van ict-applicaties in het onderwijs valt te verwachten dat deze assessmentvorm de komende jaren een belangrijker rol zal gaan vervullen.

In het navolgende bespreken we de hierboven genoemde vormen van assessment kort. We besteden vooral aandacht aan drie relatief nieuwe vormen van assessment: het criteriumgericht interview (4), het portfolio (10) en het selfassessment (15).

1 Een **afstudeeropdracht** is een eindopdracht ter afronding van de opleiding. Op het mbo heeft deze opdracht vaak de vorm van een proeve van bekwaamheid ofwel PvB (vgl. par. 2.1): een 'meesterproef' die de kandidaat zelfstandig uitvoert. Het betreft een complexe opdracht waarin kennis uit theorie en praktijk worden verbonden met praktijkgericht onderzoek. De competenties van de deelnemer worden getoetst in een zo authentiek mogelijke context. Met de PvB kan getoetst worden of een examendeelnemer na het mbo-onderwijs aan de slag kan als beginnend beroepsbeoefenaar in het beroep waarvoor hij of zij is opgeleid. Als de PvB op niveau 4 wordt gemaakt is op basis daarvan ook doorstroom naar het HBO een mogelijkheid. Een PvB is meestal onderdeel van de zogenaamde 'methodenmix' waarbij van verschillende assessmentvormen gebruik gemaakt wordt.

   In een PvB moet de examendeelnemer in een zo authentiek mogelijke omgeving opdrachten uitvoeren die hij of zij ook in het werkveld tegen zal kunnen komen. De deelnemers leveren producten op die overeenkomen met producten uit hun latere beroepspraktijk. Daarnaast moeten ze zich in een eindgesprek verantwoorden voor de door hen gemaakte keuzes en beslissingen. De beoordeling vindt vaak plaats door onafhankelijke, getrainde assessoren. De kortste PvB duurt 160 uur, de langste een half jaar. De Maa (2007) geeft aanwijzingen over hoe taaltaken systematisch opgenomen kunnen worden in een PvB .

2 Het **beroepsproduct** is het resultaat van een opdracht waarin alle fasen van het oplossen van een beroepstaak aan de orde komen. Daarmee is het vergelijkbaar met een afstudeeropdracht, maar een beroepsopdracht hoeft geen onderdeel te vormen van de examenopdracht. Vanwege de geringe verschillen met de afstudeeropdracht blijft de beroepsopdracht verder buiten beschouwing.

3 De **casustoets** bestaat uit een gevalsbeschrijving die sterk lijkt op een praktijkprobleem. De kandidaat heeft de taak om te handelen zoals hij dat zou doen in de beroepsuitoefening en daarvan verslag te doen. In dat verslag dienen zowel de resultaten als het doorlopen proces met de beslissingen daarin aan de orde te komen. De casustoets lijkt in allerlei beroeps-gerelateerde opleidingen voor te komen, vanaf vmbo tot en met universitaire opleidingen.

4 Het **criteriumgericht interview** is een gestructureerd gesprek met een of meer beoordelaars over (beroeps)situatie(s) waarin de leerling zijn/haar competenties laat zien. Dit interview kan worden ingezet bij portfolio assessment, gedragsassessment en intaketoets (bij toelatingsexamens); het instrument vormt onderdeel van het voor mbo-leerlingen verplichte LLB-examen (Leren Loopbaan Beroep). Bij het LLB-examen hebben de beoordelaars voorafgaande aan het gesprek het examenportfolio in kunnen zien en per kerntaak een prestatie-indicator gekozen waarover de kandidaat ondervraagd wordt. Het examenportfolio

ligt op tafel zodat beide partijen aan de onderdelen ervan kunnen refereren. De ene beoordelaar interviewt de kandidaat, vaak met gebruikmaking van de START(T)-methode (Situatie, Taak, Actie, Resultaat, Reflectie en eventueel Transfer); de andere beoordelaar maakt aantekeningen op procesverslagen. Op basis van de aantekeningen wordt achteraf per kerntaak een beoordelingsformulier ingevuld; deze formulieren dienen als basis voor de eindbeoordeling. In het gesprek kunnen in totaal wel 50 aspecten worden beoordeeld. Een voorbeeld van een scoreformulier bij het criteriumgerichte interview staat in figuur 2.3.

Figuur 2.3: Voorbeeld van een scoreformulier bij het criteriumgerichte interview (Uit Studentenhandleiding; Leerlijn LLB. Bron: http://mbo2010.kennisnet.nl)

Criteriumgericht interview Beoordelingsformulier	**Leren, Loopbaan en Burgerschap Kerntaak …**	**niveau 3 en 4**

Indicator: _____

Nr.	**Kern taak**	**Werk proces**	**Vraag**	**0**	**1**	**2**	**3**
1			Wat was de situatie? (Wat was er aan de hand?)	0	1	2	3
2			Wat was jouw taak? (Wat moest je doen?)	0	1	2	3
3			Welke acties ondernam je? (Wat heb je precies gedaan?)	0	1	2	3
4			Wat was het resultaat? (Wat was de uitwerking?)	0	1	2	3
5			Hoe kijk je erop terug? (Wat heb je hiervan geleerd?)	0	1	2	3
6			In welke andere situatie kan je de ervaring gebruiken?	0	1	2	3

Totaal	**Controle: 6 scores**	**…x 0**	**…x 1**	**…x 2**	**…x 3**
Totaal generaal					

5 Een **essaytoets** is een schriftelijke kennistoets die bestaat uit één of enkele open vragen waarmee de inhoudelijke kennis van de deelnemer kan worden getoetst. De toetsvorm - de deelnemer moet langere teksten schrijven - maakt het mogelijk om combinaties van inzichten, toepassingen, analyses en argumentaties te demonstreren. Daarnaast kan de kandidaat in de tekst laten zien over methodisch inzicht en reflectief vermogen te beschikken.

6 Het **gedragsassessment** is een specifieke vorm van performance assessment waarbij op de gedragingen van de kandidaat wordt gelet aan de hand van een aantal criteria. Dit assessment is een integrale toetsvorm waarbij de kandidaat taken moet uitvoeren in (nagebootste) kritische beroepssituaties om op die manier het bereikte competentieniveau te demonstreren. De gehanteerde beoordelingscriteria zijn taakafhankelijk en kunnen daardoor zeer van elkaar verschillen.

7 Met de term **intaketoets** wordt niet zozeer gedoeld op een toetsvorm alswel op een toetsfunctie: een intaketoets wordt gebruikt om te bepalen welke leerlingen in een bepaalde opleiding kunnen instromen. Daarmee liggen de vorm en de inhoud van de toets nog geenszins vast, die kunnen op allerlei manieren worden geconcretiseerd. Een intaketoets heeft een redelijk algemeen karakter en is om die reden minder geschikt om er feedback op te baseren. Wel is het denkbaar dat de toets is opgebouwd uit onderdelen die elk afzonderlijk discrete toetsdoelen toetst. In dat geval kan aan deelnemers per toetsdoel gespecificeerde feedback worden gegeven.

8 Een **kennistoets** is gericht op het toetsen van kennis. In de bekende taxonomie van Bloom wordt kennis als het meest elementaire niveau beschouwd. Het lijkt erop dat kennis in de vakliteratuur over toetsing wordt genegeerd; in de beschreven onderwijskundige vernieuwingsgolf komen kennistoetsen nauwelijks aan bod. De algemene aanname is, dat kennis automatisch wordt afgetoetst in vaardigheids- en competence-toetsen. In de dagelijkse schoolpraktijk wordt in de vorm van proefwerken en andere door de docent vervaardigde toetsen wel vaak de kenniscomponent getoetst.

9 Bij **peer assessment** beoordelen leerlingen elkaar bij het uitvoeren van een opdracht in groepsverband. De beoordeling zou betrekking moeten hebben op de participatie tijdens het gezamenlijk werken, op de kwaliteit van de inbreng en eventueel op de mondelinge presentatie of het schriftelijke verslag. Officieel is peer assessment geen toetsvorm: de beoordelaar is immers geen officieel aangewezen instantie. Ook is in de praktijk wel duidelijk dat vaak andere dan de te beoordelen aspecten zwaarwegende invloed hebben op het uit te spreken oordeel. Bovendien ontbreekt in de praktijk enig protocol dat richting geeft aan de manier waarop de oordelen tot stand zouden moeten komen. Toch komt het vaak voor dat de resultaten van de peer assessment in een groepsbeoordelingsgesprek aan de orde komen en zo – al dan niet bewust – invloed uitoefenen op de officiële beoordeling.
Dit alles laat onverlet dat peer assessment een krachtig onderwijsmiddel kan zijn; met name in het schrijfonderwijs zijn er positieve effecten mee behaald (zie bv. Galbraith & Rijlaarsdam (1999).

Het verband tussen peer assessment en formele evaluatie is met name bestudeerd in de context van het hoger onderwijs. In dit type onderzoek wordt vaak gebruik gemaakt van onder andere essays, mondelinge presentaties en meerkeuzetoetsen. De bevindingen zijn zeer divers: sommige auteurs rapporteren redelijk hoge correlaties (tot r = .70) waar andere studies signaleren dat de interrater betrouwbaarheid bij tutoren en professionele assessoren hoger was dan bij de peer evaluation. Deze bevinding staat in een vreemd daglicht omdat de gevonden correlaties bij professionals varieerden tussen .40 en .53 (vgl. Topping 2003). Anders gezegd, de betrouwbaarheid van peer evaluatie is laag.

Recent onderzoek (vgl. Braaksma et al 2004; Rijlaarsdam et al. 2008) laat zien dat het samenwerken met en observeren van peer writers leidt tot snellere verwerving van higher

order skills zoals plannen tijdens het schrijven. Dit soort observatie zou een middel kunnen zijn "to close the gap" tussen de onderwijsdoelen en de positie waarin een leerling zich op een gegeven moment bevindt (Black & Wiliam 1998b).

10 Een **portfolio** is een persoonlijk dossier, met teksten en andersoortige producten van en over de leerling, inclusief beoordelingen en zelfevaluaties. Een - eventueel elektronisch opgeslagen - portfolio is bedoeld voor het vastleggen en tonen van leerresultaten en eventueel het leerproces aan zichzelf of aan anderen. Aan het portfolio als tastbaar "bewijs" van het leren worden drie hoofdfuncties toegekend (vgl. Wolf et al. 1997):

- het *ontwikkelingsgerichte portfolio*: dit heeft als doel de professionele groei van de samensteller te documenteren en het leerproces zichtbaar te maken, zowel voor de omgeving als voor het individu zelf (is dus ook een bewustwordingsproces);
- het *assessmentportfolio* (ook wel: beoordelingsportfolio): dit bevat materiaal op basis waarvan uitspraken gedaan kunnen worden over bekwaamheden die de leerling moet bezitten;
- het *presentatieportfolio* (ook wel 'showcase'): dit heeft vooral een toonfunctie, als hulp bij bv. sollicitaties. In dit type portfolio worden uitspraken aannemelijk gemaakt die de samensteller over zijn of haar eigen competenties / capaciteiten doet.

Met name bij universiteiten, hbo- en mbo-instellingen heeft het portfolio een hoge vlucht genomen (vgl. Inspectie van het Onderwijs 2003). Vanaf 2000 bieden tal van ict-bedrijven software voor elektronische portfolio's aan waarin leerlingen hun prestaties ter beoordeling kunnen opslaan (vgl. bv. Booth et al. 2003, p. 66).

Davies & LeMahieu (2003) constateren op basis van literatuurstudie over portfolio-gebruik:

- Betrouwbaarheid van beoordeling wordt bemoeilijkt doordat de inhoud en opbouw van portfolio's erg kan verschillen. Leerkrachten zouden na een beoordelaarstraining wel meer consistent en betrouwbaar beoordelen.
- De bijdrage van portfolio's aan het onderwijs wordt belangrijk geacht: de algemene bevinding is dat leerlingen hun interesses en competenties in de volle breedte kunnen demonstreren in het portfolio. De betrokkenheid van leerlingen wordt erdoor vergroot en dit beïnvloedt het leerrendement in positieve zin. Bovendien bevordert het portfolio het communiceren over leren tussen leerlingen.
- Specifieke descriptieve en motiverende feedback op portfolio's bevordert het leren. Het leren wordt met name bevorderd met descriptieve feedback die aangeeft
    - wat de leerling goed doet en verder kan uitbouwen,
    - op welke criteria het werk van de leerling nog niet voldoet, en
    - wat de leerling het beste kan doen om de performance te verbeteren.

Op basis hiervan constateren Davies & LeMahieu (2003) dat het portfolio een krachtig middel is bij de *assessment for learning*.

Inmiddels zijn er voor allerlei beroepsrichtingen specifieke taalportfolio's in ontwikkeling. Doel van deze taalportfolio's is om een deugdelijke assessment te verkrijgen van de taalcompetenties van leerlingen in beroepsmatige settings. De taken in taalassessments sluiten aan bij de leerdoelen van kandidaten, lokken functioneel taalgedrag uit in (nagebootste) beroepssituaties, integreren meestal verschillende vaardigheden en hebben heldere beoordelingscriteria. Specifiek voor de talige component moeten de taaltaken

beantwoorden aan de niveauspecificaties die, met verwijzing naar het Europese Referentiekader (vgl. Council of Europe 2001), in het kwalificatiedossier zijn vastgelegd.

Een voorbeeld van een beroepsspecifiek taalportfolio is te vinden in Bharosa et al. (2008) onder de veelzeggende titel *Uiterlijke verzorging, toets je taal!* De opdrachten in deze prototypisch bedoelde bundel zijn volgens de auteurs zowel formatief als summatief te gebruiken (Bharosa et al., 2008, 5).: "de ontwikkeling van opdrachten die zowel voor formatieve als summatieve beoordeling gebruikt kunnen worden". Per taak zijn er beoordelingscriteria gegeven, uitgesplitst naar vorm, opdrachtspecifieke inhoudelijke normen en algemene taalnormen voor het betreffende genre. Bij wijze van voorbeeld bevat figuur 2.4 een kopie van een beoordelingsblad uit het portfolio.

Het taalportfolio wordt zelden als alleenstaand instrument ingezet. Driessen (2007) rapporteert over een experiment waarbij ze 59 taaldocenten enkele mogelijke scenario's voor een mbo-taalexamen heeft voorgelegd. Twee varianten waarbij een gestandaardiseerde taaltoets en een taalportfolio werden gecombineerd (de toets al dan niet als onderdeel van het portfolio of als zelfstandig onderdeel) konden te zamen rekenen op 23 stemmen. Het scenario waarbij de gestandaardiseerde taaltoets onderdeel vormt van een taalportfolio waarover een gesprek wordt gevoerd plus een aanvullende proeve van bekwaamheid had met 21 stemmen het grootste aantal voorstanders onder de docenten.

Brown (2002) vat een aantal bevindingen over het portfolio en de voor- en nadelen ervan samen. Door gegevens te verzamelen is het mogelijk het eigen leerproces te sturen aan de hand van gewenste en reeds behaalde prestaties (vaardigheden, competenties); de leerling kan zijn resultaten (en eventueel het leerproces) aan anderen tonen. Met name de betrouwbaarheid van de beoordeling van portfolio's is problematisch en een beoordelingsprocedure kan nauwelijks gestandaardiseerd worden omdat de competence performance die erin gereflecteerd wordt "may vary in depth, in approach, and in the specificity of the professional work adressed". De betrouwbaarheid van de beoordeling zou vergroot kunnen worden door clearcut criteria tussen beoordelaars af te spreken en door ervoor te zorgen dat de performance indicators representatief zijn voor de competenties die beheerst moeten worden. In Nederland heeft met name Straetmans (2004, 2006) geprobeerd om de portfolio scoring te standaardiseren met behulp van protocollaire beoordelingsprocedures (vgl. ook Sluijsmans et al. 2008). In hoeverre de ontwikkelde methoden daadwerkelijk in praktijk worden gebracht, en met welk effect, is niet bekend.

Figuur 2.4: Beoordelingsmodel Spreken / Presenteren voor Uiterlijke verzorging (Bron: Bharosa 2007, p. 32)

Beoordelingsmodel De diabetische voet

Spreken B2

A. Vormspecificaties		**V**	**O**	**Verbeterpunten**	**Beslissing**
Het is een presentatie. Toelichting: het is geen presentatie als • de deelnemer de tekst voorleest; • de presentatie 2x de voorgeschreven tijd duurt; • de presentatie minder dan de helft van de voorgeschreven tijd duurt.					Voldoende = voor alle criteria een 'V'. Bij onvoldoende stoppen.
Het is verstaanbaar					
B. Opdrachtspecifieke kenmerken (gaan over vakinhoud)					
De deelnemer beschrijft op een duidelijke manier • het probleem; • de behandeling; • het advies.					Voldoende = voor drie criteria een 'V'. Bij onvoldoende stoppen.
De deelnemer heeft vragen van klasgenoten en docent inhoudelijk adequaat beantwoord.					
De deelnemer geeft de inhoud juist weer. Toelichting: haalt feiten en meningen niet door elkaar.					
C. Taalspecificaties (Selecteer uit kenmerken van taakuitvoering Raamwerk Nederlands voor (V)MBO, Spreken B2)					
Samenhang	Maakt publiek opbouw en structuur duidelijk en volgt deze ook.				Voor vier criteria een 'V', anders geen B2.
Afstemming op doel	Presentatie is informatief en overtuigend.				
Afstemming op gesprekspartners	Past woordgebruik aan publiek aan. Wijkt af van voorbereide tekst op grond van vragen.				
Woordgebruik en woordenschat	Varieert en is trefzeker.				
Grammatica	Toont een goede beheersing van de grammatica: kleine onvolkomenheden komen voor, maar worden meestal zelf hersteld.				
Verstaanbaarheid	(zie vorm).				

Datum:	Naam deelnemer:	Klas:
Opdracht:		
	Formatief (voortgang)	Naam en handtekening beoordelaar(s):
	Summatief/ kwalificerend	
	Voldoende (voor onderdeel A, B en C)	
	Onvoldoende	

11 De **presentatie** wordt zeker met grote regelmaat gebruikt als assessmentvorm in vmbo en mbo, maar kennelijk wordt ze in de praktijk niet ervaren als een zelfstandige toetsvorm. Wel is de presentatie vaak een cruciaal onderdeel van het assessment. De beoordelingsnormen bij een presentatie zijn vaak afhankelijk van het inzicht van de beoordelaar.

12 In een **procesverslag** beschrijft de leerling van begin tot eind welke stappen zijn ondernomen om een bepaalde taak uit te voeren of een probleem op te lossen. Alle fasen worden systematisch doorlopen en beschreven: de oriëntatie op het probleem, de analyse in deelstappen, het genereren van oplossingen en het maken van een keuze uit de alternatieven inclusief het realiseren en implementeren van de oplossing. Kennelijk wordt ook het procesverslag niet als een zelfstandig hanteerbare assessmentvorm gezien, in elk geval is het niet vaak als zodanig benoemd in de literatuur en door geraadpleegde docenten.

13 Bij een **projectopdracht** is er sprake van een probleem dat door een bedrijf, organisatie of instelling is aangedragen. De leerling dient dit probleem individueel of in groepsverband op te lossen of te beantwoorden. De kwaliteitseisen die aan het product worden gesteld, moeten in toenemende mate door de leerling zelf worden verantwoord.

De beoordeling van individuele bijdragen aan groepswerk is lastig. In elk geval dient de beoordelaar op de hoogte te zijn van de taakverdeling en het tijdstip van uitvoering van de activiteiten. Verder lijkt regelmatig voortgangsoverleg tussen beoordelaar en betrokkenen noodzakelijk. Daarbij kan een logboek een handig hulpmiddel zijn, zodat leerlingen kunnen vastleggen welke projectactiviteiten ze op een dag hebben gedaan. Aan de hand van de rapportages en de daarbij behorende (aparte) beoordelingsvoorschriften kunnen de individuele prestaties worden beoordeeld en gewaardeerd.

De projectopdracht wordt in het mbo vaak gecombineerd met een kennistoets en een assessment waarbij het project in een verkorte versie wordt overgedaan. Een project omvat een aantal competenties, maar die zijn als vak niet meer herkenbaar. Overigens heeft een projectopdracht meestal betrekking op een beroepsgerelateerde activiteit waarbinnen taalgebruik niet afzonderlijk wordt beoordeeld (zie bv. OER ROC West-Brabant, Crebocode: 90400, Cohort: 2005 – 2009, bron: http://herontwerpmbo.kennisnet.nl/attachments/session=cloud_mmbase+909423/A2-OER_Mediavormgever_05-09-Radiuscollege.doc).

14 Een **reflectie-opdracht** is niet meer of minder dan een werkstuk waarin de leerling zelfkritisch verslag doet van eigen ervaringen in bepaalde leer- en beroepssituaties. Daarbij kunnen de communicatieve vaardigheden aan de orde komen, maar dat hoeft niet. Er zijn geen voorbeelden van reflectie-opdrachten aangetroffen waarbij aan de leerling expliciet gevraagd wordt naar bespiegeling van de eigen communicatieve competenties bij het uitvoeren van beroepsgerelateerde taken.

15 Bij de **self-assessment** beoordeelt de leerling zichzelf aan de hand van een aantal tevoren gegeven criteria. Waarschijnlijk omdat aan de self-assessment techiek veel voordelen voor het onderwijsleerproces worden onderkend, zijn er veel publicaties met praktische aanwijzingen over gebruik van zelfbeoordeling in taalonderwijs. Veel publicaties bevatten

praktische tips over hoe de zelfevaluatie in te zetten op klasniveau. Exemplarisch is een publicatie van Klein (2007) over het dagelijks gebruik van een zelfevaluatie-instrument in het vreemde talenonderwijs. In deze en soortgelijke publicaties (bv. Arter 2003) wordt beschreven hoe zelfevaluatie een plaats krijgt in het onderwijsproces. Zelfevaluatie wordt beschreven als een middel om leerlingen zelf hun leerdoelen te leren definiëren. Daarnaast wordt gerept over het motivatieverhogende effect ervan op leerlingen, het positieve effect op leerrendement en de waarde ervan als instrument om leerlingen zelf het curriculum mede te laten bepalen.

Veelgenoemde doelen van selfassessment zijn:

- Creating awareness of the own communication skills
- Permit the student to be involved in the assessment process
- Provide formative assessment information about strenghts and learning problems
- Assist the student in becoming self-directed in the learning process

Greenan 1985 is een van de eerste empirische studies op het gebied van zelfassesment. Hij wilde onderzoeken of studenten hun eigen communicatieve vaardigheden konden inschatten. De gebruikte instrumenten bleken alle betrouwbaar en consistent te meten: consistency alfa was .93 en de test-retest reliability (Pearson) was .81. De correlatie tussen zelfinschatting en de gebruikte performance test was .42. Ondanks deze lage correlatie benadrukt Greenan dat de zelfbeoordelingen nuttige informatie verschaffen en dat er geïnvesteerd moet worden in begeleiding zodat de zelfinschatting de in de loop der tijd consistent kan worden met de feitelijke performance.
Op basis van een meta-analyse van self-assessment studies onder tweede taalleerders constateert Ross (1998) zeer uiteenlopende, maar vooral tegenvallende correlaties bij selfassessment taken.

