

Docentenreeks

# Toetsen in het hoger onderwijs

Derde, geheel herziene druk

redactie:

Henk van Berkel

Anneke Bax

Desirée Joosten-ten Brinke



Bohn  
Stafleu  
van Loghum

Springer Media

**Redactie:**

Henk van Berkel

Anneke Bax

Desirée Joosten-ten Brinke

**Toetsen in het hoger onderwijs**

**Redactie:**

Henk van Berkel

Anneke Bax

Desirée Joosten-ten Brinke

# Toetsen in het hoger onderwijs

Derde, geheel herziene druk



Bohn  
Stafleu  
van Loghum

---

Springer Media

Houten 2014

**ISBN 978-90-368-0238-3**

© 2014 Bohn Stafleu van Loghum, onderdeel van Springer Media BV

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën of opnamen, hetzij op enige andere manier, zonder voorafgaande schriftelijke toestemming van de uitgever.

Voor zover het maken van kopieën uit deze uitgave is toegestaan op grond van artikel 16b Auteurswet j° het Besluit van 20 juni 1974, Stb. 351, zoals gewijzigd bij het Besluit van 23 augustus 1985, Stb. 471 en artikel 17 Auteurswet, dient men de daarvoor wettelijk verschuldigde vergoedingen te voldoen aan de Stichting Reprorecht (Postbus 3060, 2130 KB Hoofddorp). Voor het overnemen van (een) gedeelte(n) uit deze uitgave in bloemlezingen, readers en andere compilatiewerken (artikel 16 Auteurswet) dient men zich tot de uitgever te wenden.

Samensteller(s) en uitgever zijn zich volledig bewust van hun taak een betrouwbare uitgave te verzorgen. Niettemin kunnen zij geen aansprakelijkheid aanvaarden voor drukfouten en andere onjuistheden die eventueel in deze uitgave voorkomen.

NUR 916

Ontwerp omslag: Studio Bassa

Automatische opmaak: Crest Premedia Solutions (P) Ltd., Pune, India

Eerste druk 2002

Tweede, herziene druk 2006

Derde, geheel herziene druk 2014

Bohn Stafleu van Loghum

Het Spoor 2

Postbus 246

3990 GA Houten

[www.bsl.nl](http://www.bsl.nl)

## Voorwoord

---

Alweer een nieuw boek over toetsen in het hoger onderwijs. Maar dan wel één waarvan wij denken dat het een aanvulling is op de bestaande boeken, en niet zozeer een verdubbeling. Het boek gaat in op nieuwe ontwikkelingen op het gebied van toetsen, op vragen van docenten en het bevat veel concrete handvatten.

De belangrijkste reden om dit boek te schrijven, was de roep van docenten om meer informatie over nieuwe ontwikkelingen op het gebied van toetsen. Tijdens onze contacten met docenten bleek dat zij met de gebruikelijke toetsvormen (vooral met meerkeuzevragen) onvoldoende zicht konden krijgen op de kennis, kunde en vaardigheden van hun studenten. Dit heeft vooral te maken met nieuwe doelen die aan onderwijs worden gesteld én met innovatieve onderwijsvormen die een meer aangepaste en authentieke toetsing vergen. Te denken valt hierbij aan het afnemen van toetsen met de computer, de overalltoets, toetsen met portfolio's, voortgangstoetsen, toetsen van competenties en het toetsen van groepsproducten. Deze onderwerpen zijn in het boek opgenomen.

Een andere belangrijke reden voor dit boek is dat er nog steeds veel misgaat bij toetsing en examinering. Uit analyses van tentamens en uit de cursussen die wij voor docenten verzorgen, komen veelvuldig onnodige slordigheden boven tafel bij het construeren van toetsen en bij het formuleren van vragen. Meestal zijn deze makkelijk oplosbaar, mits docenten over de noodzakelijke informatie beschikken. Zo is het bijvoorbeeld met behulp van een werkwoordentabel relatief eenvoudig vragen te formuleren die verder gaan dan het bevragen van reproducerende kennis. Daarnaast is het voor een docent ook eenvoudig de toets te laten voldoen aan de validiteitseis door een specificatietabel te gebruiken. Verder is de kwaliteit van de vragen sterk te verbeteren door de checklists aan te houden die hiervoor bestaan.