Reinders & Lazaro (2007) verrichtten onderzoek in 46 self-access centres in 5 landen (Germany, Hong Kong, New Zealand, Spain, and Switzerland) om de assessment praktijk in beeld te brengen. Hun belangrijkste bevinding was dat meer dan de helft van de instellingen helemaal niets deden aan assessment. De overige instellingen hadden een grote variëteit aan assessment instrumenten in gebruik, waarbij de self-assessment de meest gebruikte soort is. Prapphal (2008) signaleert dat in Thailand de computer wordt ingezet voor self-assessment en niet voor het computerbased testen. Onderzoek van Dlaska & Krekeler (2008) laat zien dat volwassen leerders van Duits als tweede taal grote moeite hebben om hun eigen uitspraakvaardigheid te beoordelen: ze identificeerden nauwelijks de helft van de klanken die volgens ervaren beoordelaars problematisch waren: de fonologische regels van de moedertaal veroorzaken daarbij de meeste problemen. Ook was er op individueel niveau sprake van overschatting en onderschatting. Topping (2003) vat resultaten van eerdere vergelijkingen tussen peer assessment en selfassessment als volgt samen: “peer assessment seems likely to correlate more highly with professional assessment than does self-assessment”. Op basis hiervan kan samenvattend kan gesteld worden dat self-assessment een onvoldoende betrouwbaar is waar het gaat om assessment of learning.

Een interessante ontwikkeling op het gebied van self assessment is beschreven in Alderson & Huhta (2005): in het internationale Dialang-project is voor 14 europese talen een online diagnostisch taalassessment systeem geconstrueerd. In dit systeem is voor 5 taaldomeinen (reading, listening, writing, vocabulary and Structures) gebruik gemaakt van zowel traditionele test items als self assessment statements (de can do-statements op basis van

het Europese Referentiekader). Het systeem is adaptief: de resultaten van de selfassessment bepalen te zamen met de resultaten op een woordenschattest de moeilijkheidsgraad van de overige testonderdelen. De opgaven - per onderdeel ongeveer 50 opgaven - zijn gecalibreerd op basis van data uit een pilot test. Na afloop rapporteert het systeem het behaalde niveau in termen van het ERK en het geeft feedback in de vorm van een overzicht van goede en foutieve antwoorden.

Vergelijking van test scores en selfassessments levert correlaties op die variëren tussen .47 en .58 "between overall self-assessment and performance on a test of that skill". De hoogste correlatie is gevonden bij onderdeel schrijven, wellicht omdat bij die vaardigheid de meeste taal wordt geproduceerd die voor reflectie in aanmerking komt. Bedacht moet worden dat hiermee nog geen 35% van de gevonden variantie in de data kan worden verklaard.

In Nederland hebben Neuvel et al. (2004) op het mbo onderzoek gedaan naar de relatie tussen zelfbeoordeling en taalvaardigheid. Ze hebben 163 leerlingen op niveau 1 en 2 en 182 leerlingen op niveau 3 en 4, afkomstig van verschillende opleidingen van vijf verschillende ROC's, gevraagd om twee taken uit te voeren:

- beoordeling van de eigen taalvaardigheid: de deelnemers kregen een lijst van taaltaken voorgelegd en moesten zelf beoordelen in welke mate zij deze taken konden volbrengen. De lijst van taaltaken was gebaseerd op het Europese Raamwerk voor moderne talen. Leerlingen op niveau 1 en 2 kregen een lijst met taaltaken op de niveaus A2, B1 en B2. Leerlingen op niveau 3 en 4 op de niveaus B1, B2 en C1.
- toetsing van de leesvaardigheid: De deelnemers op niveau 1 en 2 maakten aan de computer de adaptieve Nedcat-toets van het CITO voor leesvaardigheid. De leerlingen op niveau 3 en 4 moesten een samenvatting maken van een tekst op Mavo c/d-niveau (ongeveer B2).

Tevoren was de 47 deelnemende docenten gevraagd om de taalvaardigheid van hun leerlingen in te schatten. Respectievelijk rond 80 en 90 procent van de docenten was van mening dat hun leerlingen het schrijven en lezen op niveau A2 of ergens tussen niveau A2 en B1 beheersten. De deelnemers oordeelden aanzienlijk positiever over hun eigen taalvaardigheid dan zou moeten volgens de toetsresultaten. Meer dan tweederde van de leerlingen op niveau 1 en 2 dacht zelfs de taaltaken op niveau B2 voldoende tot goed uit te kunnen voeren; de toetsresultaten lieten zien dat 83% van de deelnemers onder dat niveau zat. Een soortgelijk beeld werd gevonden bij de leerlingen op niveau 3 en 4: gemiddeld genomen dachten de leerlingen dat hun leesvaardigheid tegen niveau C1 aan lag, terwijl afgaande op de oordelen van docenten (zie hierboven) het feitelijke niveau gemiddeld zelfs onder niveau B1 uitkwam. De toetsresultaten kwamen meer overeen met de schatting van de docenten dan met de zelfbeoordeling van de leerlingen. Opmerkelijk was het dat leerlingen op niveau 1 en 2 hun taalvaardigheid soms nog hoger aansloegen dan leerlingen op niveau 3 en 4.

Ook Boers (2008) rapporteert resultaten die pessimistisch stemmen over de zelfbeoordeling als toetsinstrument: ze vergeleek bij NT2-leerders toetsresultaten met zelfinschattingen door de cursisten en inschattingen door de docent. Er waren 21 proefpersonen, allen hoogopgeleide volwassen NT2-leerders. Er werd gebruik gemaakt van onderdelen van de gestandaardiseerde NIVOR-toetsen NT2 en van de zelfevaluatie-instrumenten die horen bij Raamwerk NT2 voor de vaardigheden Lezen en Luisteren; alle instrumenten en

beoordelingen waren gebaseerd op de niveau-indeling van het ERK. De resultaten laten zien, dat zowel de cursisten als de docenten voor het merendeel van de proefpersonen op beide taalvaardigheden het bereikte niveau te hoog inschatten: voor lezen schatten 12 cursisten zichzelf te hoog in, 2 schatten zichzelf goed of te laag in en van 7 cursisten was dit niet goed te bepalen. Voor luisteren lagen de resultaten nauwelijks anders (resp. 13, 1 en 7 proefpersonen). Docenten deden het iets beter maar niet significant beter. De resultaten uit het onderzoek van Boers zijn des te opmerkelijker, omdat de gebruikte niveau-indeling weinig precies is: het bereik vanaf nulniveau tot niveau *native speaker* omvat zes stappen, dus er wordt zeer grof gemeten…

Op basis van bovenstaande kan worden geconcludeerd dat selfassessment bij gebrek aan betrouwbaarheid geen optie is waar het gaat om summatieve toetsing. Wel biedt het mogelijkheden voor formatieve assessment en met name voor het verdere inrichten van vervolginstructie: zo benadrukken Black & Wiliam (2006c, 93) het belang van externalisatie van reflectie op eigen kunnen (vgl. ook Kavaliauskiene et al. 2007). Zeker wanneer de afstand tussen bereikt resultaat en de gestelde onderwijsdoelen in onderwijssettings expliciet aan de orde wordt gesteld bij de leerlingen, kan dit leiden tot begripsvergroting.

16 De **simulatie** is een assessment-vorm die rond de jaren ’90 ontstaat is dankzij de voortdurend verbeterde prestaties van de comupter. Hoewel de term ook wel gebruikt wordt om bv. het traditionele rollenspel aan te duiden, bedoelen wij met simulatie werkomstandigheden die nagebootst zijn gebruikmaking van de computer. Geluid en beeld suggereren een beroepsmatige kritische situatie waarin de leerling moet handelen of moet aangeven hoe die zou handelen. In de meer geavanceerde apparatuur heeft dit handelen direct effect op de sutuatie die wordt weergegeven, zodat er van echte interactie sprake is. Er zijn met name simulaties gebouwd voor beroepen waarin het lastig, tijdrovend of ethisch niet verantwoord is om de kritische situaties daadwerkelijk te forceren – denk aan het opzettelijk creëren van een gevaarlijke situatie op de snelweg om na te gaan of een kandidaat-wegbeheerder weet hoe in zo’n geval op te treden (vgl. Van Gelooven & Veldkamp 2006). Voor taaltoetsing zijn geen simulaties voorhanden, de situaties waarin taalgedrag te zien gegeven moet worden, worden doorgaans nagespeeld met behulp van leerlingen.

17 De **stage-opdracht en bijbehorend verslag** is een veelgebruikt assessment-instrument. Basis is een opdracht die de leerling uitvoert voor en bij een instelling of bedrijf, met een eindverslag en eventueel een eindproduct als resultaat. Taalgebruik vormt een integraal, maar niet zelfstandig onderscheiden deel van de opdracht. Er zijn geen landelijke criteria afgesproken waarop het verslag wordt beoordeeld.

18 De **vaardigheidstoets** is een toets die controleert of de leerling bepaalde (beroepsmatige) vaardigheden correct en adequaat kan uitvoeren. Met enige regelmaat wordt van taalvaardigheidstoetsen gebruik gemaakt in vmbo en mbo (zie bijlage 1 voor een korte beschrijving van veelgebruikte toetsen). Overigens wordt met name gebruik gemaakt van door leerkrachten zelf gemaakte toetsen (H. Hacquebord, persoonlijke communicatie); over validiteit, betrouwbaarheid en normering van deze instrumenten bestaan geen gegevens.

19 De **voortgangstoets** is bedoeld om te meten of een leerling voldoende vooruitgaat. De inhoud kan betrekking hebben op kennis of op vaardigheden. Naast enkele gepubliceerde taaltoetsen (zie Bijlage 1) zijn in het vmbo de huiswerkcontrole, de schriftelijke overhoring en de methodegebonden toets de meest voorkomende manieren om voortgang te meten (bron: P. Litjens, pers. communicatie; vergelijk Inspectie van het Onderwijs 2007; Visscher 2008; vergelijk ook Lee 2007 die opmerkt dat leerkrachten uit assessments geen consequenties trekken voor het verdere verloop van het onderwijs). Over validiteit noch betrouwbaarheid zijn gegevens bekend.

20 Er zijn geen aanwijzingen gevonden dat de **computergestuurde toets** veelgebruikt is in vmbo en mbo. We besteden er hier wel enige aandacht aan omdat op basis van het voorspelde lerarentekort verwacht mag worden dat dit beeld in de toekomst ingrijpend zou kunnen gaan veranderen. Computerbased testing is een toetssoort die vanaf 1990 in opkomst is. Belangrijkste voordelen ervan zijn (vgl. o.a. Eggen 2009; Eggen et al 1996):

- afname en beoordeling zijn arbeidsextensief;
- afname is tijd- en ruimte-onafhankelijk;
- naast taal kan van andere multimedia-mogelijkheden gebruik gemaakt worden, inclusief simulaties van beroepssituaties;
- scores en deelscores voor diagnostische doelen zijn direct beschikbaar;
- ingeval van computergestuurd *adaptief* toetsen zijn er beduidend minder opgaven nodig om tot een resultaat met eenzelfde betrouwbaarheid te komen.

Daarnaast is het mogelijk om aan een adaptieve assessment direct een remediërend programma te koppelen (vgl. bv. Schuurs & Verschoor 2004).

Voor **computergestuurd adaptief toetsen** is in het vmbo en mbo vooralsnog weinig belangstelling te constateren. Exemplarisch daarvoor is wellicht een oproep op de mailing list van De Digitale School (www.digitaleschool.nl) onder de onderwerpsregel “Werken met een ELO op het vmbo” op 27 januari. Een leerkracht vraagt om tips van collega’s, want ze gaat werken als docent Nederlands aan een nieuwe school: “Momenteel geef ik alleen les aan vmbo tl en ik heb helemaal geen ervaring met basis, kader en gemengde leerweg. Daarbij komt dat we met een ELO gaan werken, waar ik ook geen ervaring mee heb.” Op haar oproep is geen enkele reactie gekomen…

De geringe belangstelling is opmerkelijk omdat de voordelen van computergestuurd toetsen evident zijn: uit de testtheorie is bekend dat de betrouwbaarheid van een beoordeling toeneemt naarmate het aantal waarnemingen groter is (vgl. bv. Eggen & Sanders 1993; vgl. ook Van den Brink & Mellenbergh 1998. Naarmate de moeilijkheidsgraad van de opgaven echter beter aansluit op het daadwerkelijke niveau dat de leerling op dat moment heeft, wordt de betrouwbaarheid van de meting groter en zijn er relatief gezien minder waarnemingen - dus minder opgaven - nodig. In computergestuurde adaptieve toetsen wordt na elke opgave een schatting gemaakt van het vaardigheidsniveau van de leerling en wordt een dicht bij dat niveau passende opgave geselecteerd. Onderzoek wijst uit dat met een computergestuurde adaptieve toets maar ongeveer de helft van het aantal opgaven nodig is om met dezelfde mate van nauwkeurigheid uitspraken over de kandidaat te kunnen doen (zie Eggen et al. 1996). Deze winst in efficiëntie is vooral groot bij extreem laag vaardige en extreem hoog vaardige leerlingen (vgl. Eggen 2009; Wainer 2000. Met name voor kennistoetsen is de computergestuurde adapatieve toetsvorm dan ook heel geschikt,

maar inmiddels bestaan er ook de technische mogelijkheden om adaptieve algoritmen toe te passen in simulaties (vgl. Roelofs & Straetmans 2006).

2.4 Een voorlopige kijk op assessment

Aan dit onderzoek liggen drie vragen ten grondslag. De eerste onderzoeksvraag luidde: Welke toetsvormen zijn geschikt om zowel de *assessment of learning* functie te vervullen als de *assessment for learning* functie? Op basis van de beschrijving van assessmentvormen in dit hoofdstuk kan geconstateerd worden dat een groot aantal assessmentvormen in principe bruikbaar is voor zowel *assessment of learning* als voor de *assessment for learning*. Aan elke vorm van assessment kunnen principieel steeds beide toetsfuncties worden toegedicht: tot welke soort een assessment moet worden gerekend, is afhankelijk van de intentie om op basis van de resultaten het onderwijs bij te stellen. Hoe algemener de toetsdoelen van een bepaalde assessment echter zijn geformuleerd, hoe lastiger het is om op basis van de resultaten ervan zinnige en gerichte feedback te formuleren om het verdere onderwijs te richten.

De vraag is dan niet of een bepaalde assessmentvorm beide functies kan vervullen, want dat is principieel wel mogelijk: de vraag is eerder of aan bepaalde vormen van assessment die kenmerken kunnen worden meegegeven die wenselijk resp. noodzakelijk zijn in de situatie waarin getoetst wordt en de doelen waarmee getoetst wordt. Wiliam & Black (1996) schetsen het beeld waarbij summatieve en formatieve assessment-functies worden gekarakteriseerd als de uiteinden van een continuüm waarop alle assessments kunnen worden gelokaliseerd. In die optiek worden de extremen van dat continuüm gevormd door

- highstake assessments met een certificerende functie: aan dit soort assessment moeten de hoogste eisen ten aanzien van betrouwbaarheid worden gesteld. Een verkeerde beslissing kan er immers toe leiden dat iemand ten onrechte een diploma wordt onthouden of dat ten onrechte een formele toekenning van bewezen competentie wordt verstrekt.
- Assessments met een diagnostische functie: dit soort assessment heeft een lokaal werkingsgebied en een verkeerde beslissing heeft minder verstrekkende gevolgen. Dit is zeker het geval “if the teacher can detect and correct for it in continuing interaction with the learner” (in de woorden van Black & Wiliam 2006b, 129).

Deze voorbeelden maken duidelijk dat de betrouwbaarheid van een toets in het ene geval van groter belang is dan in het andere geval. Waar het de validiteit betreft: bij een diagnostische toets - bij uitstek assessment for learning – is het doorgaans gemakkelijker om de inhoud van de toets dichtbij de inhoud van het geboden onderwijs te brengen dan bij assessments met een groter inhoudelijk meetbereik zoals een afsluitend examen.

Het is zinnig om te inventariseren aan welke eisen een assesment zou moeten voldoen gezien de functie die moet worden vervuld. Hierover gaat het volgende hoofdstuk.

Hoofdstuk 3: De validiteit van assessment

In hoofdstuk 2 is geconstateerd dat er uiteenlopende eisen aan assessment-instrumenten kunnen worden gesteld. Hoewel het lastig is om hierover in algemene zin te rapporteren, is het mogelijk om een aantal belangrijke eisen uit de literatuur te filteren en uit te diepen. In paragraaf 3.1 bespreken we een aantal kenmerken van toetsen zoals die in de literatuur aan de orde zijn gekomen; op basis hiervan is de tweede onderzoeksvraag te beantwoorden. In paragraaf 3.2 brengen we de genoemde kenmerken van assessment-instrumenten in relatie met de functie die de toets moet vervullen. Op basis hiervan kan een antwoord op de derde onderzoeksvraag worden geformuleerd.

3.1.1 Kwaliteitseisen aan assessment

Bij een assessment worden twee karakteristieken steevast van groot belang geacht: de betrouwbaarheid en de validiteit. De betrouwbaarheid zegt iets over de mate waarin gemeten waarden overeenkomen met de feitelijke waarden. Om betrouwbaar genoemd te kunnen worden, moet een assessment nauwkeurig meten en bovendien bij herhaaldelijk meten tot hetzelfde resultaat komen. Een assessment met lage betrouwbaarheid laat grote verschillen in scores zien als herhaaldelijk hetzelfde wordt gemeten. Naarmate een assessment betrouwbaarder is, zijn kleinere verschillen in scores nog betekenisvol en op een zinnige manier te interpreteren.

De betrouwbaarheid is de belangrijkste eis die aan een assessment te stellen is (vgl. Haertel 2006). Als een assessment niet voldoende betrouwbaar is, heeft het weinig zin om de validiteit aan de orde te stellen. De validiteit van een assessment is de mate waarin het instrument meet wat gemeten moet worden. Waar betrouwbaarheid een objectief gegeven is, moet de validiteit van een assessment toets worden beredeneerd: aangetoond moet worden dat het beoogde construct wordt gemeten, maar dat construct zelf kan alleen maar indirect zichtbaar worden gemaakt via metingen. In een retrospectief overzicht merkt Spolsky (2008, 448) op: “Once statistical methods of establishing reliability were found, replacing single individual measures with large numbers of items lending themselves to appropriate statistical treatment, testers could argue that their test was reliable: in other words, that it would have much the same result when repeated on other occasions or other candidates. More difficult has been agreement on the validity, essentially the meaning rather than the stability of the result.”

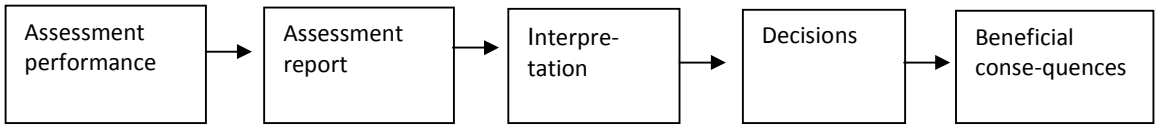
Door de interpretatieve verschillen die in de loop der tijd optreden ten aanzien van het te meten construct, kan de functie van een toets veranderen. Van den Bergh (2007) illustreert dit aan het doel van de samenvatting: in de jaren ’50 van de vorige eeuw was de samenvatting bedoeld als toets voor schrijfvaardigheid. De samen te vatten brontekst was een manier om verschillen in voorkennis over het onderwerp bij de leerlingen te minimaliseren. Later is men de samenvatting gaan beschouwen als een toets voor leesvaardigheid. De opvatting over welk construct gemeten wordt, is kennelijk gaandeweg veranderd.

Het lijkt erop dat veel vormen van alternative assessment ontwikkeld zijn vanuit onvrede met de validiteit van de toetsen die op dat moment in gebruik waren. Het beeld dringt zich op als zouden gestandaardiseerde toetsen met een acceptabele betrouwbaarheid niet goed passen bij het gegeven onderwijs en om die reden zijn verruild voor toetsen die

naar het oordeel van leerkrachten wel valide zijn maar waarvan de betrouwbaarheid ontoereikend of onbekend is.

Met name sinds de opkomst van de *assessment of learning* worden er naast betrouwbaarheid en validiteit een groot aantal andere eisen genoemd die men kan stellen aan toetsen. Bachman 2005 gaat terug naar de essentie als hij opmerkt: *The primary use of all language assessments is to gather information to help us make decisions that will lead to beneficial consequences for stake holders*. Bachman schetst de keten vanaf testafname tot en met de consequenties van de beslissing op basis van de testresultaten als volgt.

Figuur 3.1: Assessment Design and Development (bron: Bachman 2005)



Omdat beslissingen op basis van testresultaten vaak high stake zijn en foutieve beslissingen vaak lastig terug te draaien zijn als ze al ontdekt worden, moeten we in staat zijn om het testgebruik te rechtvaardigen en belanghebbenden ervan kunnen overtuigen dat het beoogde gebruik van de test gerechtvaardigd is. In de door Bachman voorgestelde Assessment Design and Development moet elke stap uit figuur 1 achteraf in omgekeerde volgorde te verantwoorden zijn. Bachman bepleit daarom een uitgebreide Assessment Use Argument, een 'toetsgebruikargumentatie' die is opgezet volgens het argumentatiemodel van Toulmin (inclusief data, een claim, een warrant en een backing om de warant te rechtvaardigen; liefst bevat de argumentatie ook een rebuttal waarin de uitzonderingen worden genoemd).
Deze argumentatie kan alleen zuiver zijn wanneer de geschetste keten voldoet aan de volgende voorwaarden:

- de report en de scores zijn consistent;
- de interpretaties zijn betekenisvol, zijn gebaseerd op de toets als geheel, zijn generaliseerbaar, relevant en toereikend;
- de beslissingen moeten billijk zijn;
- de consequenties moeten bruikbaar zijn.

Over alle door Bachman genoemde aspecten zijn in de literatuur wensen geventileerd. In het vervolg van deze paragraaf inventariseren we veelgenoemde desiderata.

Segers et al. (2003) benoemen zes dimensies waarlangs de veranderingen in assessment in de laatste twee decennia zich bewegen. Deze dimensies zijn in latere publicaties geoperationaliseerd in eisen waaraan authentieke assessments zoveel mogelijk zouden moeten voldoen. De zes dimensies zijn:

1 *Mate van authenticiteit*: er is een verandering vanaf contextloos toetsen op afzonderlijke eenheden uit he onderwijsaanbod naar authentieke assessments in betekenisvolle context. Dit is een direct gevolg van het streven naar leren in authentieke contexten. Binnen het authentieke testen wordt minder gebruik gemaakt van objectieve tests met itemsoorten zoals het korte antwoord, de *fill in the blank*, de meerkeuze-opgave en de goed/fout-opgave. Een meer voor de hand liggende vorm is de performance assessment waarbij de leerling

gedrag vertoont in een semi-authentieke situatie; dat gedrag wordt op een aantal criteria beoordeeld.
Vanuit een functioneel perspectief beoogt men een reële afspiegeling in de toets van wat de kandidaat in de reële beroepssituatie moet kunnen. Met name in het mbo blijven de assessmenttaken qua opzet zo dicht mogelijk bij de werkelijkheid. Aangezien in werkelijkheid vaardigheden ook samen voorkomen, worden in één assessmenttaak meerdere vaardigheden en taalaspecten geïntegreerd getoetst.