Een derde reden is eveneens ontleend aan de praktijk: veel klachten van studenten hebben te maken met de beoordeling, de wijze van nakijken, de zak/slaaggrens en de beslissingen die over hen worden genomen. Ook voor deze problemen bestaan relatief eenvoudige oplossingen. Het hanteren van een antwoordmodel bijvoorbeeld, voorkomt al veel problemen.

Toetsen zijn vrijwel nooit honderd procent perfect, waardoor er ook altijd onzuiverheden in de beoordeling zullen zitten. De manier van het toekennen van cijfers houdt daar nauwelijks rekening mee. Het is nodig dat docenten zich realiseren dat het helemaal niet zo 'zeker' is dat een 5,5 echt een 5,5 is, maar dat een 5,5 net zo goed een 5,4 of 5,6 had kunnen zijn; de zekerheid van terecht zakken of slagen heeft niet altijd een stevige basis.

Het idee voor dit boek is ontstaan naar aanleiding van ons vorige boek, *Beoordelen in het onderwijs, een handleiding voor het construeren van toetsen en het evalueren van leerdoelen en onderwijsvormen (1993)*. In dat boek worden diverse nieuwe toetsinstrumenten – waaraan onder docenten behoefte blijkt te bestaan – echter niet behandeld. Daarom is in plaats van een herdruk, gekozen voor een geheel nieuwe opzet. Voor ieder hoofdstuk zijn een of meer auteurs benaderd die deskundig zijn op het betreffende gebied. Wij danken alle auteurs voor hun bijdragen.

Januari 2002

#### ■ Bij de tweede druk

Toetsing in het hoger onderwijs staat volop in de belangstelling. De Nederlands-Vlaamse Accreditatie Organisatie (NVAO) heeft recentelijk uitgesproken dat de wijze waarop opleidingen nadere invulling geven aan toetsen een speerpunt zal zijn bij het beoordelen of opleidingen aan de accreditatie-eisen voldoen. Dat impliceert dat opleidingen (nog) meer aandacht moeten geven aan de manier waarop studenten worden getoetst. Sinds het verschijnen van dit boek in 2002 zijn er ook andere ontwikkelingen die een geactualiseerde druk nodig maken.

Studenten die in het vwo met het studiehuis in aanraking zijn gekomen, hebben inmiddels het hoger onderwijs bereikt (en soms alweer verlaten). Op het vwo hebben zij bijvoorbeeld geleerd werkstukken te maken met behulp van bronnen van het internet. Daar is niets op tegen, mits zij niet zomaar teksten van het internet halen en de suggestie wekken dat ze er zelf de auteur van zijn. Dat zou plagiaat zijn; bronvermelding is een vereiste. Het raadplegen van het internet heeft in het algemeen, onder alle scholieren en studenten, een hoge vlucht genomen. Alleen schort het er nogal eens aan om de bronvermelding op juiste wijze toe te passen. Het brengt studenten ook hoe langer hoe meer in de verleiding andermans werk voor eigen werk uit te geven, hoewel dit soms uit onwetendheid gebeurt. Om het docenten mogelijk te maken na te gaan of er sprake is van plagiaat, is hoofdstuk 21 toegevoegd dat gaat over het opsporen van plagiaat.

Verder heeft de bachelor-masterstructuur ('bama') in het hoger onderwijs haar intrede gedaan. Hoewel de bama in het wetenschappelijk onderwijs wordt gezien als één geheel – de overgrote meerderheid van afgestudeerde bachelors gaat door voor master – zal het niet lang meer duren dat een universitaire bachelor het eindstation is van een opleidingsperiode. Dit naar het voorbeeld van de ervaringen in Angelsaksische landen. Studenten met een bachelordiploma gaan daar bijvoorbeeld eerst een bepaalde periode werken om vervolgens alsnog te kiezen voor een mastersopleiding, maar dat kan dan problemen geven voor toelating omdat geleerde kennis verouderd blijkt te zijn. Inmiddels is in het onderwijs steeds meer sprake van *life long learning* en ervaringskennis. De zogenaamde EVC-procedure (Erkenning Verworven Competenties) is een methode om ervaringskennis in kaart te brengen en te relateren aan de instroomeisen. Daarom is hoofdstuk 22, Toetsen met EVC, aan de tweede druk toegevoegd.

Vooral binnen het hbo is een toenemende mate van vraaggestuurd en competentiegericht onderwijs te zien. Doelstellingen zijn geformuleerd in termen van beroepscompetenties en authentieke (levensechte) beroepstaken, en praktijksituaties vormen de basis van een curriculum. Naast toetsvormen als (performance)assessments, peer- en coassessment zijn ook de meer traditionele toetsvormen (waaronder kennistoetsen met meerkeuzevragen) van belang, maar met een andere toepassing. Kennis is juist een voorwaarde om competenties te kunnen verwerven, waarbij het er vooral om gaat kennis te gebruiken in een beroepscontext. Dit alles heeft vergaande implicaties voor toetsing, kwaliteitscriteria en beoordeling. In het nieuwe hoofdstuk 23 worden deze aspecten beschreven.

Behalve dat drie nieuwe hoofdstukken zijn toegevoegd zijn sommige hoofdstukken sterk geactualiseerd, in de tekst maar ook in de geannoteerde literatuur achteraan.

**Amsterdam/Heerlen, Anneke Bax**

**Maastricht, Henk van Berkel**

Juni 2006

#### ■ Bij de derde druk

‘Toetsing in het hoger onderwijs staat volop in de belangstelling.’ Deze zin stond ook al in het voorwoord bij de tweede druk, in 2006. Toch is de inhoud van dit boek nog steeds actueel, zelfs meer dan in het verleden. De kwaliteitscontrole van het hoger onderwijs wordt, namens de overheid, uitgevoerd door de Nederlands-Vlaamse Accreditatie Organisatie (NVAO). Die hanteert hierbij richtlijnen, ook wel standaarden genoemd. Een van die standaarden is:

#### » Standaard 3:

De opleiding beschikt over een adequaat systeem van toetsing en toont aan dat de beoogde eindkwalificaties worden gerealiseerd. «

In de toelichting staat:

» Het gerealiseerde niveau blijkt uit de tussentijdse en afsluitende toetsen, de afstudeerwerken en de wijze waarop afgestudeerden in de praktijk of in een vervolgopleiding functioneren. De toetsen en de beoordeling zijn valide, betrouwbaar en voor studenten inzichtelijk. «

Indien niet aan deze standaard is voldaan, verliest de opleiding haar accreditatie. Vroeger was het nog mogelijk een eventuele onvoldoende voor het toetssysteem te compenseren met positieve oordelen over andere aspecten. Tegenwoordig is dat onmogelijk. Het is dus voor een opleiding gewoon een halszaak om haar toetsingssysteem op orde te hebben.

De redactie heeft gemeend bij het uitgeven van de derde druk de inhoud van het boek kritisch tegen het licht te houden. Zijn alle hoofdstukken nog wel relevant? Moeten er

geen nieuwe bij? Is de inhoud van sommige hoofdstukken nog wel actueel genoeg, of is aanpassing nodig om nieuwe ontwikkelingen te schetsen? Inderdaad, sommige hoofdstukken waren niet meer actueel – onder andere de overalltoets en vraaggestuurd competentiegericht maatwerkonderwijs komen nauwelijks meer voor in het hoger onderwijs. Andere onderwerpen zijn daarentegen wel sterk in opkomst: toetsbeleid, formatief toetsen, rubrics, onderzoekvaardigheden, (generieke) competenties en landelijke kennis-toetsen. Daaraan zijn nieuwe hoofdstukken gewijd. Alle resterende hoofdstukken zijn zo veel mogelijk aangepast aan de actualiteit, maar ook op basis van reacties van lezers.

De redactie is uitgebreid met Desirée Joosten-ten Brinke. Zij is universitair hoofddocent bij de Open Universiteit en lector Eigentijds toetsen en beoordelen aan de Fontys lerarenopleiding in Tilburg.