2 *Het aantal beoordeelde aspecten*: Er is een tendens om leerlingen niet meer op één aspect of met één cijfer te beoordelen, maar om de uitvoering van een complexe cognitieve taak te beoordelen op een groter aantal aspecten. De beoordeling van de prestaties van de leerlingen op authentieke opdrachten moet liever niet integratief plaatsvinden: die beoordeling wint aan betrouwbaarheid en validiteit wanneer ze achtereenvolgens betrekking heeft op afzonderlijke deelaspecten (vgl. bv. Clayton et al. 2003; Straetmans 2006; Van den Bergh 2007).

3 In plaats van lagere orde-vaardigheden worden de zogenaamde *higher-order skills* getoetst. Waar vroeger de nadruk lijkt te liggen op toetsing en reproductie van kennis, wordt nu een assessment uitgevoerd op de beheersing van complexe vaardigheden (vgl. Hartley & Wentling 1996; Bharosa et al. 2008).

4 Vanwege het streven naar een integrale toetsing van gedrag komen er niet meer alleen cognitieve aspecten voor toetsing in aanmerking, *ook metacognitieve en sociale en psychomotorische aspecten* (kunnen) worden (mee-)beoordeeld.

5 Stond vroeger de assessment relatief los van het onderwijs, tegenwoordig is er in toenemende mate sprake van *integratie van instructie, leren en toetsing*: assessment wordt gezien als “a tool for dynamic ongoing learning” (vgl. Black & Wiliam 1998b).

6 Terwijl vroeger de leerkracht verantwoordelijk was voor zowel het leerproces als de toetsing, heeft nu de leerling in toenemende mate *zelfverantwoordelijkheid* in het assessment proces. Deze zelfverantwoordelijkheid kan betrekking hebben op zowel de condities waaronder de toets wordt afgenomen als de beoordelingscriteria die worden gehanteerd. Er is reden om te veronderstellen dat het expliciet benoemen van beoordelingscriteria vooraf het leerproces faciliteert (Black & Wiliam 1998a; Leahy et al. 2005; Matsuno 2009).

Dysthe et al. 2008 laten zien dat masterstudenten het gebruik van expliciete criteria juist als beperkend ervaren en de voorkeur geven aan de mogelijkheid om te onderhandelen over de normen waaraan hun prestaties moeten voldoen. Het lijkt erop dat het competence niveau en persoonsvariabelen zoals leerstijl mede bepalen in hoeverre het expliciet benoemen van beoordelingscriteria faciliterend werkt.

Bachman & Palmer (1996) introduceerden het begrip bruikbaarheid van een test. Deze *usefullness* hangt in hun optiek af van zes karakteristieken van een toets:

1. *Reliability*: de betrouwbaarheid kan worden gedefinieerd in termen van consistentie van testresultaten over verschillende versies van de test en verschillende beoordelaars.
2. *Construct validity*: de constructvaliditeit betreft de mate van nauwkeurigheid van inferenties over het construct dat men wil meten en het doel van de test.
3. *Authenticity*: Bachman and Palmer (1996: 23) definiëren authenticiteit als "the degree of correspondence of the characteristics of a given language task to the features of target language use task". Een test moet in hun ogen authentiek zijn omdat de testresultaten geïnterpreteerd moeten worden in relatie tot het doel van de test: het functioneren buiten de testomgeving.
4. *Interactivity*: met interactiviteit wordt de *personal involvement* van de deelnemer in de taakstelling bedoeld. Als een test door de kandidaat wordt geïnterpreteerd als een kunstmatige en onechte taakomgeving, hebben de resultaten geen zeggingskracht (vgl. Straetmans 2006, 27).
5. *Impact*: dit kenmerk is gerelateerd aan de *consequential validity* (vgl. Messick 1989; Gielen et al. 2003). De vorm en inhoud van een test hebben invloed op de didactiek en inhoud van het onderwijs en op de leerstrategieën die de leerders gebruiken. Positieve impact is een belangrijk kenmerk van authentic assessment (vgl. Wiggins 1989; Gulikers et al. 2006a; Sluijsmans et al. 2008).
6. *Practicality*: een assessmentafname moet praktisch zijn en inpasbaar in de dagelijkse gang van zaken. Het streven naar een hoge betrouwbaarheid kan ervoor zorgen dat de assessmentafname te veel tijd vergt en daardoor onbruikbaar wordt (vgl. Meuffels & Maat 2009). Kosten en baten moeten in een acceptabele verhouding tot elkaar staan: practicality is zeker een kenmerk dat met gebruikmaking van de computer kan worden verbeterd, zowel door adaptieve algoritmes te gebruiken (vgl. Eggen 2009) als door simulatie in de testomgeving te brengen (Roelofs & Straetmans 2006).

Booth et al. 2003 geeft een literatuuroverzicht van criteria waaraan online toetsing zou moeten voldoen. Genoemd worden onder andere:

1. *variety:* including both quantitative and qualitative methods
2. *authenticity:* using open-ended tasks that simulate workplace tasks, as well as appropriate quantitative tasks
3. *collaboration:* allowing for interaction between learners and others, and using appropriate communication technologies
4. *feedback:* ensuring appropriate feedback mechanisms are possible using peer feedback and peer tutoring
5. *online resources:* making full use of available quantitative packages as well as other internet resources
6. *learner responsibility:* providing options and opportunities for accountability within assessment tasks

Rowlands (2001, 54) bepleit het gebruik van onderstaande criteria bij het evalueren van online assessment:

1. Are assessments authentic, based on real life applications?
2. Are assessment items flexible, and are multiple forms of assessment possible?
3. Are students allowed to present evidence of knowledge and skill that is meaningful to them and unique to their learning preferences?

4. Is the assessment introduced before or simultaneously with content material?
5. Is assessment continuous?
6. Is self-assessment or peer assessment available?

Ten onzent noemt Klarus (2000) de volgende karakteristieken van een model voor het beoordelen van competenties:

- ***Oordelende*** (judgemental) benadering van het waarderen van competentiebewijzen.
- ***Authentiek*** beoordelen: work samples of praktijkopdrachten. Reële beroepssituatie.
- ***Congruent*** beoordelen: handelingsvolgorde in de beroepssituatie verloopt via planning, uitvoering, evaluatie/bijstelling. In de beoordelingsprocedure dient dezelfde volgorde aangehouden te worden.
- ***Geïntegreerd*** beoordelen: theorie en praktijk, kennis door toepassing van kennis. Competenties verschijnen als handelingen waarin declaratieve, procedurele en conditionele kennis zijn geïntegreerd.
- ***Criterium-gerelateerd*** beoordelen: criteria worden ontleend aan een vooraf gestelde kwalificatiestandaard en niet zoals bij normgerelateerde beoordeling met de scores van andere kandidaten.
- ***Leerwegonafhankelijk***: de criteria zijn niet gerelateerd aan een bepaald leertraject, maar gebaseerd op competentie-eisen zoals de beroepspraktijk die stelt.

Baartman (2008a, 2008b) richt zich in haar onderzoek op het geheel aan beoordelingsvormen waarmee opleidingen hun programma van toetsing en afsluiting invullen. Ze heeft een raamwerk van twaalf kwaliteitscriteria ontwikkeld voor competentiegerichte assessmentprogramma's waarin zowel 'oude, beproefde' criteria zijn opgenomen als 'nieuwe' criteria, passend bij de aard van competentiegericht beroepsonderwijs. Dat raamwerk van criteria is zowel theoretisch onderbouwd als gevalideerd op praktijkrelevantie. De criteria zijn, naast betrouwbaarheid en validiteit:

1 authenticiteit (vergelijkbaar met beroepstaken)
2 cognitieve complexiteit (het denkproces moet beoordeelbaar zijn)
3 vergelijkbaarheid met doelen & herhaalbaarheid van beslissingen (consistent scoorbaar)
4 tijd en kosten (opwegen tegen de opbrengsten)
5 onderwijsgevolgen (goed interpreteerbaar in relatie tot onderwijsdoelen en beroep)
6 positief effect op onderwijs (positieve gevolgen voor onderwijsproces)
7 eerlijkheid (voor alle groepen en individuen in beginsel even moeilijk)
8 betekenisvolheid (nuttig in de ogen van de studenten en andere betrokkenen)
9 reproduceerbaar (beoordelaarsonafhankelijk)
10 transparantie & acceptatie (alle betrokkenen moeten de leerdoelen en scoringsregels duidelijk zijn, vgl. Bachman 2005)

Eisen die volgens Kuhlemeijer (2005) aan assessment gesteld worden:

1. *Integraal onderdeel onderwijsleerproces*. Een competentiegerichte toets vormt een integraal onderdeel van het onderwijsleerproces. Instructie, oefening en toetsing vormen een geheel. De toetsing vindt niet meer alleen 'achteraf' plaats, maar wordt ook voorafgaand aan en tijdens het onderwijsleerproces ingezet. Toetsen worden geacht het leerproces in positieve zin te sturen.

2. *Authenticiteit*. De vraag- en probleemstelling van een competentiegerichte toets is realistisch of authentiek. De leertaken hebben een grote overeenkomst met problemen, taken en dilemma's uit het alledaagse leven of de latere beroepspraktijk. Ook de eisen waaraan de leerlingprestatie moet voldoen kunnen authentiek zijn in de zin dat ze gelijkenis vertonen met de criteria die ook gelden voor beginnende beroepsbeoefenaren.
3. *Leergangonafhankelijkheid*. Traditionele toetsing is meestal gebonden aan een specifieke opleiding en een bepaalde periode. Competentiegerichte toetsing is daarentegen vaak leerwegonafhankelijk. De beoordelingscriteria zijn niet gerelateerd aan een bepaald leertraject, maar gebaseerd op competentie-eisen zoals de maatschappelijke of professionele praktijk die stelt.
4. *Samenwerking tussen leerlingen*. Competentiegerichte toetsen worden vaak uitgevoerd door groepjes leerlingen. Leerlingen oefenen daarbij allerlei samenwerkingsvaardigheden, zoals naar elkaar luisteren, met elkaar rekening houden, samen een planning maken en taken verdelen. De kwaliteit van de samenwerking vormt een integraal onderdeel van de beoordeling.
5. *Meer tijd*. Leerlingen krijgen meer tijd om hun competentie aan te tonen. Een gewone toets of proefwerk neemt hooguit één of twee lesuren in beslag. Een competentiegerichte toets kan zich uitstrekken over een periode van meerdere weken tot maanden. Leerlingen krijgen bijvoorbeeld voldoende tijd om hun werk te plannen, uit te voeren, verschillende bronnen te raadplegen, zichzelf te evalueren en hun werk te reviseren.
6. *Zelfreflectie en zelfevaluatie*. Competentiegerichte toetsvormen stimuleren de leerlingen tot zelfreflectie en zelfevaluatie. De leerling oefent metacognitieve vaardigheden zoals het kritisch beschouwen van eigen werk en het herzien van eigen werk op grond van andermans commentaar.
7. *Intersubjectiviteit*. Toetsing is niet langer het alleenrecht van de individuele docent. In de beoordeling worden meer personen en gezichtspunten betrokken. Medebeoordelaars kunnen collegadocenten of deskundigen uit het bedrijfsleven zijn, maar ook leerlingen die een oordeel uitspreken over hun werk.

Tot slot heeft Straetmans (2006) in zijn lectorale rede op basis van zijn eerdere werk samenvattend een aantal eisen geformuleerd waaraan in zijn ogen een goede assessment-procedure moet voldoen. Die eisen komen op het volgende neer: een goede toets

1 heeft een acceptabele graad van betrouwbaarheid (Straetmans 2006, 21)
2 is aantoonbaar valide voor het gestelde doel (ibid, 21)
3 is competentiegericht (ibid,15)
4 reflecteert beroepstaken in inhoud en moeilijkheidsgraad (ibid,15)
5 is representatief voor taken uit het beroep, zowel in volledigheid (comprehensiveness) als natuurgetrouwheid (fidelity) (ibid,22)
6 is representatief voor het criteriumdomein (ibid,23)
7 stuurt interpretaties van individuele assessoren zoveel mogelijk in dezelfde richting (ibid,26).
8 wordt door de kandidaat als realistisch ervaren en niet als kunstmatige situatie. Dit betreft de obtrusive vs unobtrusive observations; vgl. de automobilist die zich bij het rijexamen aan de snelheidslimiet houdt, maar daarna nooit meer (ibid,27).
9 heeft een grensscore voor de zak/slaagbeslissing die criterion-referenced is (ibid, 37).

10 heeft zoveel mogelijk die vorm waarmee gedrag wordt uitgelokt dat maximaal lijkt op het gedrag in het criteriumdomein zijn (ibid,41).
11 past binnen de beschikbare hoeveelheid met de beschikbaarheid van tijd en geld (ibid, 42).
12 lokt gedrag uit dat beoordeeld kan worden met de eerder geformuleerde beoordelingsaspecten (ibid,44).

Ten overvloede wijst Straetmans erop dat tegenvallende prestaties van kandidaten terugvoerbaar kunnen zijn op een gebrek aan benodigde kennis (Straetmans 2006, 46). Om de reden blijft de beschikbaarheid van kennistoetsen belangrijk.

Op basis van bovenstaande beschrijvingen van criteria die aan assessments gesteld kunnen worden, is een lijst van 20 relevante kenmerken samengesteld:

1. De taken in de taaltoets komen overeen met taken in de praktijk.

2. De taken in de taaltoets zijn competentiegericht (kennis, vaardigheid, houding geïntegreerd).

3. De moeilijkheidsgraad van de taaltoets is vergelijkbaar met de moeilijkheidsgraad van taken in de praktijk.

4. De toetsscore geeft een goed beeld van het prestatieniveau in de praktijk.

5. De prestaties op de toets worden vergeleken met van tevoren vastgestelde kwaliteitseisen.

6. Leerlingen zijn op de hoogte van wat er van ze verwacht wordt bij de taaltoets.

7. De toets meet daadwerkelijk wat er gemeten moet worden (validiteit).

8. De beoordeling van prestaties op de toets is beoordelaars-onafhankelijk en/of wordt door meer beoordelaars gedaan.

9. De toets heeft een aangetoonde hoge mate van betrouwbaarheid.

10. De toetsdoelen komen overeen met de onderwijsdoelen.

11. De taaltoets is toegankelijk voor de leerling op het moment dat de leerling eraan toe is.

12. De taaltoets geeft inzicht in het handelen van de leraar en is leerzaam voor de leerling.

13. De toets is leerwegonafhankelijk: het maakt niet uit waar de leerling de competenties heeft verworven – op school, op de werkplek of elders.

14. De kosten en tijd van de afname staan in verhouding tot de opbrengst van de toets.

15. De toets levert een bijdrage aan het leerproces dankzij de feedback aan de leerling.

16. De toets heeft invloed op de inhoud van het verdere onderwijsleerproces.

17. De toets bevordert het vermogen tot zelfbeoordeling bij de leerling.

18. De toetsscores voorspellen voor Nederlands de verdere prestaties van de leerling in het onderwijs.

19 De toetsscores zijn generaliseerbaar en zeggen iets over de prestaties op alle andere taken die voorgelegd hadden kunnen worden.

20. De toetsscores zijn een aantoonbaar goede voorspeller van de prestaties in de beroepspraktijk.

Opvallend is dat in de literatuur bij het bespreken van wenselijke of noodzakelijke kenmerken van assessments doorgaans geen aandacht wordt besteed aan de functie van de betreffende assessment. Hierop zijn twee algemene uitzonderingen te melden:

1) In een aantal publicaties wordt het belang van de validiteit benadrukt en wordt expliciet gemeld dat het validiteitsbegrip impliceert dat de toetsconstructeur rekening houdt met de doelen van de toetsing. Het validiteitsbegrip is in de loop der tijd verruimd en heeft inmiddels minstens vijf betekenissen:
   1 **Indruksvaliditeit** (*face validity*): dit betreft niet meer of minder dan de intuïtieve indruk van de toetsconstructeur of andere betrokkenen of de toets meet wat die zou moeten meten.
   2 **Inhoudsvaliditeit** (*content validity*): dit heeft betrekking op de vraag of de toets een verantwoorde representatie vormt van het te meten kennisdomein.
   3 **Criteriumvaliditeit** (*criterion validity*): dit betreft de vraag of de toetsresultaten voorspellende waarde hebben ten aanzien van hoe de kandidaten zich in de werkelijkheid gaan gedragen (predictieve validiteit) of hoe de kandidaten zouden presteren op vergelijkbare instrumenten die hetzelfde construct proberen te meten (concurrente validiteit).
   4 **Ecologische validiteit** (*ecological validity*): dit betreft de vraag in hoeverre de toetsresultaten overeenkomst vertonen met de alledaagse praktijk. Dit begrip heeft niet betrekking op validiteit als psychometrische eigenschap van een instrument.
   5 **Constructvaliditeit ofwel begripsvaliditeit** (*construct validity*) betreft de vraag of de resultaten samenhang vertonen met resultaten op soortgelijke instrumenten of met observatie (convergente validiteit). Ook kan het begrip betrekking hebben op de vraag of de resultaten systematisch afwijken van resultaten op instrumenten die een ander construct beogen te meten (discriminante validiteit).

   Met de publicatie van Messick (1989) behoren zelfs de beslissingen van derden op basis van toetsresultaten tot het validiteitsbegrip. Borsboom et al. (2004) bepleiten om het begrip validiteit terug te brengen tot die processen die de verschillen in prestaties op een toetsinstrument kunnen verklaren.

2) In een aantal publicaties staat de interactie tussen instructie, leren en toetsing centraal (vgl. Biggs 1996; Black & Wiliam 1998a; Gielen et al. 2003; Black & Wiliam 2006c). In deze en soortgelijke publicaties staat steeds de formative assessment centraal.

De bevindingen uit de literatuur zijn in tabel 3.2. samengevat. De cijfers in de tabel verwijzen naar de volgorde waarin per auteur de wensen genoemd zijn in voorgaand overzicht.

Tabel 3.2: Wenselijke kenmerken van assessment-instrumenten

	Segers ea 2003	Bachman Palmer 1996	Booth ea 2003	Rowlands 2001	Klarus 2000	Baartman 2008a	Kuhle-meijer ea 2005	Straet-mans 2006
1. De taken in de taaltoets komen overeen met taken in de praktijk	1	3	2	1	2	1	2, 5	4
2. De taken zijn competentiegericht en geïntegreerd	3		1		4	2		3
3. De taken in de taaltoets en taken in de praktijk zijn vergelijkbaar moeilijk	1	3	2	1	2	1	2	4
4. De toetsscore geeft een goed beeld van het prestatieniveau in de praktijk	1	3	2	1	2	1	2	4
5. De toets is criterion-referenced (tevoren vastgestelde kwaliteitseisen)					5			9
6. Leerlingen zijn op de hoogte van wat er van ze verwacht wordt						10		
7. De toets meet daadwerkelijk wat er gemeten moet worden (validiteit)		2				2	2	2
8. Beoordeling is beoordelaaronafhankelijk / gebeurt door meer beoordelaars						9	7	7
9. De toets heeft een aangetoonde hoge mate van betrouwbaarheid		1						1
10. De toetsdoelen komen overeen met de onderwijsdoelen					3	3	(1)	5
11. De taaltoets is toegankelijk voor de leerling op een zelfgekozen moment	6			3				
12. De toets geeft inzicht aan de leerkracht en is leerzaam voor de leerling	5					6		
13. De toets is leerwegonafhankelijk					6		3	
14. De kosten en tijd staan in verhouding tot de opbrengst van de toets		6				4		11
15. De toets draagt door feedback bij aan het leerproces			4			6	(1)	
16. De toets heeft invloed op de inhoud van het verdere onderwijsleerproces		5				5	(1)	
17. De toets bevordert het vermogen tot zelfbeoordeling bij de leerling							6	
18. De toetsscores hebben predictieve validiteit binnen het onderwijs							(1)	5
19 De toetsscores zijn generaliseerbaar naar het taakdomein						(8)		5
20. De toetsscores hebben predictieve validiteit voor de beroepspraktijk	1	3	2	1	2	1	2	6

De gegevens in tabel 3.2 bieden indirect een antwoord op de tweede onderzoeksvraag. Die vraag luidde: Op welke manieren kunnen de opbrengsten van taalonderwijs in het (v)mbo op een gestandaardiseerde manier worden gemeten? Uit tabel 3.2 kan worden opgemaakt dat er in de literatuur kennelijk eenstemmigheid heerst over vier kenmerken waaraan een assessment zou moeten voldoen:

- De taken in de taaltoets komen overeen met taken in de praktijk
- De taken in de taaltoets en taken in de praktijk zijn vergelijkbaar moeilijk
- De toetsscore geeft een goed beeld van het prestatieniveau in de praktijk
- De toetsscores hebben predictieve validiteit voor de beroepspraktijk

Opvallend in dit verband is veel auteurs niet benadrukken dat een assessment valide en/of betrouwbaar zou hoeven zijn. Probleem hierbij is, dat wanneer een auteur een bepaald kenmerk niet genoemd heeft, dit niet wil zeggen dat er geen waarde aan het betreffende kenmerk wordt gehecht. Intussen wijst de grote heterogeniteit bij het benoemen van al dan niet wenselijke kenmerken van assessments er wel op, dat een verregaande standaardisering in de betekenis van uniformering nauwelijks mogelijk is: de experts zijn het er niet over eens aan welke eisen voldaan moet worden.

In de tweede onderzoeksvraag kan het begrip 'gestandaardiseerd' worden opgevat in de betekenis van: met centraal landelijke toetsen waarvan validiteit en betrouwbaarheid tevoren zijn bepaald. Wanneer dat wordt bedoeld, kan het antwoord worden gevonden in de landelijke examens die weer zullen worden ingevoerd voor het mbo en in de overige talige eisen die vanaf 2012 worden gesteld aan deelnemers aan het mbo, onder verwijzing naar het raamwerk dat Commissie Meijerink heeft geformuleerd (zie Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008).

Wanneer met het begrip 'gestandaardiseerd' ook wordt gedoeld op enigerlei uniformering in de vele assessment-instrumenten die momenteel worden ingezet voor *assessment for learning*, is het antwoord minder eenvoudig. In dat geval vereist standaardisering in elk geval aanpassing van de assessmentinstrumenten in twee opzichten:

- elk instrument - al dan niet op school ontwikkeld - dat bedoeld is voor *assessment of learning* zou duidelijk moeten verwijzen naar een algemeen aanvaard en herkenbaar stelsel van taalniveaus. Enige tijd leek het Europese Raamwerk en daarvan afgeleide indelingen (zie bv. Bohnenn et al. 2007) daarvoor het meest aangewezen stelsel; sinds 2008 ligt het voor de hand om het raamwerk van Commissie Meijerink (Expertgroep Doorlopende Leerlijnen Taal en Rekenen 2008) als referentiepunt te nemen.
- een praktisch probleem bij deze raamwerken is dat ze weliswaar volop aanknopingspunten bieden om assessment-opdrachten te ontwikkelen, maar ze bieden vooralsnog weinig houvast waar het gaat om het beoordelen van de manier waarop assessment-taken worden uitgevoerd. Ook de beoordelingsprocedures zouden moeten worden gestandaardiseerd. Met name bij *assessment for learning*-assessments is dat bijna onbegonnen werk.