**Amsterdam, Anneke Bax**

**Landgraaf, Desirée Joosten-ten Brinke**

**Maastricht, Henk van Berkel**

September 2013



# Inhoud

---

1	<b>Het toetsproces ontleed</b> .....	1
	<i>Henk van Berkel, Anneke Bax en Desirée Joosten-ten Brinke</i>	
1.1	Fasen in het toetsproces.....	2
1.2	Stappen in het toetsproces.....	4
2	<b>Toetsen: toetssteen of dobbelsteen</b> .....	15
	<i>Henk van Berkel en Anneke Bax</i>	
2.1	Kwaliteitseisen aan toetsen .....	16
2.2	Kwaliteit van toets en toetsvragen .....	16
2.3	Rechtvaardigheidseisen.....	17
2.4	Psychometrische analyse: het berekenen van indicatoren .....	19
2.5	Reparatiemogelijkheden.....	24
2.6	Ten slotte .....	27
3	<b>Ontwikkelen van toetsbeleid</b> .....	29
	<i>Riet Martens en George Moerkerke</i>	
3.1	De aandacht voor toetsbeleid .....	30
3.2	Integraal toetsbeleid: een samenhangend stelsel .....	31
3.3	Toetsbeleid en kwaliteit van toetsing op de verschillende organisatieniveaus .....	33
3.4	Borging van kwaliteit van tentamens en examens door de examencommissie.....	37
3.5	De docent en het toetsbeleid.....	38
3.6	Ten slotte .....	40
4	<b>Toetsen met landelijke kennistoetsen</b> .....	41
	<i>Arnoud van Leuven en Wim Lansu</i>	
4.1	Waarom een landelijke kennistoets: de gemeenschappelijke kennisbasis .....	42
4.2	Voorbeeld: de rol van de docent.....	43
4.3	Mogelijkheden.....	46
4.4	Voordelen.....	47
4.5	Nadelen.....	48
4.6	Voorwaarden voor een succesvolle toepassing.....	49
5	<b>Kwaliteit van toetsprogramma's</b> .....	51
	<i>Liesbeth Baartman, Raymond Kloppenburg en Frans Prins</i>	
5.1	Wat is een toetsprogramma?.....	52
5.2	Waarom toetsprogramma's?.....	53
5.3	Ontwerp van een toetsprogramma.....	57
5.4	De kwaliteit van toetsprogramma's.....	58
5.5	Zelfevaluatie-instrument KIT.....	61
5.6	Ten slotte .....	62

6	<b>Toetsen zonder plagiëren</b> .....	65
	<i>Henk van Berkel</i>	
6.1	Inleiding .....	66
6.2	Voorkomen is beter dan genezen .....	67
6.3	Soms zijn studenten onwetend .....	67
6.4	Programma's voor het opsporen van plagiaat .....	68
6.5	SafeAssign (Blackboard) .....	69
6.6	WCopyfind .....	70
6.7	Besluiten tot het oordeel plagiaat .....	71
7	<b>Normeren en cijfergeven</b> .....	73
	<i>Henk van Berkel en Wynand Wijnen</i>	
7.1	Twee methoden van normeren: relatief en absoluut .....	74
7.2	Relatief normeren .....	74
7.3	Absoluut normeren .....	76
7.4	De relatieve en absolute methode vergeleken .....	76
7.5	Compromismethoden .....	77
7.6	Het proces van cijfergeven .....	78
8	<b>Formatief toetsen</b> .....	81
	<i>Desirée Joosten-ten Brinke en Dominique Sluismans</i>	
8.1	Wat is formatief toetsen? .....	82
8.2	Waarom is formatief toetsen belangrijk? .....	83
8.3	Effectieve methoden van formatief toetsen .....	84
8.4	De rol van docenten bij formatief toetsen .....	88
8.5	Controlepunten voor docenten .....	89
9	<b>Digitaal toetsen</b> .....	91
	<i>Silvester Draaijer, Patris van Boxel en Marcel van Brunschot</i>	
9.1	Twee voorbeelden van toetsafname met de computer .....	92
9.2	De computer als toetsmachine .....	94
9.3	Itembanken .....	96
9.4	Nakijken van open vragen en werkstukken .....	97
9.5	Echt andere toetsvormen: de toekomst? .....	99
10	<b>Creatieve toetsvormen</b> .....	101
	<i>Titus Geerligts, Henk Schmidt, Irma Kokx, Erik de Graaf en Henk van Berkel</i>	
10.1	Creatieve toetsvormen .....	102
10.2	Ten slotte .....	111
11	<b>Toetsen met gesloten vragen</b> .....	113
	<i>Henk van Berkel en Anneke Bax</i>	
11.1	Meerkeuzevragen .....	114
11.2	Juist/onjuist- en ja/nee-vragen .....	116