Samenvattend: voor de *assessment of learning* kan het onderwijsrendement het beste worden gemeten met behulp van landelijk uitgerolde, gestandaardiseerde toetsen die alle verwijzen naar één onderliggende meetlat, bijvoorbeeld door gebruik te maken van equivaleringsprocedures met behulp van Item Response Theorie. Om te voorkomen dat er per opleiding andere assessments komen - dat zou leiden tot hoge kosten en waarschijnlijk

tot een klein aantal gegevens om de equivalering deugdelijk uit te voeren – kan het beste gezocht worden naar enkele assessmentvormen met generiek bruikbare inhouden.

Voor de instrumenten ten behoeve van de *assessment for learning* is met name de relatie tussen inhoud van assessment en inhoud van onderwijs van belang. Deze instrumenten zijn ervoor bedoeld om invloed te hebben op de verdere inhoud van het onderwijs, om door middel van feedback bij te dragen aan het leerproces. Als daar het primaat ligt, wordt het praktisch onhaalbaar om deze instrumenten naar niveau te equivaleren en te standaardiseren: de toetsresultaten moeten dan immers vertaald worden naar het niveau van de individuele leerling.

Dit betekent niet dat de instrumenten voor *assessment for learning* geen rol van betekenis kunnen hebben in het onderwijs. Maar die betekenis ligt meer in het praktische gebruik van de resultaten en de interpretatie van de betekenis daarvan voor toekomstig onderwijs dan in het beoordelen van individuele leerlingen in summatief opzicht.

3.1.2 Review van relevante studies naar assessment

In het kader van onderhavig onderzoek hebben we een computergestuurde inventarisatie verricht van mogelijk relevant onderzoek naar *assessment-for-learning* en *assessment-of-learning*. Daarbij zijn uiteenlopende zoektermen gebruikt, zowel in algemene zin (*formative – summative* / *assessment – language testing*) als specifiek: er is bv. in combinatie met de algemene zoektermen gezocht naar specifieke toetssoorten en assessmentvormen en naar diverse soorten van validiteit, waaronder de predictieve validiteit. In de zoektocht is onder andere gebruik gemaakt van de volgende computerbestanden:

- Educational Resources Information Center (ERIC-documents);
- British Educational Index;
- Psychological Index;
- Linguistics and Language Behavior Abstracts;
- Dissertation Abstracts;
- Educational testing Service Test Collection.
- Psycho-info
- Google Scholar

Een groot aantal aangetroffen studies bleek bij nader inzien niet goed te passen binnen het opgestelde zoekprofiel en is uit de database verwijderd, bijvoorbeeld omdat ze bij nader in zien niet relevant leken of omdat ze tot doublures leidden. Via de sneeuwbalmethode zijn we op het spoor gekomen van ander onderzoek dat relevant leek; een enkele keer betrof dat onderzoek dat reeds voor 2000 was gepubliceerd. Na dit proces van toevoegen en deleren hebben we de meest relevante onderzoeken ingedeeld in 4 categorieën, te weten

- Experimentele studies
- Quasi-experimentele studies
- Literatuurstudies
- Praktische publicaties

Resultaten en bevindingen uit alle vier categorieën zijn verwerkt in voorafgaande hoofdstukken. In deze paragraaf presenteren we een overzicht van de publicaties uit de eerste twee categorieën, te weten 22 publicaties die voldoen aan een van volgende criteria:

- de publicatie is een metastudie, gebaseerd op minimaal 30 andere studies;
- de publicatie doet verslag van experimenteel of quasi-experimenteel onderzoek naar assessment of learning resp. assessment for learning én betreft de rol van een of meer specifieke assessmentvormen.

De bevindingen worden eerst tabelgewijs gepresenteerd, daarna volgt van iedere publicatie een korte bespreking.

Tabel 3.3: Overzicht relevante publicaties (metastudies, experimenteel en quasi-experimenteel onderzoek na 2000)

	1	2	3	4	5	6
	Alderson & Hutha 2005	Black & Wiliam 2006b	Booth et al. 2003	Brown 2002	Chen et al. 2006	Davies & LeMahieu 2003
Soort studie						
meta-analyse *				*		*
experimenteel					*	
quasi-experimenteel	*	*	*			
assessment for learning		*				*
assessment of learning		*				
not specified	*		*	*	*	
formative	*	*	*	*	*	*
summative		*	*	*	*	
not specified						
Toetsfunctie						
intake test						
placement test						
vorderingentoets			*	*	*	*
diagnostic	*		*			
certificering				*	*	
program evaluation			*			
not specified		*				
Toetsinhoud						
kennistoets			*			
vaardigheidstoets	*		*	*		
competence-assessment		*		*	*	*
Specifieke toetssoort						
essay						
portfolio assessment				*	*	*
performance assessment						
computerbased	*		*			
adaptive	*					
peer assessment						
self-assessment	*					
reflective assessment						
360 graden feedback						
anderszins		*				

Tabel 3.3: Overzicht relevante publicaties (metastudies, experimenteel en quasi-experimenteel onderzoek na 2000) – vervolg

	7	8	9	10	11	12
	Dlaska & Krekeler 2008	Evans 2009	Greenan 1985	Gulikers et al. 2006b	Hoogestraat 2009	Lee 2007
Soort studie						
meta-analyse *						
experimenteel	*	*		*	*	
quasi-experimenteel			*			*
assessment for learning	*	*				*
assessment of learning				*		*
not specified			*		*	
formative	*	*	*	*		*
summative		*				*
not specified					*	
Toetsfunctie						
intake test						
placement test					*	
vorderingentoets		*	*	*		*
diagnostic	*	*		*		*
certificering						
program evaluation						
not specified						
Toetsinhoud						
kennistoets						
vaardigheidstoets	*	*	*	*		*
competence-assessment				*	*	
Specifieke toetssoort						
essay						*
portfolio assessment						
performance assessment				*		
computerbased						
adaptive						
peer assessment						
self-assessment	*		*			
reflective assessment		*				
360 graden feedback					*	
anderszins						

Tabel 3.3: Overzicht relevante publicaties (metastudies, experimenteel en quasi-experimenteel onderzoek na 2000) – vervolg

	13	14	15	16	17	18
	Lopes Bonilla et al. 2003	Matsuno 2009	Meuffels & Maat 2009	Neuvel 2004	Ross 2005	Shute et al. 2008
Soort studie						
meta-analyse *						
experimenteel		*		*		*
quasi-experimenteel	*		*		*	
assessment for learning			*		*	*
assessment of learning	*				*	*
not specified		*		*		
formative	*		*		*	*
summative					*	*
not specified		*		*		
Toetsfunctie						
intake test						
placement test						
vorderingentoets		*		*	*	*
diagnostic	*		*		*	
certificering						
program evaluation						
not specified						
Toetsinhoud						
kennistoets						
vaardigheidstoets	*	*	*	*	*	*
competence-assessment						
Specifieke toetssoort						
essay						
portfolio assessment						
performance assessment						
computerbased			*			*
adaptive						
peer assessment		*				
self-assessment		*		*		
reflective assessment						
360 graden feedback						
anderszins	*			*	*	

Tabel 3.3: Overzicht relevante publicaties (metastudies, experimenteel en quasi-experimenteel onderzoek na 2000) – vervolg

	19	20	21	22
	Sluijsmans et al. 2008	Smith & Tillema 2007	Topping 2003	Wiggele-Worth & Storch 2009
Soort studie				
meta-analyse *			*	
experimenteel				
quasi-experimenteel	*	*		*
assessment for learning	*	*	*	*
assessment of learning	*	*		
not specified				
formative	*	*	*	*
summative	*	*		
not specified				
Toetsfunctie				
intake test				
placement test				
vorderingentoets	*	*	*	*
diagnostic				
certificering	*	*		
program evaluation				
not specified				
Toetsinhoud				
kennistoets				
vaardigheidstoets		*	*	*
competence-assessment	*			
Specifieke toetssoort				
essay				*
portfolio assessment	*	*		
performance assessment				
computerbased				
adaptive				
peer assessment			*	
self-assessment			*	
reflective assessment				
360 graden feedback				
anderszins				

Alderson & Huhta (2005) beschrijven het project Dialang: op basis van het Europese referentiekader is voor 14 talen een adaptief diagnostisch taalassessment systeem geconstrueerd. Op basis van can do statements is een self-assessment instrument ontwikkeld. De correlaties tussen de self-assessment en taaltoetsen variëren tussen .47 en .58.

Black & Wiliam (2006b) bespreken validiteit en betrouwbaarheid als de kernwaarden van elke assessment. De auteurs demonstreren de kwetsbaarheid van gestandaardiseerde assessments: ze wijzen op de toevalsfactor waar het de opname van test items betreft en ze demonstreren dat een doorgaans klakkeloos geaccepteerd betrouwbaarheidsinterval (bij een betrouwbaarheid tussen .70 en .80) in de praktijk leidt tot een groot aantal verkeerde beslissingen. Omdat toetsverlenging geen realistische optie is, bepleiten de auteurs de inzet van teacher assessment ofwel assessment for learning met een continue karakter.

Booth et al. (2003) onderzochten met behulp van interviews welke assessmentvormen bruikbaar zijn binnen een online learning system. Allerlei soorten assessment werden gebruikt, maar er bleek een lichte voorkeur voor formatieve assessment. Er blijkt behoefte aan self-assessment en group assessment-procedures, bijvoorbeeld met gebruikmaking van online chatten en van bulletin boards. Aangeraden wordt om van meerdere soorten instrumenten gebruik te maken (de 'methode mix').

Brown (2002) inventariseert functies en toepassingsmogelijkheden voor portfolio's. Op basis van een review wordt geconstateerd dat de kwaliteit van de beoordeling zorgelijk is maar kan verbeteren door middel van assessor-training, vaststelling van performance indicatoren en afstemming van assessment en onderwijs.

Chen et al. (2006) laten zien dat het leerproces met behulp van electronische portfolio's geautomatiseerd in beeld kan worden gebracht. Daartoe analyseerden ze 583 portfolio's van basisschoolleerlingen volgens een methode die gebruik maakt van vier schema's voor kunstmatige intelligentie. Analyse van de resultaten liet zien dat met de geautomatiseerde analysemethode een betrouwbaarheid van .70 kon worden bereikt in de beoordeling.

Davies & LeMahieu (2003) geven een review van literatuur over portfolio-gebruik. Belangrijkste functies van portfolio's zijn: het aantonen van groei, assessment for learning, inzichtelijk maken van de relatie tussen leerproces en resultaat. De review laat zien dat Over het algemeen wordt het portfolio gezien als een krachtig instrument in assessment for learning. Bij assessment of learning is de tegenvallende betrouwbaarheid in beoordeling van portfolio een bron van zorg; met beoordelaarstraining zijn eenduidiger resultaten te behalen. Portfolio-gebruik bevordert de motivatie en draagt daarmee indirect bij aan het leren, doordat de leerder eigenaar is van het portfolio, zelf keuzen kan maken en verantwoordelijk is voor het eigen portfolio. Leerkrachten moeten getraind worden in het geven van feedback die het leren bevordert. Leerkrachten blijken dankzij de ingebruikname van portfolio's meer inzicht te krijgen in de onderwijsdoelen en mogelijke beoordelingscriteria.

Dlaska & Krekeler (2008) laten op basis van onderzoek bij 46 volwassen leerders van Duits als tweede taal dat zij grote moeite hebben om hun eigen uitspraakvaardigheid te beoordelen: de T2-leerders identificeerden nauwelijks de helft van de klanken die volgens

ervaren beoordelaars problematisch waren: de fonologische regels van de moedertaal zijn de belangrijkste oorzaak van de problemen.

Evans (2009) laat in een quasi-experimentele studie met highschool-leerlingen zien, dat "reflective assessment" als onderdeel van het curriculum moedertaalonderwijs leidt tot betere resultaten en tot langer beklijven.

Greenan (1985) behoort tot de eerste empirische studies over formatief gebruikte self-assessment. Doel van het onderzoek was na te gaan of studenten hun eigen communicatieve vaardigheden konden inschatten. Test items hadden betrekking op onder andere woordenschat, lees- en schrijfvaardigheid en de mondelinge vaardigheden; voor de self-assessments werd van een vierpunts-Likertschaal gebruik gemaakt. De gebruikte instrumenten waren van goede kwaliteit (interne consistentie =.93. en Pearson r =.81). De gevonden correlatie tussen self-assessment en de Performance test was middelmatig (.42). Greenan benadrukt dat zelfbeoordeling het onderwijs ten goede komt en dat de studenten kunnen worden geoefend in zelfbeoordeling.

Gulikers et al. (2006b) onderzoeken de hypothese dat authentieke assessment leidt tot beter begrip en hogere onderwijsopbrengsten. Daartoe bevragen ze 118 studenten Sociaal werk over de mate van authenticiteit van assessments. Het blijkt dat drie aspecten van authenticiteit - authentieke, complexe leertaken, een authentieke fysieke leeromgeving en de assessmentvorm - wel leiden tot beter leerrendement, maar als authentiek ervaren beoordelingscriteria niet. De verklaring hiervoor wordt gezocht in de operationalisering van de taken: oppervlakkig leren – een studiehouding gericht op memorisatie - bleek al voldoende om de taken goed te verrichten. Een 'deep study approach' – gericht op begrip - leidde niet tot betere prestaties. Wellicht staan daarom te concrete beoordelingscriteria een studiehouding gericht op begrip in de weg bij meer ervaren studenten.

Hoogestraat (2008) doet validatie-onderzoek naar het 360 graden feedback instrument van PMG (Philips Management Group). Bij dit type onderzoek wordt voor verschillende functies in het bedrijfsleven geprobeerd om overeenstemming te vinden over gedragsindicatoren die aangeven wie goed kan functioneren in een specifiek banenprofiel. Gedemonstreerd wordt dat met een beperkt aantal algemene indicatoren - bijvoorbeeld probleemoplossend vermogen, communicatieve vaardigheden, leiderschapsvaardigheden zoals visie-ontwikkeling en kunnen delegeren - al een hoge mate van overeenstemming bij beoordelaars kan worden bereikt.

Lee (2007) onderzoekt de aard van de feedback die 26 docenten Engels geven op geschreven teksten van 174 leerlingen (12-16 jaar). 91% van de feedback heeft betrekking op vormfouten, 4% betreft inhoud. In de meeste gevallen worden de correcties er direct bij geschreven. Bovendien kregen de leerlingen een cijfer voor de tekst. Uit de inventarisatie blijkt dat de assessment steeds een summatief karakter had. In interviews geven leerkrachten te kennen dat ze geen direct verband zien tussen deze assessments en de inhoud van het onderwijs. Assessment blijkt steeds het karakter van *assessment of learning* te hebben.

López Bonilla & Rodríguez Linares (2003) ontwikkelen een authentieke leesassessment voor lezen met open vragen. Na een pilot werd het instrument gebruikt bij 95 highschool studenten. Het instrument bleek geschikt om verschillende deelvaardigheden van lezen van elkaar te onderscheiden (de publicatie geeft echter geen psychometrische onderbouwing).

Matsuno (2009) vergelijkt met behulp van een Rasch multifacet model docentbeoordelingen met peer- en self-assessments van schrijfvaardigheid bij 79 Japanse T2-leerders van het Engels op universiteitsniveau. Zelfbeoordelingen waren lager dan de docentoordelen, met name bij vaardige studenten; als geheel is de kwaliteit van zelfbeoordeling ontoereikend, zodat dit geen rol kan spelen in formele assessments. Oordelen van peers waren systematisch milder dan docentoordelen, maar wel consistent en niet afhankelijk van de eigen schrijfvaardigheid. Om die reden kan overwogen worden om peer assessment een functie te geven in het onderwijs.

Meuffels & Maat (2009) is een empirische studie naar inkorting van een diagnostische toets. De schrijftoets, bedoeld voor leerling-registeraccountants, wordt afgenomen voordat de leerling-accountants een opleidingstraject ingaan. Er deed zich een praktisch probleem voor: de duur van de toetsafname (3 uur) stond niet in aanvaardbare verhouding tot de duur van het opleidingstraject (8 uur). Door de betrouwbaarheid per diagnostisch onderdeel te berekenen en een passingsmodel te hanteren voor de convergente validiteit, kon op verantwoorde manier de toetslengte worden bekort met 25% zonder in te boeten op de gestelde eisen aan betrouwbaarheid en validiteit.

Neuvel et al. (2004) onderzochten bij 345 leerlingen op het mbo de relatie tussen zelfbeoordeling, leerkrachtbeoordeling en prestaties op leestoetsen. De deelnemers oordeelden aanzienlijk positiever over hun eigen taalvaardigheid dan zou moeten volgens de toetsresultaten. Terwijl meer dan tweederde van de leerlingen op niveau 1 en 2 dacht zelfs de taaltaken op niveau B2 voldoende tot goed uit te kunnen voeren, lieten de toetsresultaten zien dat 83% van de deelnemers onder dat niveau zat. Een soortgelijk beeld werd gevonden bij de leerlingen op niveau 3 en 4: gemiddeld genomen dachten de leerlingen dat hun leesvaardigheid tegen niveau C1 aan lag, terwijl afgaande op de oordelen van docenten (zie hierboven) het feitelijke niveau gemiddeld zelfs onder niveau B1 uitkwam. De toetsresultaten kwamen meer overeen met de schatting van de docenten dan met de zelfbeoordeling van de leerlingen. Hoe lager het feitelijke taalniveau, hoe groter de zelfoverschatting bleek te zijn.

Ross (2005) demonstreert in een longitudinale studie naar taalverwerving bij Japanse studenten English for Academic Purposes (n = 2215) dat er een sterkere groei in taal-ontwikkeling optreedt bij groepen die gebruikmaken van formatieve assessment (n=1102) dan bij groepen die traditioneel summatief getoetst worden (n=1113). Groepsgemiddelden van formatieve toetsen en TOEFL-scores (Lezen, Luisteren) werden gebruikt. Ross constateert dat in zijn programma geen verschil in betrouwbaarheid (internal consistancy) van de summatieve en de formatief gebruikte assessments (resp. .80 en .79). Multivariate analyse met beginscores als covarianten laat zien dat de groepen met formatieve assessment snellere taalontwikkeling te zien geven.

Overigens reageerde Stapleton (2006) negatief op de studie van Ross: die zou niet aannemelijk hebben gemaakt dat andere verklaringen voor de gevonden effecten mogelijk zijn en storende variabelen zouden niet adequaat zijn uitgesloten.

De studie van Shute et al. (2008) heeft betrekking op algebra op highschool-niveau, maar is interessant omdat is nagegaan of toevoeging van feedback negatieve effecten heeft op de kwaliteit van assessment. Er werd gebruik gemaakt van een adaptief, diagnostisch assessment systeem dat feedback kan geven bij foutief gemaakte algebra-oefeningen. Er werd geëxperimenteerd met highschool studenten (n=268) in vier condities:

1 uitgebreide feedback (foutmelding plus uitleg) en adaptieve aanbieding
2 beperkte feedback (vermelding goed of fout) en adaptieve aanbieding
3 uitgebreide feedback (foutmelding plus uitleg) en lineaire aanbieding
4 controlegroep zonder gebruik van het assessment-instrument

De resultaten laten zien dat uitgebreide feedback (foutmelding plus uitleg) effectiever voor het leerproces dan alleen vermelding van goed of fout. Adaptieve resp. lineaire aanbieding leverde geen significant verschil op. Tot slot bleken validiteit noch betrouwbaarheid van de assessment aangetast door de toevoeging van feedback. Dit suggereert dat toevoeging van feedback aan bestaande assessments het leerproces efficiënter kan laten verlopen zonder dat de assessmentfunctie wordt aangetast.

Sluijsmans et al. (2008) beschrijven presenteren een methode om de beoordeling van authentieke taken in realistische settings valide en zo betrouwbaar mogelijk te maken, zodat de resultaten als summatieve beoordelingen in een elektronisch portfolio kunnen worden opgenomen. Deze methode wordt gekenmerkt door

- een brede variëteit aan authentieke taken;
- een vaste set van beordelingscriteria;
- evaluatie per beoordelingscriterium over meerdere taken en per taak over verschillende criteria.

Smith & Tillema (2007) bestuderen de vraag hoe transparant de criteria zijn die gebruikt worden bij de beoordeling van summatief gebruik van portfolio's van studenten Engels in Israël en Nederland. Daartoe werden 35 leerkrachten schriftelijk én mondeling bevraagd over de aard van de criteria die ze gebruikten bij beoordeling. De vragen hadden betrekking op drie thema's, namelijk de doelen van een portfolio, de beoordeling van portfolio's en de manier waarop een cijfermatig oordeel tot stand komt. Belangrijkste bevinding is, dat veel leerkrachten werken met een tevoren vastgelegde set criteria die ongevoelig is voor de context en het doel waarmee de portfolio's zijn gemaakt. Opvallend is dat het verzorgen van feedback aan de studenten het minst genoemd is als doelstelling bij de beoordeling.

Topping (2003) biedt een review van empirische studies naar self-assessment en peer-assessment, met een meta-analyse op betrouwbaarheid, validiteit en bruikbaarheid van deze twee soorten van assessment. De betrouwbaarheid van self-assessment is laag; met name komt vaak zelfoverschatting voor. De betrouwbaarheid van peer assessment zou van dezelfde grootorde zijn als die van leerkrachtoordelen. Het effect van beide assessmentvormen op het onderwijsrendement is positief.

Wigglesworth & Storch (2009) onderzochten het effect van pair training / peer evaluation in schrijflessen bij 144 tweede taalleerders op universitair niveau. 48 paren van schrijvers schreven een betoog. De teksten werden geanalyseerd op onder andere fluency , zinscomplexiteit en spelfouten; de interbeoordelaarsbetrouwbaarheden daarbij lagen boven .84. De interacties tussen de paren werden geanalyseerd. Ruim de helft van de interactie-episodes hadden betrekking op de inhoud van de essays, 335 had betrekking op taalgerelateerde zaken; daarvan had 55% betrekking op woordkeuze, 38% op grammaticale zaken en 8% op spelling. Dit resulteerde wel in een significant hoger percentage foutloze zinnen in de experimentele groep. De auteurs stellen op basis hiervan, dat samenwerkend schrijven het leren bevordert.

Samengevat komen de belangrijkste bevindingen op het volgende neer.
Er is een kennelijke behoefte om gebruik te maken van self-assessment (Booth 2003); deze assessmentvorm correleert echter matig tot slecht met andere, aantoonbaar betrouwbare indicaties van taalvaardigheid (vgl. Greenan 1985; Alderson & Huhta 2005; Dlaska & Krekeler 2008; Matsuno 2009; Neuvel et al. 2004; Topping 2003).

De betrouwbaarheid van peer assessment is iets beter dan van self-assessment (vgl. Matsuno 2009; Topping 2003; Wigglesworth & Storch 2009) en bovendien lijkt peer assessment een groter en beter aantoonbaar effect op het onderwijs te hebben.

Er is in het onderwijs ook een grote behoefte om gebruik te maken van portfolio's. Gebruik ervan in het onderwijs is aantoonbaar van nut (Davies & LeMahieu 2003), maar de beoordeling van portfolio's is problematisch: leerkrachten weten niet goed wat ze beoordelen (vgl. Smith & Tillema 2007), maar na training van beoordelaars kan verbetering optreden (Brown 2002, Sluijsmans et al. 2008) en de beoordeling kan zelfs geautomatiseerd worden (Chen et al. 2006).