11.3	Andere gesloten vraagvormen .....	117
11.4	Samenstellen van de toets en controleren van vragen .....	117
11.5	Toetsen van hogere cognitieve vaardigheden met gesloten vragen .....	119
11.6	Aanwijzingen en tips voor het construeren van gesloten vragen .....	122
11.7	Checklist .....	127
12	<b>Toetsen met open vragen</b> .....	129
	<i>Tom Erkens</i>	
12.1	Vormen van open vragen .....	130
12.2	Voor- en nadelen van open vragen .....	131
12.3	Correctievoorschrift .....	133
12.4	Constructieregels .....	134
12.5	Oorzaken voor verschillen in beoordeling .....	139
12.6	Oplossingen voor beoordelaarsverschillen .....	140
12.7	Checklist .....	140
13	<b>Toetsen met korte casussen</b> .....	143
	<i>Lambert Schuwirth</i>	
13.1	Korte casussen geschikter dan lange .....	144
13.2	Construeren en beoordelen van een korte-casustoets .....	145
13.3	Schrijven van casussen .....	147
13.4	Tips bij het schrijven van casussen .....	148
13.5	Ten slotte .....	154
13.6	Tips .....	155
14	<b>Toetsen met een mondelinge toets</b> .....	157
	<i>Henk van Berkel en Anneke Bax</i>	
14.1	Overwegingen voor mondeling toetsen .....	158
14.2	Voordelen .....	159
14.3	Kwaliteitseisen voor de beoordeling en haalbaarheid .....	160
14.4	Beroepsmogelijkheden .....	165
14.5	Conclusies .....	167
14.6	Tips .....	168
15	<b>Toetsen met voortgangstoetsen</b> .....	169
	<i>Arno Muijtjens en Wynand Wijnen</i>	
15.1	De rol van de docent bij voortgangstoetsing .....	170
15.2	Een voorbeeld: de interuniversitaire Voortgangstoets Geneeskunde (iVGT) .....	171
15.3	Mogelijkheden van voortgangstoetsing .....	176
15.4	Voor- en nadelen .....	177
15.5	Voorwaarden .....	179
15.6	Ten slotte .....	180

16	<b>Toetsen met performance assessment methodieken</b> .....	181
	<i>Gerard Straetmans</i>	
16.1	Inleiding .....	182
16.2	Argumenten voor de inzet .....	183
16.3	Knelpunten bij het gebruik.....	184
16.4	Enkele cruciale ontwerpvragen.....	187
16.5	Ten slotte .....	192
17	<b>Toetsen met rubrics</b> .....	193
	<i>Paul van den Bos, Carel Burghout en Desirée Joosten-ten Brinke</i>	
17.1	Definities en soorten .....	194
17.2	Rubrics voor formatief én summatief toetsen.....	199
17.3	Curriculumbrede inzet .....	201
17.4	Het beoordelen van de kwaliteit .....	202
17.5	Een stappenplan voor de constructie.....	202
18	<b>Toetsen met EVC</b> .....	205
	<i>Ruud Klarus en Desirée Joosten-ten Brinke</i>	
18.1	Wat is EVC?.....	206
18.2	Toetsmethoden voor een EVC-procedure.....	207
18.3	Het ontwikkelen van een EVC-procedure .....	208
18.4	De kwaliteitscode EVC.....	208
18.5	Instrumenten .....	209
18.6	Voorbeelden van EVC-procedures .....	211
18.7	Aandachtspunten voor EVC .....	215
18.8	Ten slotte .....	216
19	<b>Toetsen van competenties</b> .....	217
	<i>Judith Gulikers en Niek van Benthum</i>	
19.1	Wat is competentietoetsing? .....	218
19.2	De meerwaarde van competentietoetsing.....	219
19.3	Competentietoetsing vraagt ‘andersom denken’ .....	219
19.4	Het wat en hoe van competentietoetsing .....	221
19.5	Wat wordt getoetst?.....	221
19.6	Hoe wordt getoetst? .....	223
19.7	Ten slotte .....	227
20	<b>Toetsen van werkplekleren</b> .....	229
	<i>Marcel van der Klink en Jo Boon</i>	
20.1	Vormen van leren op de werkplek.....	230
20.2	Dilemma’s en problemen in werkplekleren .....	231
20.3	Instrumenten voor het beoordelen van werkplekleren .....	234
20.4	Ten slotte .....	238