Bij genoemde en andere vormen van assessment (vgl. Hoogestraat 2008; Evans 2009) is met name de feedback die gegeven wordt van belang voor het onderwijsrendement (vgl. Black & Wiliam 2006b; Ross 2005; Lee 2007). Shute et al. (2008) tonen aan dat de toetsfunctie niet belemmerd wordt door feedback toe te voegen aan een assessment.

Opmerkelijk is, dat in de onderzochte literatuur geen onderzoek is aangetroffen dat betrekking heeft op *predictieve validiteit* van taaltoetsen die bedoeld zijn voor gebruik in vmbo en mbo. De weinige studies die aandacht schenken aan predictieve validiteit hebben betrekking op talig materiaal voor kleuters (bv. Braams & Bosman, 2000 ; Verhoeven & Vermeer, 2003; Ciat & Roy, 2008) of op cognitieve vaardigheden (vgl. bv. Van Batenburg & Van der Werf 2004; Van den Bergh & Bleichrodt 2000). Evenmin is er onderzoek aangetroffen naar de *cross-validiteit* van instrumenten voor *assessment-for-learning* en *assessment-of-learning.*

3.3 Conclusies

In dit derde hoofdstuk is gepoogd een antwoord te geven op de tweede en de derde onderzoeksvraag. De tweede onderzoeksvraag luidde: Op welke manieren kunnen de opbrengsten van taalonderwijs in het (v)mbo op een gestandaardiseerde manier worden gemeten? In de literatuur lijkt eenstemmigheid te bestaan over vier kenmerken waaraan een assessment in ieder geval zou moeten voldoen:

- De taken in de taaltoets komen overeen met taken in de praktijk
- De taken in de taaltoets en taken in de praktijk zijn vergelijkbaar moeilijk
- De toetsscore geeft een goed beeld van het prestatieniveau in de praktijk
- De toetsscores hebben predictieve validiteit voor de beroepspraktijk

In de tweede onderzoeksvraag kan het begrip 'gestandaardiseerd' worden opgevat in de betekenis van: met centraal landelijke toetsen waarvan validiteit en betrouwbaarheid tevoren zijn bepaald. De herinvoering van de landelijke examens voor het mbo en de overige talige eisen die vanaf 2014 worden gesteld aan deelnemers aan het mbo op basis van Commissie Meijerink (zie Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2008), vormen een gedeeltelijk antwoord op de tweede onderzoeksvraag. Dit deel van het antwoord betreft de summatieve assessment ofwel de *assessment of learning*.

Wanneer met het begrip 'gestandaardiseerd' ook wordt gedoeld op enigerlei uniformering in de vele assessment-instrumenten die momenteel worden ingezet voor *assessment for learning*, is het antwoord minder eenvoudig. In dat geval zijn er strikt genomen twee maatregelen nodig:

- elk instrument - al dan niet op school ontwikkeld - dat bedoeld is voor *assessment of learning* zou duidelijk moeten verwijzen naar een algemeen aanvaard en herkenbaar stelsel van taalniveaus. Het ligt voor de hand om de indeling van Commissie Meijerink als referentiepunt te nemen.
- Ook de beoordelingsprocedures bij de assessment-instrumenten zouden moeten worden gestandaardiseerd. Met name bij *assessment for learning*-assessments is dat bijna onbegonnen werk. Wel kunnen algemeen bruikbare richtlijnen worden ontwikkeld waarvan aannemelijk is dat makers van *assessment-for-learning*-instrumenten zich daaraan zullen conformeren. Voorwaarden daarbij liggen in de sfeer van de eenvoudige toepasbaarheid en duidelijke meerwaarde voor betrouwbaarheid en objectiviteit.

Dit alles betekent dat voor de *assessment of learning* het onderwijsrendement het beste kan worden gemeten met behulp van landelijk uitgerolde, gestandaardiseerde toetsen die alle verwijzen naar één onderliggende niveau-indeling. Voor de instrumenten ten behoeve van de *assessment for learning* is met name de relatie tussen inhoud van assessment en inhoud van onderwijs van belang. Deze instrumenten zijn ervoor bedoeld om invloed te hebben op de verdere inhoud van het onderwijs, om door middel van feedback bij te dragen aan het leerproces. Ze zijn er niet voor bedoeld en ook niet voor geschikt om onderwijsrendementen in beeld te brengen. De betekenis van deze instrumenten ligt in het praktische gebruik van de resultaten voor het aanpassen van het onderwijs aan de behoeften van de leerlingen.

Uit het in paragraaf 3.2 gepresenteerde overzicht blijkt, dat er met name onderzoek is gedaan naar de waarde van self-assessment, peer-assessment en het portfolio.

Onderwijskundig bezien hebben deze instrumenten zeker waarde (*assessment-for-learning*) maar toetstechnisch bezien voldoen ze niet aan de meest elementaire eisen (*assessment-of-learning*). Er heeft echter nauwelijks onderzoek plaatsgevonden naar de vraag hoe deze en vergelijkbare instrumenten die bijdragen aan het onderwijs, zo geconstrueerd kunnen worden dat ze voldoen aan de gewenste kenmerken van summatieve assessment (zie echter Wools 2009; Wools et al., te verschijnen). Om die reden lijkt het momenteel niet verstandig om instrumenten die bedoeld zijn als *assessment- for-learning* in te zetten met het oog op de *assessment of learning*-functie.

Daarmee is niet gezegd dat de instrumenten die ingezet worden als, of bedoeld zijn voor *assessment-of-learning* wél aan de minimale eisen voldoen: de literatuur geeft geen duidelijk antwoord op de vraag of aan de standaardeisen voldaan wordt. Evenmin is hiermee gezegd dat de *assessment-for-learning*-instrumenten geen rol van belang mogen spelen in het onderwijs: die kunnen zeker van nut zijn, maar dan in de functie waarvoor ze bedoeld zijn - om het onderwijs inhoudelijk nauwkeurig af te stemmen op de behoeften van de leerlingen - en niet als summatieve instrumenten.

Het literatuuroverzicht in par 3.2 geeft een duidelijk, negatief antwoord op de derde onderzoeksvraag. Die vraag luidde: Wat is er bekend over predictieve validiteit van moderne toetsvormen? In de onderzochte literatuur is er geen onderzoek aangetroffen dat betrekking heeft op predictieve validiteit van taaltoetsen bedoeld voor gebruik in vmbo en mbo. De weinige studies die aandacht schenken aan predictieve validiteit hebben betrekking op talig materiaal voor kleuters (bv. Braams & Bosman, 2000 ; Verhoeven & Vermeer, 2003) of op cognitieve vaardigheden (vgl. bv. Van Batenburg & Van der Werf 2004; Van den Bergh & Bleichrodt 2000); wel wordt met enige regelmaat de noodzaak van onderzoek naar predictieve validiteit onderstreept (vgl. bv. Gysen & Van Avermaat 2005; Perie et al. 2009). Evenmin is er onderzoek aangetroffen naar de cross-validiteit van instrumenten voor *assessment-for-learning* en *assessment-of-learning.*

Hoofdstuk 4 Bevraging bij docenten en experts

Om na te gaan of de in de literatuur beschreven inzichten overeenkomen met denkbeelden in de praktijk, is besloten om een bescheiden veldbevraging te organiseren en om experts op het gebied van assessment te raadplegen. Van beide bevragingsronden wordt in dit hoofdstuk kort verslag gedaan.

4.1 Opzet veldbevraging en expertbevraging

Om na te gaan of de in de literatuur beschreven inzichten overeenkomen met denkbeelden in de praktijk, is besloten om enkele telefonische interviews te houden met enkele leerkrachten mbo en om schriftelijk een aantal leerkrachten in vmbo en mbo te bevragen. In de veldbevraging is sprake van random selectie. Organisatorisch ging de veldbevraging als volgt: met de leerkrachten is contact gelegd via een open call in oktober 2009 op de mailinglist list-nederlands@digischool.nl . Een aantal leerkrachten is gevraagd om een vragenlijst in te vullen die was geplaatst op de besloten website www.ledenpanel.nl. De vragenlijst kon worden ingevuld vanaf 20 november tot en met 20 december.

Kern van de schriftelijke vragenlijst voor leerkrachten betreft de in hoofdstuk 3 besproken lijst met 20 eisen die aan assessment-instrumenten gesteld kunnen worden. Over elk van deze eisen is gevraagd hoe belangrijk de leerkracht deze eis vindt; daarbij is onderscheid gemaakt naar instrumenten met een formatieve functie en met een summatieve functie. Aanvullend zijn er aan de docenten vragen gesteld over de soort assessment-instrumenten die ze gebruiken en over de vaardigheden die met de betreffende instrumenten in beeld gebracht worden. Bijlage 2 bevat de vragenlijst die aan leerkrachten is voorgelegd. De vragen zijn beantwoord door 7 leerkrachten uit het mbo en 8 leerkrachten Nederlands uit het vmbo.

Bij de experts is nadrukkelijk selectief contact gelegd, daarbij werd gelet op instelling en op persoonlijke kwaliteiten cq. bewezen jarenlange ervaring met assessment. Enkele experts zijn in een individueel gesprek bevraagd, anderen hebben deelgenomen aan een gezamenlijke bijeenkomst op 12 januari 2010 in Nijmegen. Voorafgaande daaraan hebben in totaal zes experts via internet een schriftelijke vragenlijst ingevuld; een groot deel van de vragen had betrekking op dezelfde zaken als aan de orde waren gekomen bij de leerkrachten. Dit maakt vergelijking van een groot deel van de responses uit de bevraagde groepen mogelijk.

4.2 Resultaten

Aan de docenten is gevraagd om van een aantal soorten toetsen op een vierpuntsschaal aan te geven hoe vaak ze er gebruik van maken. Antwoordmogelijkheden waren (bijna) nooit / af en toe / regelmatig / (erg) vaak. De procentuele scores op de eerste twee en de laatste twee antwoordmogelijkheden zijn samengevoegd. Tabel 4.2 geeft de resultaten weer:

Tabel 4.1: Mate van gebruik van verschillende assessmentvormen

	vmbo	**mbo**
	regelmatig tot vaak	**regelmatig tot vaak**
kennistoets	50,0%	42,9%
essaytoets	25,0%	0,0%
vaardigheidstoets	62,5%	71,4%
casustoets	0,0%	57,1%
voortgangstoets	37,5%	28,6%
peer assessment	12,5%	14,3%
self-assessment	0,0%	0,0%
stage-opdracht	25,0%	57,1%
projectopdracht	12,5%	42,9%
gedragsassessment	12,5%	42,9%
afstudeeropdracht	37,5%	28,6%
portfolio assessment	0,0%	42,9%
criteriumgericht interview	0,0%	28,6%
reflectie-opdracht	0,0%	42,9%

> 50% = met grote regelmaat
25 tot 50% = met enige regelmaat
< 25% = weing gebruik

De veldbevraging laat zien dat in het vmbo de kennistoets en de vaardigheidstoets verreweg de meestgebruikte toetsvormen zijn. Daarnast wordt met enige regelmaat gebruik gemaakt van voortgangstoetsen, essaytoetsen, stage-opdrachten en afstudeeropdrachten.
In het mbo zien we een ander beeld: daar zijn de vaardigheidstoets, de casustoets en de stageopdracht de meestgebruikte vormen van assessment. In het mbo wordt ook met enige regelmaat gebruik gemaakt van twee vormen van assessment die in het vmbo helemaal niet genoemd zijn, namelijk de portfolio assessment en de reflectie-opdracht. De in de literatuur veelbesproken assessmentvormen zoals het criteriumgericht interview, de peer assessment en de self assessment worden in beide schooltypes weinig of niet gebruikt.

In de vragenlijsten voor docenten en die voor experts is gevraagd op een vierpuntsschaal het belang aan te geven van de eerder besproken 20 kenmerken bij taaltoetsen in het vmbo/mbo. Bijvoorbeeld is gevraagd hoe belangrijk het is dat een formatieve taaltoets de toetsscores een goede voorspeller zijn van prestaties in de beroepspraktijk. Als antwoord-mogelijkheden kon worden gekozen uit Erg onbelangrijk / Onbelangrijk / Belangrijk / Erg belangrijk / Weet niet of Geen mening. Bij de gegevensverwerking zijn de scores op de eerste twee en de laatste twee antwoordcategorieën gesommeerd.

In onderstaande tabel is voor de twee onderscheiden soorten toetsen (formatief en summatief) en voor de drie groepen respondenten (leerkrachten vmbo, leerkrachten mbo, experts) aangegeven hoeveel procent heeft gekozen voor de antwoorden Belangrijk of Erg belangrijk.

Tabel 4.2: Het belang van verschillende assessment-kenmerken volgens leerkrachten vmbo, leerkrachten mbo en experts

		formatieve toetsen			summatieve toetsen		
		vmbo	**mbo**	**expert**	**vmbo**	**mbo**	**expert**
	omschrijving eis						
1	taken komen overeen met taken praktijk	87,5%	85,7%	58,0%	87,5%	100,0%	71,0%
2	taken zijn competentiegericht	62,5%	57,1%	43,0%	100,0%	71,4%	71,0%
3	moeilijkheidsgraad als in praktijk	75,0%	85,7%	58,0%	100,0%	100,0%	71,0%
4	geeft goed beeld prestatieniveau praktijk	75,0%	100,0%	58,0%	87,5%	100,0%	57,0%
5	tevoren vastgestelde kwaliteitseisen	100,0%	100,0%	71,0%	100,0%	100,0%	86,0%
6	leerlingen weten wat van ze verwacht wordt	87,5%	85,7%	72,0%	100,0%	100,0%	57,0%
7	validiteit	100,0%	85,7%	100,0%	100,0%	100,0%	100,0%
8	beoordelaars-onafhankelijkheid	62,5%	57,1%	43,0%	75,0%	85,7%	86,0%
9	betrouwbaarheid	100,0%	85,7%	72,0%	100,0%	100,0%	86,0%
10	toets- en onderwijsdoelen komen overeen	100,0%	100,0%	87,0%	100,0%	100,0%	100,0%
11	toets maken als leerling eraan toe is	87,5%	28,6%	86,0%	87,5%	42,9%	43,0%
12	toets is leerzaam voor de leerling	87,5%	85,7%	87,0%	75,0%	57,1%	14,0%
13	toets is leerweg-onafhankelijk	62,5%	71,4%	57,0%	75,0%	71,4%	73,0%
14	kosten staan in verhouding tot opbrengst	75,0%	100,0%	87,0%	75,0%	100,0%	73,0%
15	draagt bij aan leerproces door feedback	87,5%	85,7%	100,0%	87,5%	100,0%	29,0%
16	invloed op inhoud onderwijsleerproces	100,0%	85,7%	100,0%	100,0%	71,4%	29,0%
17	bevordert vermogen tot zelfbeoordeling	87,5%	71,4%	100,0%	75,0%	71,4%	43,0%
18	voorspellen prestaties leerling in onderwijs	50,0%	71,4%	43,0%	100,0%	100,0%	100,0%
19	scores zijn generaliseerbaar over taken	50,0%	42,9%	57,0%	62,5%	85,7%	100,0%
20	voorspellen prestaties in beroepspraktijk	37,5%	85,7%	43,0%	87,5%	85,7%	73,0%
	gemiddeld	78,8%	78,6%	70,4%	88,8%	87,1%	68,1%

Uit tabel 4.2 is op te maken dat zowel docenten als experts vinden dat toetsdoelen en onderwijsdoelen overeen moeten komen (10) en daarnaast validiteit (7) zeer belangrijk vinden; dit geldt zowel de formatieve als de summatieve toetsen. Waar het de betrouwbaarheid (9) betreft, ligt dit anders: docenten vinden de betrouwbaarheid vrijwel unaniem van groot belang bij zowel formatieve als summatieve toetsen. De experts vinden betrouwbaarheid in mindere mate van belang bij summatieve toetsen - wellicht in de wetenschap dat geen enkel instrument volledig betrouwbaar kan zijn - en in nog mindere mate bij de formatieve toetsen. Ter toelichting daarbij werd ten tijde van het op 12 januari 2010 georganiseerde symposium opgemerkt, dat het voornaamste doel van een formatieve toets niet is om een betrouwbare meting te doen maar om indicaties te verkrijgen rondom wat beheerst wordt en wat niet, om daar vervolgens in het onderwijs op in te spelen. In dezelfde lijn ligt de bevinding van experts dat een formatieve toets niet per se beoordelaars-onafhankelijk (8) scoorbaar moet zijn: de kosten en moeite wegen in dat geval niet op tegen de opbrengsten bij de formatieve toets.

Bij de formatieve toetsen (*assessment for learning*) zijn de experts minder veeleisend dan de leerkrachten: dit is met name het geval waar het gaat om de predictieve validiteit, zoals die tit uiting komt in de mate waarin toetstaken de taken in de beroepspraktijk weerspiegelen (4), de mate waarin de prestaties op de toets die in de praktijk voorspellen (20) en de moeilijkheidsgraad van de assessment die overeen zou moeten komen met de praktijk (3). Het zijn met name de mbo-leerkrachten die benadrukken dat de assessmenttaken even moeilijk moeten zijn als die in de praktijk en dat de assessment de prestaties in de beroepspraktijk zouden moeten kunnen voorspellen.

Bij de summatieve toetsen (*assessment of learning*) benadrukken de experts - naast de al genoemde validiteit (7) en de parallellie tussen toets- en onderwijsdoelen (10) - de voorspellende waarde van assessments voor het vervolgonderwijs (18) en de generieke werking van assessments: wat getoetst wordt, moet generaliserende uitspraken opleveren voor andersoortige taken (19). Opvallend is verder de afwijkende mening van de experts over de functie van de assessments voor het onderwijs zelf: de leerkrachten vinden in meerderheid dat *assessment of learning* leerzaam moet zijn voor de leerlingen (12), door feedback moet bijdragen aan het leerproces (15), het onderwijsleerproces inhoudelijk moet beïnvloeden (16) en het vermogen tot zelfbeoordeling moet bevorderen (17). De experts benadrukken dat dit alles juist niet de functie van summatieve toetsing is.

Meer in algemene zin lijken leerkrachten alle mogelijke kenmerken van toetsen van groot belang te vinden; experts lijken eerst te kijken naar het doel van de toetsing om op grond daarvan te definiëren aan welke kenmerken een toets zou moeten voldoen. Het meest extreem komt deze opvatting tot uiting bij de vraag of toetsen moeten bijdragen aan het leerproces door middel van feedback (15): het merendeel van de docenten vindt nodig bij zowel formatieve als summatieve toetsen, terwijl de experts dit unaniem nodig achten bij de formatieve toetsen terwijl slechts een kleine minderheid van de experts dit noodzakelijk vindt bij de summatieve toetsen.

De reacties laten zien dat in het mbo meer waardering is voor competentiegericht toetsen dan in het vmbo. Dit komt overeen met het in de eerste hoofdstukken geschetste beeld. Tegelijkertijd lijken lang niet alle leerkrachten een scherp oog te hebben voor de functie van

de assessmentvormen en de eisen die op grond daarvan gesteld zouden moeten worden aan de in te zetten instrumenten. Het is alsof de retoriek van de assessmenttechnieken wel door de leerkrachten beheerst wordt, maar deze niet steeds goed gekoppeld kan worden aan de beoogde assessment-functies.

Hoofdstuk 5 Besluit

5.1 Antwoord op de onderzoeksvragen

Aan deze studie lagen drie onderzoeksvragen ten grondslag:

1. Welke toetsvormen zijn geschikt om zowel de *assessment of learning*-functie te vervullen als de *assessment for learning*-functie?
2. Op welke manieren kunnen de opbrengsten van taalonderwijs in het (v)mbo op een gestandaardiseerde manier worden gemeten?
3. Wat is er bekend over predictieve validiteit van moderne toetsvormen?

Op basis van de beschrijving van assessmentvormen in hoofdstuk 2 is geconstateerd dat een groot aantal assessmentvormen in principe bruikbaar is voor zowel *assessment of learning* als voor de *assessment for learning*. Aan elke vorm van assessment kunnen principieel steeds beide toetsfuncties worden toegedicht: tot welke soort een assessment moet worden gerekend, is afhankelijk van de intentie van de leerkracht om op basis van de resultaten het onderwijs bij te stellen. Hoe algemener de toetsdoelen van een bepaalde assessment echter zijn geformuleerd, hoe lastiger het is om op basis van de resultaten ervan zinnige en gerichte feedback te formuleren om het verdere onderwijs te richten.

Naarmate een toets inhoudelijk dichterbij de korte termijn doelen van onderwijs staat, kan het instrument beter gebruikt worden voor *assessment for learning* (formatieve toetsing); tegelijkertijd zorgt deze korte afstand tussen assessment-instrument en de leerdoelen ervoor dat de toets inhoudelijk een kleiner domeinbereik heeft en daardoor minder geschikt is voor *assessment of learning* (summatieve toetsing). Uit de literatuur (hoofdstuk 3) en uit de respons van de experts op de enquête (hoofdstuk 4) is verder gebleken dat er aan *assessment for learning*-instrumenten andersoortige eisen worden gesteld dan aan *assessment of learning*-instrumenten. Met de kanttekening dat de verschillen gradueel zijn en niet in absolute zin moeten worden opgevat, kunnen de belangrijkste verschillen als volgt worden samengevat:

Tabel 5.1 Belangrijkste verschillen tussen *assessment for learning* en *assessment of learning*

Assessment for learning	**Assessment of learning**
Toetsen moeten leerzaam zijn en d.m.v. feedback bijdragen aan het leerproces	Toetsen hoeven niet leerzaam te zijn of bij te dragen aan het leerproces
Toetsen moeten bijdragen tot het vermogen tot zelfbeoordeling	Toetsen hoeven het vermogen tot zelfbeoordeling niet te bevorderen
Validiteit van grootste belang	Validiteit van belang
Betrouwbaarheid van minder groot belang	Betrouwbaarheid van groot belang
Beoordelaars-onafhankelijkheid van minder groot belang	Beoordelaars-onafhankelijkheid van groot belang
Taken hoeven niet per se competentie-gericht te zijn of overeen te komen met de praktijk	Taken zijn liefst competentiegericht en overeenkomen met taken uit de beroeps-praktijk
Predictieve validiteit is minder van belang	Predictieve validiteit t.a.v. school en beroepsuitoefening zo groot mogelijk

Doordat de *assessment for learning*-instrumenten inhoudelijk een directere relatie hebben met de onderwijsdoelen op korte termijn en omdat zich op dat vlak grote verschillen tussen opleidingen voordoen, lenen de *assessment for learning*-instrumenten zich er minder gemakkelijk voor om gestandaardiseerd te worden. Weliswaar worden er veelbelovende pogingen ondernomen om onder deze assessment-instrumenten één gemeenschappelijk inhoudelijk domein te leggen (resp. het Europese referentiekader, het daarvan afgeleide Raamwerk Nederlands en Commissie Meijerink), maar daarmee is niet gegarandeerd dat onder deze instrumenten ook een gemeenschappelijke meetlat ligt (zie ook Weir 2005). Het creëren van zo'n gemeenschappelijke meetlat is des te moeilijker, omdat de *assessment for learning*-instrumenten doorgaans op een kleiner domein betrekking hebben en het gehele meetbereik in het betrouwbaarheidsinterval van bestaande gestandaardiseerde summatieve toetsen zou passen.

In hoofdstuk 3 komen de tweede en de derde onderzoeksvraag aan de orde. De tweede onderzoeksvraag luidde: Op welke manieren kunnen de opbrengsten van taalonderwijs in het (v)mbo op een gestandaardiseerde manier worden gemeten? In de literatuur blijkt eenstemmigheid over vier kenmerken waaraan een assessment zou moeten voldoen:

- De taken in de taaltoets komen overeen met taken in de praktijk
- De taken in de taaltoets en taken in de praktijk zijn vergelijkbaar moeilijk
- De toetsscore geeft een goed beeld van het prestatieniveau in de praktijk
- De toetsscores hebben predictieve validiteit voor de beroepspraktijk

In de tweede onderzoeksvraag kan het begrip 'gestandaardiseerd' worden opgevat in de betekenis van: met centraal landelijke toetsen waarvan validiteit en betrouwbaarheid tevoren zijn bepaald. De herinvoering van de landelijke examens voor het mbo en de overige talige eisen die vanaf 2014 worden gesteld aan deelnemers aan het mbo op basis van Commissie Meijerink (zie Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2008), vormen een gedeeltelijk antwoord op de tweede onderzoeksvraag. Dit deel van het antwoord betreft de summatieve assessment ofwel de *assessment of learning*.

Wanneer met het begrip 'gestandaardiseerd' ook wordt gedoeld op enigerlei uniformering in de vele assessment-instrumenten die momenteel worden ingezet voor *assessment for learning*, is het antwoord minder eenvoudig. In dat geval zijn er strikt genomen twee maatregelen nodig:

- elk instrument - al dan niet op school ontwikkeld - dat bedoeld is voor *assessment of learning* zou duidelijk moeten verwijzen naar een algemeen aanvaard en herkenbaar stelsel van taalniveaus. Het ligt voor de hand om de indeling van Commissie Meijerink als referentiepunt te nemen.
- Ook de beoordelingsprocedures bij de assessmentinstrumenten zouden moeten worden gestandaardiseerd. Met name bij *assessment for learning*-assessments is dat bijna onbegonnen werk. Wel kunnen algemeen bruikbare richtlijnen worden ontwikkeld waarvan aannemelijk is dat makers van *assessment-for-learning*-instrumenten zich daaraan zullen conformeren. Voorwaarden daarbij liggen in de sfeer van de eenvoudige toepasbaarheid en duidelijke meerwaarde voor betrouwbaarheid en objectiviteit.

Dit alles betekent dat voor de *assessment of learning* het onderwijsrendement het beste kan worden gemeten met behulp van landelijk uitgerolde, gestandaardiseerde toetsen die alle

verwijzen naar één onderliggende niveau-indeling . Voor de instrumenten ten behoeve van de *assessment for learning* is met name de relatie tussen inhoud van assessment en inhoud van onderwijs van belang. Deze instrumenten zijn ervoor bedoeld om invloed te hebben op de verdere inhoud van het onderwijs, om door middel van feedback bij te dragen aan het leerproces. Ze zijn er niet voor bedoeld en ook niet voor geschikt om onderwijsrendementen in beeld te brengen. De betekenis van deze instrumenten ligt in het praktische gebruik van de resultaten voor het aanpassen van het onderwijs aan de behoeften van de leerlingen.

Ook uit het in paragraaf 3.3 gepresenteerde overzicht blijkt, dat er nauwelijks onderzoek heeft plaatsgevonden naar de vraag hoe *assessment for learning*-instrumenten zo geconstrueerd kunnen worden dat ze voldoen aan de gewenste kenmerken van assessment. Er is ook weinig empirisch onderzoek naar de vraag of deze instrumenten voldoen aan de minimale eisen waaraan een assessment zou moeten voldoen, namelijk betrouwbaarheid en validiteit. Om die reden lijkt het momenteel niet verstandig om instrumenten die bedoeld zijn als *assessment for learning* in te zetten met het oog op de *assessment of learning*-functie.

Daarmee is niet gezegd dat de instrumenten die ingezet worden als, of bedoeld zijn voor *assessment of learning* wél aan de minimale eisen voldoen: de literatuur geeft geen duidelijk antwoord op de vraag of bepaalde assessmentvormen aan de standaardeisen voldaan wordt. Evenmin is hiermee gezegd dat de *assessment-for-learning*-instrumenten geen rol van belang mogen spelen in het onderwijs: die kunnen zeker van nut zijn, maar dan in de functie waarvoor ze bedoeld zijn - om het onderwijs inhoudelijk nauwkeurig af te stemmen op de behoeften van de leerlingen - en niet als summatieve instrumenten.

Het literatuuroverzicht in paragraaf 3.3 geeft een duidelijk maar negatief antwoord op de derde onderzoeksvraag. Die vraag luidde: Wat is er bekend over predictieve validiteit van moderne toetsvormen? In hoofdstuk 3 is al aangegeven dat er meer bespiegelende publicaties dan empirische studies bestaan over de competentiegerichte toetsvormen en *assessment for learning*-instrumenten. In de onderzochte literatuur is er geen onderzoek aangetroffen dat betrekking heeft op predictieve validiteit van taaltoetsen bedoeld voor gebruik in vmbo en mbo. Evenmin is er onderzoek aangetroffen naar de cross-validiteit van instrumenten voor *assessment-for-learning* en *assessment-of-learning.*

5.2 Aanbevelingen voor verder onderzoek

1 Uit de literatuurstudie komt als belangrijkste gemis naar voren: een volledig ontbreken van onderzoek naar de predictieve validiteit van assessments. Dit betreft zowel de *assessment-for-learning* instrumenten die iets kunnen zeggen over de verdere schoolcarrière van leerlingen en mogelijke bedreigingen daarbinnen, als de *assessment-of-learning* instrumenten. Deze laatste zouden iets kunnen zeggen over de te verwachten mate van succes en de eventuele valkuilen van schoolverlaters die de beroepspraktijk ingaan. Over beide konden geen gegevens worden achterhaald. Het ligt voor de hand om op dit vlak minstens verkennend empirisch onderzoek uit te voeren.

2 Evenmin is er iets bekend over de cross-validatie van assessment-instrumenten ter begeleiding van het onderwijs (*assessment-for-learning*) en die ter afsluiting van het

onderwijs (*assessment-of-learning*). Zolang daarover niets bekend is, bevindt onderzoek naar de bruikbaarheid van formatieve toetsen voor summatieve doeleinden zich in een vacuüm. Om die reden achten wij verkennend onderzoek op het terrein van cross-validatie noodzakelijk.

3 De studie van Shute et al. (2008) suggereert dat toevoeging van feedback aan bestaande assessments het leerproces efficiënter kan laten verlopen zonder dat de assessmentfunctie hierdoor wordt aangetast. Dit onderzoek betreft rekenen/algebra. Het verdient aanbeveling om soortgelijk onderzoek te starten naar het effect van feedback bij taalassessments. Wanneer blijkt dat de assessment-component én feedback te zamen op effectieve manier kunnen worden aangeboden, ligt het voor de hand om in vervolg-onderzoek na te gaan voor welke deelgebieden deze aanpak zonder veel moeite gerealiserd kan worden, hoe een dergelijk integraal curriculum het beste kan worden vormgegeven en wat de effecten ervan zijn in vergelijking met de huidige onderwijspraktijk.

4 Het ligt voor de hand om na te gaan of het onderwijs aan kwaliteit wint door het curriculum aan te passen aan een bewezen verwervingsvolgorde. Iets dergelijks gebeurt al langer op het vlak van de woordenschatuitbreiding in het primair onderwijs: daar wordt de Woordenlijst van Schrooten & Vermeer (1994) door uitgevers van schoolboeken gebruikt als basis voor lesmaterialen voor primair onderwijs. Een eerste verkenning van deze aanpak op het gebied van schrijfonderwijs is te vinden in Schuurs (1996). Door met IRT-gecalibreerde opgaven te werken die domeinspecifiek gelabeld zijn, kan per domein de moeilijkheidsgraad en daarmee de kennelijke verwervingsvolgorde worden bepaald. Dit geeft een empirische basis aan de volgorde waarin inhoudelijke elementen in het curriculum aan de orde kunnen worden gesteld.

5 Zoals in hoofdstuk 2 al werd gemeld: de huidige assessment-praktijk is, dat er binnen het taaldomein eerst getoetst wordt op performance. Wanneer deze ontoereikend blijkt, stopt vaak het assessment-proces, zonder dat daarmee duidelijk geworden is wart de oorzaak is van de ontoereikende performance. Wanneer die te wensen overlaat, zou teruggeschakeld moeten worden naar toetsing van noodzakelijke, onderliggende vaardigheden. Wanneer die ontoereikend blijken, zou in een aantal gevallen nog eens teruggeschakeld moeten kunnen worden naar de onderliggende kenniscomponent (vgl. Straetmans 2006).

Aan deze modelmatige opzet kleven wel problemen: zo zullen bij onvoldoende performance de ene keer eerst de onderliggende vaardigheden getoetst moeten worden - als iemand bijvoorbeeld geen fatsoenlijke verkoopbrief kan schrijven, kan worden nagegaan of de formuleervaardigheid in algemene zin wel op het noodzakelijke niveau is. Maar bij onvoldoende performance op een andere taak lijkt het voor de hand te liggen om eerst de benodigde kennis te toetsen - als een kandidaat moeite heeft om een goed verlopend verkoopgesprek te voeren, zal eerst worden nagegaan of de kandidaat wel weet heeft van de opbouw van een dergelijk gesprek in plaats van de spreekvaardigheid in algemene zin te toetsen. Er bestaat kortom weinig inzicht in de verhouding tussen de componenten die te zamen het begrip competentie constitueren. We bepleiten daarom onderzoek naar de samenstelling van de onderscheiden

beroepsgebonden taalgebruikscompetenties, zodat eventueel falen meer gericht kan worden gediagnosticeerd en met instructie en oefening kan worden verbeterd.

5.3 Aanbevelingen voor onderwijspraktijk

De huidige stand van zaken levert aanbevelingen op die gedeeltelijk uiteenlopen voor de twee typen onderwijs.

1 Om het onderwijsrendement te verhogen kan het vmbo meer dan nu het geval is gebruik maken van instrumenten voor *assessment-for-learning*. Het verdient aanbeveling dat peer learning en peer evaluation, maar ook het portfolio in het vmbo vaker worden ingezet. Mits op de juiste manier gebruikt, kan dit leiden tot vergroting van het onderwijsrendement. Bovendien kan het helpen de grote cultuurverschillen tussen vmbo en mbo te verkleinen.

2 Het mbo is ermee gediend dat er weer centrale examens worden ingevoerd. Om gestandaardiseerd te meten is het noodzakelijk om te beschikken over gestandaardiseerde meetinstrumenten die op een meer dan lokaal niveau meten en voor summatieve doeleinden kunnen worden ingezet. Inmiddels is een traject gestart dat ervoor zorgt dat in 2013-2014 een vorm van centrale examinering voor Nederlands en rekenen/wiskunde wordt ingevoerd, gebaseerd op de referentieniveaus van Meijerink. De invoering van deze examens is een stapsgewijs proces. Scholen krijgen eerst de kans om twee jaar ervaring op te doen met de afname van centraal ontwikkelde examens. In 2013/2014 worden scholen verplicht om de centrale examens af te nemen bij leerlingen in het mbo-4. Voor de examinering op mbo niveau-1, -2 en -3 is een implementatieplan in de maak. Daarvoor wordt verkend welke vormen van centrale examinering bij deze doelgroepen passen.

Bovendien zijn er al vanaf voorjaar 2009 voor het mbo diagnostische toetsen beschikbaar die gebaseerd zijn op het referentiekader; voor Nederlands dekken deze toetsen de gebieden lezen, luisteren en taalverzorging (spelling en grammatica). Het ministerie van OCW stelt deze toetsen drie schooljaren gratis beschikbaar; afname van de toetsen is op vrijwillige basis. De grote belangstelling voor deze toetsen (pers. communicatie Aukje Bergsma, Cito) onderstreept de behoefte aan kwaliteitsinstrumenten voor de *assessment for learning*. Daar staat tegenover dat het - gelet op de diversiteit van opleidingen in het mbo - lastig zal zijn om inhoudelijk voldoende draagvlak te vinden voor generieke toetsen over alle mbo-richtingen heen. Het verdient daarom aanbeveling om het veld ervan te overtuigen dat gestandaardiseerde, generieke taaltoetsen een aanwinst kunnen zijn voor álle beroepsopleidingen.

3 Een kernprobleem bij formatief assessment betreft de (mogelijke) geringe kwaliteit van de instrumenten: bij gebrek aan enige controle is er een reële kans dat assessmentinstrumenten een lage betrouwbaarheid hebben en subjectief beoordeeld worden. Bovendien is een acceptabele mate van validiteit niet zonder meer gegarandeerd. Kortom, er is simpelweg geen zicht op de kwaliteit van de formatief ingezette instrumenten. Zolang dit het geval is, verdient het aanbeveling om waar mogelijk naast

zelfgemaakte *assessment-for-learning* instrumenten ook gestandaardiseerde formatieve toetsen in te zetten.

4 Het verdient aanbeveling om assessment-instrumenten zoveel mogelijk te laten verwijzen naar het werk van cie. Meijerink. Dit is een eerste voorwaarde om in elk geval inhoudelijk een gemeenschappelijk referentiepunt te creëren. Waar mogelijk wordt ook de moeilijkheidsgraad van instrumenten op één onderliggende meetlat geprojecteerd op basis van empirische gegevens; wel is duidelijk dat dit laatste nauwelijks haalbaar is voor veel *assessment-for-learning* instrumenten.

5 Het gebruik van portfolio's voorziet in een duidelijke behoefte. De beoordeling ervan is problematisch in allerlei opzichten. Het verdient daarom aanbeveling om leerkrachten duidelijk te maken
- dat portfolio's verschillende doelen dienen en dat afhankelijk van het beoogde doel andersoortige eisen moeten worden gesteld aan de portfolio's;
- dat portfolio's, bijvoorbeeld op basis van werk van Straetmans (2006) en Sluijsmans (2008), beoordeeld kunnen worden op een betrouwbare manier en met behoud van de functies van het portfolio.

6 Met name de veldbevraging heeft duidelijk gemaakt dat leerkrachten niet steeds goed zicht hebben op de verschillende functies van assessment-instrumenten. Het verdient aanbeveling om met begeleiding en bijscholing op dit vlak ervoor te zorgen dat leerkrachten beter weten welke consequenties aan assessments te verbinden, afhankelijk van de functie waarmee de instrumenten worden ingezet.

7 Het verdient aanbeveling om naar analogie van de experimentatie van Shute et al. (2008) na ta gegaan waar bij bestaande en te ontwikkelen assessment-instrumenten feedback kan worden toegevoegd. Liefst worden instrumenten zo gebouwd dat assessment en feedback resp. instructie in elkaars verlengde liggen. Een eerdere Nederlandse poging daartoe is beschreven in Schuurs & Verschoor (2004). Feedback doet aan de toetstechnische waarde van de assessment niet af en heeft een bewezen positief effect op leerrendement.

8 Het verdient aanbeveling om de door cie. Meijerink beschreven Doorgaande Leerlijnen operationeel te maken in de vorm van assessment-instrumenten. Ervaring leert dat toetsonderdelen aan zowel leerlingen als docenten effectief en efficiënt duidelijk kunnen maken welke eisen op een bepaald niveau gesteld worden en hoe de verschillende niveaugroepen in het onderwijs zich tot elkaar verhouden.

Bibliografie

Alderson, J. C. (1981). Report of the discussion on communicative language testing. In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing*. London: British Council.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly, 3*(1), 28.

Alderson, J. C., & Huhta, A. (2005). The Development of a Suite of Computer-Based Diagnostic Tests Based on the Common European Framework. *Language Testing, 22*(3), 20.

Arter, J.A. (2003). Assessment for learning: Classroom assessment to improvestudent achievement and well-being. *ERIC Documents* (2002 ERIC Document Reproduction Service No. ED 480 068).

Assessment Reform Group (2002). *Assessment for Learning: 10 principles*. From http://www.assessment-reform-group.org/CIE3.PDF

Baartman, L.K.J. (2008a). *'Assessing the assessment'. Development and use of quality criteria for Competence Assessment Programmes*. Dissertatie Universiteit Utrecht.

Baartman, L.K.J. (2008b). Assessment of learning outcomes using competence assessment programmes. *International Journal of Psychology, 43*(3-4), 743-743.

Baartman, L., Prins, F., & Kirschner, P. (2007). Kwaliteitsmeting van Competentie Assessment Programma's via zelfevaluatie. *OnderwijsInnovatie*, 10.

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17*(1), 1-43.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2 (1), 1-34.

Bachman, L. F., & Palmer, A. S. (1996). *Language assessment in practice: Developing language tests and justifying their use in the real world*. Oxford: Oxford University Press.

Bharosa, H., Bragt, B., Heins, M. J., Her, C. B., Langeraar, R., Leenders, E. (2008). *Uiterlijke verzorging, toets je taal!* De Bilt: bedrijfstakgroep Uiterlijke Verzorging, MBO Raad.

Biggs, B. T., Hinton, B. E., & Duncan, S. L. S. (1996). Contemporary approaches to teaching and learning. In N. K. Hartley & T. L. Wentling (Eds.), *Beyond tradition: Preparing the teachers of tomorrow's workforce* (pp. 113-146). Columbia: University Council of Vocational Training, Un. of Missouri.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education,* 32(3), 347-364.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessments: In search of qualities and standards* (pp. 13-36). Dordrecht: Kluwer Academic Publishers.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*, 1-74.

Black, P., & Wiliam, D. (1998b). *Inside the black box. Raising standards trhough classroom assessment.* London: Dept. of Education & Professional Studies, Kings College.

Black, P., & Wiliam, D. (2006a). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and Learning* (pp. 9-25). London: SAGE.

Black, P., & Wiliam, D. (2006b). The reliability of assessments. In J. Gardner (Ed.), *Assessment and Learning* (pp. 119-131). London: SAGE.

Black, P., & Wiliam, D. (2006c). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and Learning* (pp. 81-100). London: SAGE.

Boers, T. (2008). Hoe schat een NT2 cursist zijn eigen taalniveau in? En wat is de inschatting door de docent? *Levende Talen, 9*(1), 27-32.

Bohnenn, E., Jansen, F., Kuijpers, C., Thijssen, R., Schot, I., & Stockmann, W. (2007). *Raamwerk Nederlands / Nederlands in (v)mbo-opleiding, beroep en maatschappij*. 's-Hertogenbosch: Cinop.

Bonset, H., & Braaksma, M. (2008). *Het schoolvak Nederlands opnieuw onderzocht. Een inventarisatie van onderzoek van 1997 tot en met 2007*. Enschede: SLO.

Bonset, H., & Ebbers, D. (2007). Nederlands in het vmbo: apart of samen? *Levende Talen, 8*(4), 16-25.

Bonset, H., Ebbers, D., & Malherbe, S. (2006). *Nederlands in het vmbo / Een enquête onder docenten.* Enschede SLO.

Booth, R., Clayton, B., Hartcher, R., Hungar, S., Hyde, P., & Wilson, P. (2003). *The Development of Quality Online Assessment in Vocational Education and Training. Volume 1 [and] Volume 2*: National Centre for Vocational Education Research, Leabrook, Australia.

Borsboom, D., Mellenbergh, G., Van Heerden, J. (2004). The concept of validity. *Psychological Review*, Vol. 111, no. 4, 1061-1071.

Braaksma, M.A.H., Rijlaarsdam, G., Van den Bergh, H., Van Hout-Wolters, B.H.A.M. (2004). Observational learning and its effects on the orchestration of writing processes. *Cognition and Instruction, 22(1)*, 1-36.

Braams, T., Bosman, A.M.T. (2000). Geletterdheid, Fonologische Vaardigheden en Lees- en Spellinginstructie. *Tijdschrift voor* Orthopedagogiek, 39, 199-211.

Brennan, R. (2006). Perspectives on the evolution and future of educational measurement. In R. Brennan (Ed.), *Educational Measurement* (pp. 1-16). Westport: Praeger Publishers.

Broeder, P., & Sorce, R. (2008). Werken met taalportfolio's in het talenonderwijs aan volwassenen. *Levende Talen, 9*(2), 3-10.

Brown, C. A. (2002). Portfolio Assessment: How Far Have We Come? *ERIC Documents* (2002 ERIC Document Reproduction Service No. ED 477941).

Brown, J. D., & Hudson, T. (1998). The Alternatives in Language Assessment. *TESOL Quarterly, 32*(4), 653-675.

Canale, M., & Swain, M. (1981). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. Brennan (Ed.), *Educational Measurement* (pp. 517-530). Westport: Praeger Publishers.

Chen, C.-M., Hong, C.-M., Chen, S.-Y., & Liu, C.-Y. (2006). Mining Formative Evaluation Rules Using Web-Based Learning Portfolios for Web-Based Learning Systems. *Educational Technology & Society, 9*(3), 69-87.

Chiat, S., & Roy, P. (2008). Early phonological and sociocognitive skills as predictors of later language and social communication outcomes. *Journal of Child Psychology and Psychiatry, 49*(6), 635-645.

Cito (2008). *Het schoolexamen in het voortgezet onderwijs / Verslag van een onderzoek naar de kwaliteit van het schoolexamen bij de vakken Engels, Nederlands, biologie en wiskunde.* Arnhem, Cito, december 2008.

Clayton, B., Blom, K., Meyers, D., Bateman, A. *Assessing and certifying generic skills : what is happening in vocational education and training?* Adelaide: National Centre for Vocational Education Research (NCVER), 2003.

Council of Europe (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Craig, D.V. (2001). Alternative, dynamic assessment for second language learners. *ERIC Documents* (2001 ERIC Document Reproduction Service No. 453 691)

Custer, R. L., Schell, J., McAlister, B. D., Scott, J. L., & Hoepfl, M. (2000). *Using Authentic Assessment in Vocational Education*. Information Series No. 381: Center on Education and Training for Employment, Columbus, OH.

Davies, A., & LeMahieu, P. (2003). Assessment for learning: Reconsidering portfolios and research evidence. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessments: In search of qualities and standards* (pp. 141-169). Dordrecht: Kluwer Academis Publishers.

De Bruijn, E. (2007). *Doorleren in de beroepskolom : Product van de kenniskring 'Doorlopende leerwegen in de beroepskolom' in het kader van de Evaluatie Innovatiearrangement Beroepskolom 2003 en 2004*. 's-Hertogenbosch: Cinop.

De Jong, J. H. A. L. (1991). *Defining a variable of foreign language ability: An application of item response theory.* University Twente, Enschede.

De Maa, J. (2007). *Concept Handleiding 'Taal in de Proeve van bekwaamheid' in het mbo*. Amsterdam: ITTA.

Depauw, V., & Vangeneugden, P. (2000). Taakgericht toetsen? Taalvaardigheid meten via zinvolle taken. *VONK, 30*, 11.

Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System, 36*(4), 506-516.

Driessen, M. (2007). Welk scenario kies ik? In *Platform moderne vreemde talen / Nieuwsbrief* jaargang 4, juni 2007, nummer 15.

Dunn, K. E., & Mulvenon, S. W. (2009). *Let's Talk Formative Assessment ... and Evaluation?* Unpublished manuscript.

Dysthe, O., Engelsen, K. S., Madsem, T., & Wittek, L. (2008). A theory-based discussion of assessment criteria: The balance between explicitness and negotion. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 121-131). New York: Routledge.

Eggen, T. J. H. M. (2009). *De kwaliteit van toetsen. Inaugurele rede.* Enschede: Universiteit Twente.

Eggen, T. J. H. M., De Jong, J. H. A. L., Noijons, J., Schuurs, U., & Straetmans, G. (1996). *Adaptief toetsen in de volwasseneneducatie*. Arnhem: CITO.

Eggen, T. J. H. M., & Sanders, P. F. (Eds.). (1993). *Psychometrie in de praktijk*. Arnhem: CITO.

Ehren, P.L.M. *Assessoren competentiegericht middelbaar beroepsonderwijs : over taken en competenties, training en begeleiding, certificering en registratie van assessoren in het mbo*. De Bilt : MBO Raad, 2009.

Evans, L. (2009). Reflective assessment and student achievement in high school English. Dissertation Seattle Pacific University.

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008). *Over de drempels met taal en rekenen.* Enschede: SLO.

Fitzpatrick, J.L., J.R. Sanders, B.R. Worthen (2004). *Program Evaluation / Alternative approaches and practical guidelines*. Boston (Pearson).

Galbraith, D., & Rijlaarsdam, G. (1999). Effective strategies for the teaching and learning of writing. *Learning and Instruction, 9*, 93-108.

Gielen, S., Dochy, F., & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessments: the influence of assessment on learning, including pre-, post-, and true assessment effects. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessments: In search of qualities and standards* (pp. 37-54). Dordrecht: Kluwer Academic Publishers.

Greenan, J.P. (1985). Generizable Communications Skills Assessment. *ERIC Documents* (1985 ERIC Document Reproduction Service 261 255).

Greenleaf, C., Gee, M. K., & Ballinger, R. (1997). Authentic Assessment: Getting Started. *ERIC Documents* (1997 ERIC Document Reproduction Service 411 474).

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A Five-Dimensional Framework for Authentic Assessment. *Educational Technology Research and Development, 52*(3), 67-86.

Gulikers, J., Bastiaens, T., & Kirschner, P. (2006a). Authentic Assessment, Student and Teacher Perceptions: The Practical Value of the Five-Dimensional Framework. *Journal of Vocational Education and Training, 58*(3), 337-357.

Gulikers, J. T. M., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2006b). Relations between Student Perceptions of Assessment Authenticity, Study Approaches and Learning Outcome. *Studies in Educational Evaluation, 32*(4), 381-400.

Gulikers, J., Bastiaens, T., & Kirschner, P. A. (2008a). Defining authentic assessment: Five dimensions of authenticity. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 73-86). New York: Routledge.

Gulikers, J. T. M., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2008b). Authenticity Is in the Eye of the Beholder: Student and Teacher Perceptions of Assessment Authenticity. *Journal of Vocational Education and Training, 60*(4), 401-412.

Gysen, S., Van Avermaat (2005). Issues in functional language performance assessment: the case of the certificate Dutch as a foreign language. *Language Assessment Quarterly* 2 (1), 51-68.

Haertel, E. (2006). Reliability. In R. Brennan (Ed.), *Educational Measurement* (pp. 65-110). Westport: Praeger Publishers.

Harlen, W. (2005). Teachers' Summative Practices and Assessment for Learning -- Tensions and Synergies. *Curriculum Journal, 16*(2), 207-223.

Harlen, W. (2006). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and Learning* (pp. 103-117). London: SAGE.

Harlen, W., & Winter, J. (2004). The development of assessment for learning: learning from the case of science and mathematics. *Language Testing, 21*(3), 390–408.

Hartley, N. K. E., & Wentling, T. L. E. (1996). Beyond Tradition: Preparing the Teachers of Tomorrow's Workforce. Instructional Materials Laboratory, University of Missouri, Columbia.

Havnes, A., & McDowell, L. (2008). Introduction: Assessment dilemmas in contemporary learning cultures. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 3-14). New York: Routledge.

Hendriks, P., & Schoonman, W. (Eds.). (2006). *Handboek assessment deel 1, gedragsproeven*. Assen: Van Gorcum.
Herder, A., & Berenst, J. (2008). Perspectieven op taalbeleid. *Remediaal 6*, 9-14.

Hill, C., & Larsen, E. (1992). Testing and Assessment in Secondary Education: A Critical Review of Emerging Practices: NCRVE Materials Distribution Service, 46 Horrabin Hall, Western Illinois University, Macomb, IL.
Hoogestraat, T. L. (2008). *Validation of PMG's 360-degree feedback process*. Dissertation George Fox University, Newberg, Oregon.
Huisman, J. (2007). *Het metalen scharnierpunt: een doorlopend traject vmbo-mbo voor metaal/metalektro; tussenstand*. 's-Hertogenbosch: CINOP.
Huisman, J., & Pijnenburg, A. (2009). *Het Metalen Scharnierpunt / Het werken met praktijkopdrachten*. 's-Hertogenbosch Cinop.
Hunt, M., Neill, S., & Barnes, A. (2007). The Use of ICT in the Assessment of Modern Languages: The English Context and European Viewpoints. *Educational Review, 59*(2), 195-213.
Inger, M. (1993). *Authentic Assessment in Secondary Education*. New York, Columbia University: Institute on Education and the Economy.
Inspectie van het Onderwijs (2003). *Zicht op toetsen / Toetsing en examinering in het hoger onderwijs*. Den Haag.
Inspectie van het Onderwijs (2004). *Onderwijsverslag 2002/2003.* Utrecht: Inspectie van het Onderwijs.
Inspectie van het Onderwijs (2006). *Nederlands in het mbo*. Den Haag: Inspectie van het Onderwijs.
Inspectie van het Onderwijs (2007). *Basisvaardigheden taal in het voortgezet onderwijs / Resultaten van een inspectieonderzoek naar taalvaardigheid in de onderbouw van het vmbo en praktijkonderwijs*. Den Haag: Inspectie van het Onderwijs.
Johnson, S. D., & Wentling, T. L. (1996). An alternative vison for assessment in vocatonal teacher education. In N. K. Hartley & T. L. Wentling (Eds.), *Beyond tradition: Preparing the teachers of tomorrow's workforce* (pp. 147-166). Columbia: Un. of Missouri.
Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Westport: Praeger Publishers.
Kavaliauskiene, G., Kaminskiene, L., & Anusiene, L. (2007). Reflective Practice: Assessment of Assignments in English for Specific Purposes *Iberica, 14*, 149-166.
Kelsey, K. D. E. (2001). Theme: Evaluating Learning in Technical Agriculture. *Agricultural Education Magazine, 73*(5), 4-23.
Kenyon, D., & Van Duzer, C. (2003). *Valid, Reliable, and Appropriate Assessments for Adult English Language Learners. ERIC Q & A*: National Center for ESL Literacy Education, Center for Applied Linguistics, N.W., Washington.
Klarus, R. (2000). Beoordeling en toetsing in het nieuwe onderwijsconcept. In: Onstenk. J. (red.). *Op zoek naar een krachtige beroepsgerichte leeromgeving. Fundamenten voor een onderwijsconcept voor de bve-sector*. Cinop, Den Bosch.
Klein, L. (2007). Auto-evaluation: Daily Self-Assessment in the FSL Classroom. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes, 64*(1), 181-198.

Koopmans, H. J. M. (2006). *Professionals organiseren informeel leren. Onderzoek naar het organiseren van informeel leren door professionals en de wijze waarop managers en opleidingskundigen dat kunnen stimuleren*. Delft: Uitgeverij Eburon.

Kuhlemeier, H. (Ed.). (2005). *Competentiegericht leren en beoordelen in vmbo en mbo. Ontwikkelingen, knelpunten en oplossingsrichtingen tussen theorie en praktijk*. Utrecht: WVOI.

Land, J., Sanders, T., & Van den Bergh, H. (2006). *Wat maakt een studietekst geschikt voor vmbo-leerlingen?* Amsterdam: Stichting Lezen.

Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational Measurement* (pp. 387-431). Westport: Praeger Publishers.

Lantolf, J. P., & Poehner, M. E. (2008). Dynamic assessment. In E. Shohamy & H. Hornberger (Eds.), *Encyclopedia of language and education (2nd edition).* Volume 7: Language testing and assessment (pp. 273–284). New York: Springer Science.

Laurier, M. (2000). Can computerised testing be authentic? ReCALL, 12(1), 93-104.

Laurier, M. (2004). Evaluation and multimedia in second-language learning. *ReCALL, 16*(2), 475-487.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom Assessment: Minute by Minute, Day by Day. *Educational Leadership, 63*(3), 18-24.

Lee, I. (2007). Feedback in Hong Kong Secondary Writing Classrooms: Assessment for Learning or Assessment of Learning? *Assessing Writing, 12*(3), 180-198.

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving Learners and Their Judgements in the Assessment Process. *Language Testing, 22*(3), 321-336.

López Bonilla, G., Rodríguez, M. (2003). Alternative Assessment: Opportunities and Challenges in Evaluating Reading. *Mexican Journal of Educational Research* 8 (17), January / April 2003, pp. 67-98.

Lynch, B. K. (2001). Rethinking Assessment from a Critical Perspective. *Language Testing, 18*(4), 351-372.

Mabry, L. (1999). *Portfolios Plus. A critical guide to altenative assessment*. Thousand Oaks, California: Corwin Press.

Marzano, R., & Miedema, W. (2008). *Leren in vijf dimensies. Moderne didactiek voor het voortgezet onderwijs*. Assen: Van Gorcum.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75-100.

McMillan, J.H. (2001). *Classroom assessment. Principles and practice for effective instruction*. 2nd Edition. Needham Heights: Ally & Bacon.

Messick, S. 1989: Validity. In Linn, R.L. (ed.), *Educational measurement*. 3rd edn. New York: American Council on Education/Macmillan, 13–103.

Meuffels, B., & Maat, K. (2009). Toetsoptimalisatie onder moeilijke omstandigheden. *Tijdschrift voor Taalbeheersing, 31*(1), 18-38.

Murray, S. (2008). Summative assessment: a historical perspective. *British journal of general practice*, 2 , 894-895.

Neuvel, J., Bersee, T., Exter, H. d., & Tijssen, M. (2004). *Nederlands in het middelbaar beroepsonderwijs. Een verkennend onderzoek naar het onderwijsaanbod Nederlands en de taalvaardigheid van de leerlingen*. 's-Hertogenbosch: Cinop.

Neuvel, J., & Esch, W. v. (2006). *De doorstroom van vmbo naar mbo. Het effect van het beroepsbeeld en de toepassing van de doorstroomregeling op de schoolloopbaan in het mbo*. 's-Hertogenbosch CINOP.

Nijhof, W. (2006). *Naar 'nieuwe' examineringsvormen in het MBO?* Bezo Consult.

Orr, S. (2008). Real or imagined? The shift from norm referencing to criterion referencing in higher education. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 133-143). New York: Routledge.

Pellegrino, J. (2008). Assessment for learning: Using assesment formatively in classroom instruction. *International Journal of Psychology, 43*(3-4), 381-381.

Poehner, M. E., & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research, 9*(3), 22.

Popham, W. J. (2006). Assessment for Learning: An Endangered Species? *Educational Leadership, 63*(5), 82-83.

Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing, 25*(1), 127-143.

Prodromou, L. (1995). The backwash effect: from testing to teaching. *ELT Journal, 49*(1), 13-25.

Reinders, H., & Lazaro, N. (2007). Current approaches tot assessment in self-access language learning. *TESL-EJ: Teaching English as a Second or Foreign Language, 11*(3), 13.

Roelofs, E., & Straetmans, G. (Eds.). (2006). *Assessment in actie. Competentiebeoordeling in opleiding en beroep*. Arnhem: Cito.

Ross, S. (1998). Self-assessment in second language testing : A meta-analysis and analysis of experiential factors. Language testing 15, 1, 1-20.

Ross, S. J. (2005). The Impact of Assessment Method on Foreign Language Proficiency Growth. *Applied Linguistics, 26*(3), 317-342.

Rowlands, B. 2001, *Good practice in online education and assessment,* Department of Education and Training, NSW TAFE, Information Technology, Arts and Media Division, Sydney.

Ruijters, M. (2007). 'Goh, het lijkt net werk…' Het organiseren van informeel leren. *Leren in Organisaties, 7*(12), 14-18.

Rijlaarsdam, G., Braaksma, M., Couzijn, M., Janssen, T., Raedts, M., Van Steendam, E., Toorenaar, A., Van den Bergh, H. (2008). Observation of peers in learning to write, Practice and Research. *Journal of Writing Research*, 1 (1), 53-83.

Schrooten, W., Vermeer, A. (1994). *Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen.* Tilburg (Tilburg University Press), 1994.

Schuurs, U. (1993). STAAL: een nieuwe NT2-toets op brugklasniveau. *Levende Talen, 482*, 382-387.

Schuurs, U. (1996). Sequencing writing course content on the basis of test difficulty. In: Rijlaarsdam, G., H. van den Bergh, M. Couzijn (eds.): *Effective teaching and learning of writing.* Amsterdam Un. Press 1996, p. 300-311.

Schuurs, U. (2001). Verschillende NT2-toetsen, één meetlat. *LES, Tijdschrift voor docenten aan (jong) volwassen anderstaligen,* jaargang 19, nummer 110 (april 2001).

Schuurs, U., Verschoor, A. (2004). Paper presentation *'Formative assessment using feedback modes within an electronic learning system'.* EARLI Conference on Assessment (Bergen, Norway, June 2004, zie www.assessment2004.uib.no/).

Segers, M., Dochy, F., & Cascallar, E. (2003). The era of assessment engineering: Changing perspectives on teaching and learning and the role of new modes of assessment. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessments: In search of qualities and standards* (pp. 1-12). Dordrecht: Kluwer Academic Publishers.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., et al. (2008). On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers. *Applied Measurement in Education, 21*(4), 295-314.

Shepard, L. A. (2009). Commentary: Evaluating the Validity of Formative and Interim Assessment. *Educational Measurement: Issues and Practice, 28*(3), 32-37.

Shute, V. J., Hansen, E. G, Almond, R. G. You Can't Fatten A Hog by Weighing It – Or Can You? Evaluating an Assessment for Learning System Called ACED. *International Journal of Artificial Intelligence in Education*. Vol.18(4), 2008, pp. 289-316.

Sluijsmans, D. M. A., Straetmans, G. J. J. M., & van Merrienboer, J. J. G. (2008). Integrating Authentic Assessment with Competence-Based Learning in Vocational Education: The Protocol Portfolio Scoring. *Journal of Vocational Education and Training, 60*(2), 159-172.

Smith, K., & Tillema, H. (2007). Use of Criteria in Assessing Teaching Portfolios: Judgemental Practices in Summative Evaluation. *Scandinavian Journal of Educational Research, 51*(1), 103-117.

Smith, K., & Tillema, H. (2008). The challenge of assessing portfolios: In search of criteria. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 183-195). New York: Routledge.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2*(1), 31-40.

Spolsky, B. (2008). Language assessment in historical and future perspective. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education, 2nd edition,* Volume 7: Language Testing and Assessment, 445–454. Springer Science+Business Media LLC.

Stapleton, P. (2006). Critiquing Research Methodology: Comments on Broader Concerns about Complex Statistical Studies. A Response to Ross. *Applied Linguistics, 27*(1), 130-134.

Stiggins, R. (2005a). *Student-involved assessment FOR learning*. New Jersey: Pearson Education.

Stiggins, R. (2005b). From formative assessment to assessment FOR learning: A path to success in standard-based schools. *Phi Delta Kappan, 87*(4), 324-329.

Stiggins, R. (2009). Assessment FOR Learning in Upper Elementary Grades. *Phi Delta Kappan, 90*(6), 419-421.

Stobart, G. (2006). The validity of formative assessment. In J. Gardner (Ed.), *Assessment and Learning* (pp. 133-146). London: SAGE.

Stoel, D. (2006). Portfolio Rationalisatie / Upgrade uw opleidingen. *Leren in organisaties, 12* (dec), 25-27.

Straetmans, G. J. J. M. (2004). *Protocol portfolio scoring. Een methode voor het systematisch scoren en vaststellen van competenties*. Arnhem: Cito.

Straetmans, G. J. J. M. (2005). *Kijk op competenties. De belangrijkste wetenswaardigheden over competenties en het beoordelen daarvan*. Arnhem: Cito.

Straetmans, G. J. J. M. (2006). *Bekwaam beoordelen en beslissen. Beoordelen in*

competentiegerichte beroepsopleidingen (lectorale rede Saxion Hogescholen).
Struyven, K., Dochy, F., & Janssens, S. (2008). Students' likes and dislikes regarding student-activating and lecture-based educational settings: Consequences for students' perceptions of the learning environment, student learning and performance. *European Journal of Psychology of Education, 23*(3), 295-317.
Studentenhandleiding; Leerlijn SLB. (2009). Bron: http://mbo2010.kennisnet.nl/bronnen/details/24,1126/studentenhandleiding-leerlijn-slb.
Tan, K. (2008). Academics´ and academic developers´ views of student self-assessment. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 225-236). New York: Routledge.
Toorenaar, A., & Rijlaarsdam, G. (2007). 'Een spiekbriefje leren maken?!' Sommige schoolboeken zijn gewoon te moeilijk. *Levende Talen Magazine, 94*(2), 17-19.
Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 28.
Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessments: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academis Publishers.
Van Batenburg, Th. A., Van der Werf, M.P.C. (2004). *NSCCT Niet Schoolse Cognitieve Capaciteiten Test voor groep 4, 6 en 8 van het basisonderwijs. Verantwoording, normering en handleiding*. Groningen (GION).
Van de Watering, G., Gijbels, D., Dochy, F., & van der Rijt, J. (2008). Students' assessment preferences, perceptions of assessment and their relationships to study results. *Higher Education, 56*(6), 645-658.
Van den Bergh, H. Toetsen van taalvaardigheid. *Nieuwsbrief Moderne Vreemde Talen*, jaargang 4, juni 2007, nummer 15. Bron: http://www.cinop.nl/ezine/mvt/nb_07-15/ez_mvt_15.htm)
Van den Bergh, R.H., & Bleichrodt, N. (2000). Intelligentiemeting bij kandidaten met verschillende culturele achtergronden: de Multiculturele Capaciteiten Test (MCT-M). *Nederlands Tijdschrift voor de Psychologie*, 55, 134-147.
Van den Berg, N., De Bruijn, E. (2009). Het glas vult zich. Kennis over vormgeving en effecten van competentiegericht beroepsonderwijs; verslag van een review. Ecbo 's-Hertogenbosch.
Van den Brink, W. P., & Mellenbergh, G. J. (1998). *Testleer en testconstructie*. Amsterdam: Boom.
Van der Sanden, J. M. M., Streumer, J. N., Doornekamp, B. G., Hoogenberg, I., & Teurlings, C. C. J. (2003a). *Praktijksimulaties in het vernieuwend vmbo. Bouwstenen voor de integratie van theorie en praktijk*. Utrecht: APS.
Van der Sanden, J. M. M., van der Os, M. J. M., & Kok, H. (2003b). *Naar aantrekkelijk vmbo. Resultaten van drie jaar herontwerp*. Den Haag: Opmeer.
Van Driel, J. (2004). *Alle schoolboeken de deur uit? Ontwikkelreeks, deel 1.* Ede (Ontwikkelcentrum).
Van Gelooven, D., Veldkamp, B. (2006). Beroepsbekwaamheid van weginspecteurs: een virtual reality toets. In Roelofs, E., & Straetmans, G. (Eds.). (2006). *Assessment in actie. Competentiebeoordeling in opleiding en beroep (pp. 93-123)*. Arnhem: Cito.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One Parameter Logistic Model (OPLM)*. Arnhem: CITO.

Verhoeven, L., Vermeer, A. (2003). Zin en onzin van toetsen bij de aanvang van het basisonderwijs: Predictieve validiteit van taal- en intelligentiepeiling bij autochtone en allochtone kleuters. In: T. Koole, J. Nortier, & B. Tahitu (eds.), *Artikelen van de vierde sociolinguïstische conferentie*. Delft: Eburon, 532-541.

Visscher, A. (2008). *Onderzoeksverslag 'Taalbeleid en taaltoetsing op het mbo' in opdracht van Diataal*. Groningen: Expertisecentrum Taal, Onderwijs en Communicatie, Rijksuniversiteit Groningen.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing. A Primer / Second edition*. New York: Hillsdale.

Weir, C. J. (2005). Limitations of the Common European Framework for Developing Comparable Examinations and Tests. *Language Testing, 22*(3), 281-300.

Whitehead, D. (2007). Literacy Assessment Practices: Moving from Standardised to Ecologically Valid Assessments in Secondary Schools. *Language and Education, 21*(5), 434-452.

Wiggins, G. (1989). A true test: toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703 – 713.

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing, 26*(3), 445-466.

Wiliam, D. & Black, P. (1996). Meaning and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537-548.

Wiliam, D. (2006). Balancing dilemmas: Traditional theories and new applications. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 267-281). New York: Routledge.

Wolf, K., Lichtenstein, G., & Stevenson, C. (1997). Portfolios in teacher education. In J. H. E. Strange (Ed.), *Evaluating teaching: a guide to current thinking and best practice* (pp. 193-214). Thousands Oaks, CA: Corwin Press Inc.

Wools, S. (2009). Is dit assessment kwalitatief goed genoeg? De eerste stappen in de ontwikkeling van een beoordelingsinstrument voor competentie assessment. *Examens*(4).

Wools, S., Sanders, P., & Roelofs, E. (2007). Beoordelingsinstrument: Kwaltiteit van Competentie Assessment. Arnhem: Cito.

Wools, S., Sanders, P. F., Eggen, T. J. H. M., Baartman, L. K. J., & Roelofs, E. (te verschijnen). Evaluatie van een beoordelingssysteem voor de kwaliteit van competentie-assessments.

Wragg, E. C. (2001). *Assessment and learning in the secondary school*. London: Routledge.

Bijlagen

1) Overzicht van veelgebruikte taaltoetsen in vmbo en mbo
2) Vragenlijst voor veldbevraging
3) Geraadpleegden

Bijlage 1: Overzicht van veelgebruikte taaltoetsen in vmbo en mbo

In het Raamwerk Nederlands zijn de talige streefniveaus voor mbo uitgewerkt, met verwijzing naar het Europese Referentiekader ofwel Common European Framework . De samenstellers van het Raamwerk hebben afgezien van uitwerking van niveau A1 en van niveau C1 :

- Voor A1 geldt dat leerlingen met dit vaardigheidsniveau niet kunnen functioneren in het vmbo
- Voor C1 geldt dat een taalgebruiker op dit niveau op academisch niveau functioneert. Dit niveau komt normaliter binnen het (v)mbo niet voor.

De meeste toetsen voor vmbo en mbvo conformeren zich aan deze zienswijze.

Onderstaand is weergegeven welke toetsen momenteel in gebruik zijn op vmbo en mbo

I Vmbo

1 Er zijn voor bovenbouw v.o. en voor het mbo **Diagnostische toetsen Taal** in ontwikkeling bij Cito. Deze toetsen zijn gebaseerd op het Common European Framework (CEF). Voor Nederlands dekken deze toetsen de gebieden lezen, luisteren en taalverzorging (spelling en grmamatica). Het ministerie van OCW stelt deze toetsen drie schooljaren gratis beschikbaar; afname van de toetsen is op vrijwillige basis. Er is erg veel belangstelling voor dit materiaal voor *assessment for learning*.

2 Er is bij Cito een **Toelatingstoets lwoo en pro** ten behoeve van plaatsing in leerweg-ondersteunend onderwijs of praktijkonderwijs (vmbo). Met deze Toelatingstoets is vast te stellen welke leerlingen in aanmerking komen voor plaatsing in het leerwegondersteunend onderwijs of het praktijkonderwijs. De Toelatingstoets is speciaal gemaakt voor afname bij zorgleerlingen. De toets bevat de wettelijk voorgeschreven onderdelen van het didactische deel van de vereiste toetsing:

- Begrijpend lezen (drie taken)
- Technisch lezen (één taak)
- Spelling (twee taken)
- Inzichtelijk rekenen (twee taken)

De Toelatingstoets lwoo en pro is klassikaal af te nemen; per taak is de afnametijd ongeveer 20 minuten (bij technisch lezen 6 minuten). De taken voor begrijpend lezen en spelling bevatten uitsluitend meerkeuzevragen. De leerlingen maken deze taken op een antwoordblad. De Toelatingstoets LWOO en PO is door de Cotan beoordeeld en heeft op alle relevante punten het oordeel ‘goed’ gekregen.

3 Het **Cito Volg- en AdviesSysteem (VAS) voor het v.o.** heeft taaltoetsen leesvaardigheid en woordenschat op drie niveaus (entreetoets, volgtoets eind jaar 1, een adviestoets halverwege tweede leerjaar en toets 3, eind 3^{e} leerjaar). In hoeverre er gebruik van wordt gemaakt op vmbo en mbo is onbekend.

4 **Diataal** van het ETOC (Hilde Hacquebord) is een taaltoetspakket voor groep 7 en 8 van het basisonderwijs en de eerste drie klassen van het voortgezet onderwijs. Met de online af te nemen toetsen van Diataal kan een diagnostisch taalvaardigheidsonderzoek van de leerlingen plaatsvinden. Diataal bestaat uit:
   - de adaptieve tekstbegriptoets Diatekst, bruikbaar in primair onderwijs en v.o. t-m klas 3
   - de woordenschattoets Diawoord, voor primair onderwijs en v.o. klas 1
   - de luistertoets Diafoon, voor v.o. klas 1
   - Diaplus materiaal voor taalondersteuning, voor primair onderwijs en v.o. klas 1 en 2

   De toetsen zijn adaptief en worden online afgenomen, zodat de leerkracht geen nakijkwerk heeft. Genoemde onderdelen geven naast de individuele score ook de prestatie weer ten opzichte van leerlingen van hetzelfde onderwijstype (vaardigheidsgroep): de toetsen van Diataal maken duidelijk hoe de leerling scoort ten opzichte van de klasgenoten. Ook wordt het duidelijk waar de problemen precies liggen als een leerling slecht scoort. Het programma biedt bovendien achtergrondinformatie en oefenmateriaal voor taalondersteuning.
   De toetsen van Diataal zijn niet verbonden aan de CEF-niveaus; wellicht is dat de reden dat Diataal nauwelijks gebruikt wordt in het mbo.

5 Zelfontworpen toetsen: er zijn veel scholen die zelf ook taaltoetsen ontwerpen en deze intern gebruiken. Hierover is verder nauwelijks wat bekend.

II MBO

1 De **TOA** (Toolkit Onderwijs en Arbeid) van bureau ICE (Bureau Interculturele Evaluatie) is een 'gereedschapskist vol met toetsinstrumenten'. Met de verschillende toetsinstrumenten kan er online het niveau van de taal-, reken- en studievaardigheid vastgesteld worden en kunnen de competenties gemeten worden die horen bij de kerntaken voor Leren, Loopbaan en Burgerschap. Bijvoorbeeld voor een intake of tijdens of na afloop van een studietraject. Per kandidaat kan worden bepaald welke toetsen voorgelegd worden. Er zijn geen vaste combinaties en alle onderdelen zijn los te selecteren.

De Toolkit is gekoppeld aan het CEF. De toolkit bevat taaltoetsen op alle niveaus van het CEF voor lezen, luisteren, schrijven, spreken en gespreksvaardigheid (A1 t/m C1) voor drie talen: Nederlands, Engels en Duits. Voor de zeven kerntaken voor leren, loopbaan- en burgerschap zoals die geformuleerd zijn voor het mbo (brondocument Leren, Loopbaan en Burgerschap) bevat de TOA een intaketoets (één toets voor alle zeven kerntaken) en zeven examens (één per kerntaak). *(bron: website www.bureau-ice.nl)*

2 De **TNT** (TaalNiveauTest) van uitgeverij Deviant is een online instrument om het niveau Nederlands van deelnemers te bepalen op de schaal van het CEF, van A1 t/m C2. TNT bestaat uit 192 taaltaken die uitgevoerd kunnen worden. Er zijn vier onderdelen (luisteren, lezen, taalstructuur en woordkennis) en voor elk onderdeel zijn maximaal 48 taaltaken beschikbaar (dit is gelijk aan 8 taaltaken per CEF-niveau per vaardigheid). Een leerling die bijvoorbeeld op niveau A1 zit, maakt de hierbij behorende taaltaken en dan blijkt of de leerling voldoende heeft gescoord. Als dit het geval is, kan de leerling verder met taaltaken van niveau A2. De taaltaken zijn de zogenaamde toetsvragen. Wanneer in elke CEF-categorie 75% goed is geantwoord voor elke vaardigheid, wordt het predicaat 'passend in de betreffende CEF- categorie' toegekend.

3 **Methode VIA** is een complete methode van uitgeverij Deviant waarin ook toetsen opgenomen zijn. Het gaat hierbij om summatieve taalopdrachten waarin de vijf taalvaardigheden (lezen, luisteren, spreken, gesprekken voeren en schrijven) per CEF-niveau geïntegreerd worden aangeboden binnen het taalprofiel van de opleiding. Elk van deze toetsen kent een verantwoording naar kerntaak en werkproces op de verschillende CEF-niveaus en is voorzien van scoremodel, normering en bewijslast.

4 **Taalblokken** is een interactief taalsysteem voor het mbo, uitgegeven door uitgeverij Malmberg. Hierin zijn drie soorten toetsen verwerkt:
- de 'Bepaal je niveau-toetsen' zijn instaptoetsen. Deze geven een beeld van het niveau waarop de leerling op dat ogenblik zit;
- een 'Test jezelf-toets' die helpt bepalen of de taalbeheersing van een leerling op het gewenste CEF-niveau is;
- met de 'Checklists taalniveau' kan de leerling het eigen taalniveau inschatten middels can-do statements (self-assessment).

Taalblokken is gerelateerd aan de CEF-niveaus en de toetsen zijn allemaal gespecificeerd naar de sectoren:

- Economie
- Economie, uitgebreid met extra taaltaken voor Horeca, Toerisme en Recreatie
- Techniek
- Zorg en Welzijn

5 **Taalperfect** is een online oefen- en assessmentsysteem, ontwikkeld door enkele docenten aan het Zoomvliet College. Het biedt oefenmogelijkheden voor woordenschatuitbreiding, werkwoordsspelling, zinsontleding wen de verbetering van de schrijfvaardigheid (voegwoordgebruik, veel voorkomende stijlfouten, spreekwoorden). DE didactische aanpak is divers: zo wordt bij de woordenschatmodule WoordPerfect de helft van de woorden in een betekenisvolle context aangeboden, de andere puur op woord en betekenis. Het systeem houdt voor elke deelnemer bij welke woorden deze kent. Woorden die nog niet tot de woordenschat behoren, worden met een hogere frequentie aangeboden. Doordat het systeem gelinkt is aan het CEF, kan ook een indicatie van het taalvaardigheidsniveau worden verkregen.

6 Zelfontworpen toetsen: er zijn veel scholen die zelf ook taaltoetsen ontwerpen en deze intern gebruiken. Hierover is verder nauwelijks wat bekend.

7 Cito ontwikkelt in samenspraak met de AOC-raad en enkele AOC's afsluitende taaltoetsen voor Lezen en Luisteren op niveau B1; deze vaardigheden worden als indicatief beschouwd voor de overige vaardigheden. De toets Lezen is gebaseerd op Raamwerk Nederlands en is geprest. De toets Luisteren betreft vooralsnog een voor de gelegenheid samengestelde toets waarvoor Cito de opgaven beschikbaar heeft gesteld. De toetsen worden online afgenomen met behulp van Question Mark Perception.

8 Cito TaalSchaal Lezen Nederlands, vanaf 2010 beschikbaar, zijn digitale taaltoetsen voor mbo niveau 1 en 2 waarmee de leesvaardigheid Nederlands op niveau A2 en B1 in kaart gebracht wordt.
TaalSchaal Lezen bestaat uit twee toetsen; een intake- en een voortgangstoets. De intaketoets kan in het eerste leerjaar de basisvaardigheid lezen van de deelnemers in kaart brengen. Op deze manier zijn zwakke lezers op te sporen zodat bijscholing op maat kan worden aangeboden. Vervolgens kan de voortgangstoets TaalSchaal Lezen de vorderingen in beeld brengen. Combinatie van de resultaten van de intake- en de voortgangstoets geeft inzicht in het rendement van uw leesonderwijs. De scores op TaalSchaal Lezen geven een indicatie van de leesvaardigheid van een deelnemer ten opzichte van niveau A2 en B1 van het Raamwerk Nederlands. Naast de individuele score wordt ook de landelijke score gegeven.

Bijlage 2: Vragenlijst voor veldbevraging

Beste docent,

Via deze vragenlijst willen we u enkele vragen stellen over taaltoetsen. Er bestaan verschillende soorten toetsen, die allemaal op een ander aspect van taalvaardigheid betrekking hebben. Ook worden er verschillende eisen aan toetsen gesteld. We zijn benieuwd naar uw mening hierover en naar de praktijksituatie in uw geval. We leggen deze vragenlijst voor aan docenten en ook aan een aantal toetsdeskundigen die niet in het onderwijs werkzaam zijn. Het invullen van de vragenlijst neemt ongeveer vijftien minuten in beslag.

Aan het eind van de vragenlijst kunt u aangeven of u een samenvatting van de resultaten wilt ontvangen.

Alvast hartelijk dank voor uw medewerking!

Uriël Schuurs
Expertisecentrum Nederlands

Schoolgegevens
Naam school:
Plaats school:

Schooltype …….. **Eventueel: Sector ……..**

Neem bij het beantwoorden van de vragen het leerjaar of de klas in gedachten waaraan u het meest lesgeeft.

Ik denk aan klas/leerjaar ……

Vragen over verschillende soorten toetsen

Welke soorten taaltoetsen gebruikt u gedurende dit schooljaar in deze klas? (meerdere antwoorden mogelijk)

0 Summatieve toets, om de leerprogressie aan het eind van een onderwijsperiode vast te stellen
0 Formatieve toets, om het onderwijs tussentijds te kunnen bijstellen op basis van de toetsresultaten
0 Plaatsingstoets, om leerlingen naar niveau in (sub-)groepen te plaatsen
0 Diagnostische toets om sterke en zwakke kanten van taalbeheersing in kaart te brengen
0 Voortgangstoets om te bepalen of een leerling voldoende vooruitgaat
0 Overgangstoets om te bepalen wie naar een volgend leerjaar of onderwijsblok mag
0 Examenonderdeel om formeel af te sluiten en te certificeren
0 Anders, namelijk ……..

Verschillende toetsfuncties

Hoe belangrijk vindt u taaltoetsen voor deze verschillende functies (omcirkel één cijfer per toets)?

Summatieve toets	zeer onbelangrijk	1	2	3	4	zeer belangrijk
Formatieve toets	zeer onbelangrijk	1	2	3	4	zeer belangrijk
Plaatsingstoets	zeer onbelangrijk	1	2	3	4	zeer belangrijk
Diagnostische toets	zeer onbelangrijk	1	2	3	4	zeer belangrijk
Voortgangstoets	zeer onbelangrijk	1	2	3	4	zeer belangrijk

Overgangstoets	zeer onbelangrijk	1	2	3	4	zeer belangrijk
Examenonderdeel	zeer onbelangrijk	1	2	3	4	zeer belangrijk
Door u genoemde toets	zeer onbelangrijk	1	2	3	4	zeer belangrijk

Verschillende soorten taaltoetsen

Toetsen kunnen nogal verschillen naar verschijningsvorm.
Wilt u aangeven welke soort taaltoetsen u zoal gebruikt in de klas die in gedachten heeft genomen, en hoe vaak u daar gebruik van maakt?

bijna nooit 1 2 3 4 erg vaak

1 Kennistoets: Een toets met vragen over beroepsgerichte, theoretische en vakgerichte kennis.

2 Essaytoets: Een schriftelijke toets die bestaat uit één of meer open vragen die lange antwoorden uitlokt.

3 Vaardigheidstoets: Een toets die controleert of de leerling bepaalde (beroepsmatige) vaardigheden correct en adequaat kan uitvoeren.

4 Casustoets: Een probleem of gevalsbeschrijving, ontleend aan de beroepspraktijk, waar de leerling een respons op moet geven.

5 Voortgangstoets: Een kennistoets of vaardigheidstoets waarmee wordt gemeten of een leerling voldoende vooruitgaat.

6 Peer assessment: Een beoordeling van een product, prestatie of proces waarbij medeleerlingen de beoordeling geven.

7 Self assessment: Een beoordeling van een product, prestatie of proces waarbij de leerling de prestaties van zichzelf beoordeelt.

8 Stage- en praktijkopdracht: een opdracht die de leerling uitvoert voor en bij een instelling of bedrijf, met een eindverslag en eventueel een eindproduct als resultaat.

9 Projectopdracht: Een bedrijf, organisatie of instelling heeft een probleem of vraag die de leerling met een groep of alleen dient op te lossen of te beantwoorden. De kwaliteitseisen die aan het product worden gesteld, moeten in toenemende mate door de leerling zelf worden verantwoord.

10 Gedragsassessment: Een integrale toetsing, waarin de leerling in kritische beroepssituaties taken moet uitvoeren en daarbij het gewenste competentieniveau kan demonstreren.

11 Afstudeeropdracht: Een eindopdracht ter afronding van de opleiding als proeve van bekwaamheid. Een 'meesterproef' die de leerling geheel zelfstandig uitvoert.

12 Portfolio assessment: Een gesprek over de competentieontwikkeling met de leerling. Een beoordelingsgesprek of een voortgangsgesprek tussen de leerling en de beoordelaar(s). Het gesprek kan de vorm hebben van een criteriumgericht interview.

13 Criteriumgericht interview: Een gestructureerd gesprek over (beroeps)situatie(s) waarin de leerling zijn/haar competenties laat zien. Deze toets wordt ingezet bij portfolio assessment, gedragsassessment en intaketoets.

14 Reflectieopdracht: een werkstuk waarin de leerling zelfkritisch verslag doet van eigen ervaringen in bepaalde leer- en beroepssituaties.

Eisen die aan formatieve taaltoetsen worden gesteld

Dit deel van de vragenlijst gaat alleen over formatieve taaltoetsen, dus over toetsen waarmee tussentijds kan worden gemeten of het onderwijs bijstelling behoeft.

We noemen twintig eisen die vaak aan toetsen gesteld worden. Kunt u per eis aangeven hoe belangrijk u die vindt?

1. De taken in de taaltoets komen overeen met taken in de praktijk.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

2. De taken in de taaltoets zijn competentiegericht: kennis, vaardigheid, houding geïntegreerd.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

3. De moeilijkheidsgraad van de taaltoets is vergelijkbaar met de moeilijkheidsgraad van taken in de praktijk.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

4. De toetsscore geeft een goed beeld van het prestatieniveau in de praktijk.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

5. De prestaties op de toets wordt vergeleken met van tevoren vastgestelde kwaliteitseisen.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

6. Leerlingen zijn op de hoogte van wat er van ze verwacht wordt bij de taaltoets.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

7. De toets meet daadwerkelijk wat er gemeten moet worden (validiteit).

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

8. De beoordeling van prestaties op de toets is beoordelaars-onafhankelijk en/of wordt door meer beoordelaars gedaan.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

9. De toets heeft een *aangetoonde* hoge mate van betrouwbaarheid.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

10. De toetsdoelen komen overeen met de onderwijsdoelen.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

11. De leerling kan de taaltoets doen op het moment dat hij of zij daaraan toe is.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

12. De taaltoets geeft inzicht in het handelen van de leraar en is leerzaam voor de leerling.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

13. De toets is leerwegonafhankelijk: het maakt niet uit waar de leerling de competenties heeft verworven – op school, op de werkplek of elders.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

14. De kosten en tijd van de afname staan in verhouding tot de opbrengst van de toets.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

15. De toets levert een bijdrage aan het leerproces dankzij de feedback aan de leerling.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

16. De toets heeft invloed op de inhoud van het verdere onderwijsleerproces.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

17. De toets bevordert het vermogen tot zelfbeoordeling bij de leerling.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

18. De toetsscores voorspellen voor Nederlands de toekomstige prestaties van de leerling in het onderwijs.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

19 De toetsscores zijn generaliseerbaar en zeggen iets over de prestaties op alle andere taken die in de toets voorgelegd hadden kunnen worden.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

20. De toetsscores zijn een aantoonbaar goede voorspeller van de prestaties in de beroepspraktijk.

Dit vind ik voor formatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

Eisen die aan summatieve taaltoetsen worden gesteld

Dit deel van de vragenlijst gaat alleen over summatieve taaltoetsen, dus over toetsen die aan het eind van een onderwijsperiode worden ingezet. We noemen weer twintig eisen die vaak aan toetsen gesteld worden. Kunt u per eis aangeven hoe belangrijk u die vindt voor summatieve taaltoetsen?

1. De taken in de taaltoets komen overeen met taken in de praktijk.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

2. De taken in de taaltoets zijn competentiegericht: kennis, vaardigheid, houding geïntegreerd.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

3. De moeilijkheidsgraad van de taaltoets is vergelijkbaar met de moeilijkheidsgraad van taken in de praktijk.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

4. De toetsscore geeft een goed beeld van het prestatieniveau in de praktijk.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

5. De prestaties op de toets wordt vergeleken met van tevoren vastgestelde kwaliteitseisen.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

6. Leerlingen zijn op de hoogte van wat er van ze verwacht wordt bij de taaltoets.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

7. De toets meet daadwerkelijk wat er gemeten moet worden (validiteit).

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

8. De beoordeling van prestaties op de toets is beoordelaars-onafhankelijk en/of wordt door meer beoordelaars gedaan.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

9. De toets heeft een *aangetoonde* hoge mate van betrouwbaarheid.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

10. De toetsdoelen komen overeen met de onderwijsdoelen.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

11. De leerling kan de taaltoets doen op het moment dat hij of zij daaraan toe is.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

12. De taaltoets geeft inzicht in het handelen van de leraar en is leerzaam voor de leerling.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

13. De toets is leerwegonafhankelijk: het maakt niet uit waar de leerling de competenties heeft verworven – op school, op de werkplek of elders.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

14. De kosten en tijd van de afname staan in verhouding tot de opbrengst van de toets.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

15. De toets levert een bijdrage aan het leerproces dankzij de feedback aan de leerling.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

16. De toets heeft invloed op de inhoud van het verdere onderwijsleerproces.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

17. De toets bevordert het vermogen tot zelfbeoordeling bij de leerling.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

18. De toetsscores voorspellen voor Nederlands de toekomstige prestaties van de leerling in het onderwijs.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

19 De toetsscores zijn generaliseerbaar en zeggen iets over de prestaties op alle andere taken die in de toets voorgelegd hadden kunnen worden.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

20. De toetsscores zijn een aantoonbaar goede voorspeller van de prestaties in de beroepspraktijk.

Dit vind ik voor summatieve taaltoetsen erg onbelangrijk 1 2 3 4 erg belangrijk

Aspecten van taalvaardigheid

In dit laatste gedeelte staat centraal wat er in een taaltoets wordt gemeten.

Een taaltoets kan verschillende aspecten van taalvaardigheid beoordelen. Hieronder worden een tiental aspecten genoemd. Kunt u aangeven
- in welke mate deze aspecten in de door u gehanteerde taaltoetsen beoordeeld worden en
- hoe belangrijk u deze aspecten vindt?

1. Lezen

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

2. Schrijven

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

3. Luisteren

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

4. Gesprekken voeren / presenteren

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

5. Taalverzorging (spelling, interpunctie, zinsbouw)

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

6. Uitspraak

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

7. Teksten schrijven

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

8. Productieve woordenschat

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

9. Woordbegrip / receptieve woordenschat

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

10.Tekstbegrip

Beoordeeld	in kleine mate	1	2	3	4	in grote mate
Dit vind ik	zeer onbelangrijk	1	2	3	4	zeer belangrijk

Zijn er aspecten van taalvaardigheid die niet genoemd zijn, maar die u wel van groot belang vindt?

0 Nee
0 Ja, namelijk **……**

Zijn er aspecten van taalvaardigheid waarvoor u geen toetsinstrument heeft terwijl dat naar uw idee wel nodig is?

0 Nee
0 Ja, namelijk **……**

Samenvatting van onderzoeksresultaten

Als u daar prijs op stelt , sturen we u in februari 2010 een samenvatting van de resultaten van dit onderzoek toe.

Wilt u die samenvatting ontvangen? JA / NEE

Te gebruiken mailadres ……

Bijlage 3: Geraadpleegden

Er zijn telefonische interviews gehouden met zie map

De docentenenquête is ingevuld door vertegenwoordigers van

Koning Willem I College	Den Bosch
Driestar	Lekkerkerk
Willem Blaeu	Alkmaar
ALFA_COLLEGE	GRONINGEN
Don Bosco College	Volendam
Graafschapcollege	doetinchem
csg reggesteyn	rijssen
School voor Kunst, Cultuur en Media	Tilburg
AOC Oost	Twello
ROC De Rooi Pannen	Tilburg
College de Heemlanden	Houten
CITAVERDE College	Nederweert
Connect College	Echt
Gerrit Rietveld College	Utrecht
ROC Zadkine, afd. Handel	Rotterdam

De volgende experts zijn geraadpleegd:

Huub van den Bergh, hoogleraar Levende Talen te Utrecht en Amsterdam; methodoloog	+	*
Aukje Bergsma, productmanager BVE Cito	+	
Theo Eggen, hoogleraar Psychometrische Aspecten van Examinering aan Universiteit Twente; methodoloog	+	*
Hilde Hacquebord, onderzoeker en toetsontwikkelaar ETOC, Rijksuniversiteit Groningen	+	*
Piet Litjens, senior consultant bij Cinop		*
Esther Steenbeek, psycholinguïst RU en EN	+	*
Gerard Straetmans, lector assesment Saxion Hogescholen Deventer & wetenschappelijk medewerker Cito		
Saskia Wools, toetsdeskundige bij Cito	+	*

+ = enquête ingevuld
* = deelname aan seminar op het Expertisecentrum Nijmegen op 12 januari 2010