21	<b>Toetsen van onderzoekvaardigheden</b> .....	241
	<i>Quinta Kools en Ilona Mathijsen</i>	
21.1	<b>Wat is onderzoekvaardigheid?</b> .....	242
21.2	<b>Waarop beoordelen?</b> .....	243
21.3	<b>Instrumenten voor beoordeling</b> .....	246
21.4	<b>Kwaliteit van beoordeling</b> .....	247
21.5	<b>Formatieve toetsing: de rol van feedback</b> .....	248
22	<b>Toetsen van academische vaardigheden</b> .....	251
	<i>Albert Pilot</i>	
22.1	<b>Kenmerkende competenties voor het hoger onderwijs</b> .....	252
22.2	<b>Professionele vorming</b> .....	253
22.3	<b>Academische vorming</b> .....	254
22.4	<b>Werkwijze om geschikte toetsvormen te selecteren</b> .....	255
22.5	<b>Drie voorbeelden van een toetsvorm</b> .....	257
22.6	<b>Competenties en toetsvormen</b> .....	261
22.7	<b>Ten slotte</b> .....	261
23	<b>Toetsen van schriftelijke werkstukken</b> .....	263
	<i>Sjoerd Romme</i>	
23.1	<b>Onbetrouwbaarheid en terugkoppelingseffecten</b> .....	264
23.2	<b>Beoordelingscriteria: wat wordt beoordeeld?</b> .....	265
23.3	<b>Wie beoordeelt?</b> .....	268
23.4	<b>Hoe wordt getoetst? De afstudeerkring als voorbeeld</b> .....	270
24	<b>Toetsen van groepswerk</b> .....	273
	<i>Anneke Bax, Jacob Perrenet en Henk van Berkel</i>	
24.1	<b>Overwegingen voor een groepsproduct</b> .....	274
24.2	<b>Beoordelen van groepsproducten</b> .....	275
24.3	<b>Opzetten van een beoordelingssysteem</b> .....	276
24.4	<b>Methoden voor beoordeling: het betrekken van medestudenten</b> .....	280
24.5	<b>Uitwerking van enkele beoordelingsmethoden</b> .....	281
24.6	<b>Kwaliteitseisen aan de beoordeling</b> .....	283
	<b>Beoordelaarseffecten</b> .....	287
	<b>De redacteurs</b> .....	291
	<b>De auteurs</b> .....	293

<b>Lexicon</b> .....	299
<b>Referenties</b> .....	315
<b>Register</b> .....	325

# Het toetsproces ontleed

*Henk van Berkel, Anneke Bax en Desirée Joosten-ten Brinke*

## Inleiding

Uit welke fasen en stappen bestaat het toetsproces?

Wat zijn de doelen van toetsen?

Wat zijn doelstellingen?

Hoe kunnen doelstellingen nader worden geconcretiseerd?

Wat is een specificatietabel?

Welke toetsvormen zijn er?

Wat zijn inspiratiebronnen voor het construeren van toetsvragen?

## 1.1 Fasen in het toetsproces – 2

### 1.2 Stappen in het toetsproces – 4

- 1.2.1 Stap 1: Waarom beoordelen en toetsen: formuleren van de bedoeling(en) – 4
- 1.2.2 Stap 2: Het formuleren en concretiseren van de doelen – 5
- 1.2.3 Stap 3: Het kiezen van de toetsvorm – 8
- 1.2.4 Stap 4: Het construeren en vooraf controleren van toetsvragen of toetstaken – 8
- 1.2.5 Stap 5: Het samenstellen van de toets – 11
- 1.2.6 Stap 6: Het afnemen van de toets – 11
- 1.2.7 Stap 7: Het nakijken van de toets, inclusief een kwaliteitscontrole – 11
- 1.2.8 Stappen 8 en 9: Cijfergeven – 12
- 1.2.9 Stappen 10 en 11: Beslissen en rapporteren – 12

Het afnemen van toetsen impliceert meer dan het construeren van toetsvragen; het is een proces dat verschillende fasen en stappen kent. Elk toetsproces is in essentie te herleiden tot een viertal fasen waarin steeds een ander aspect op de voorgrond staat (zie ook ■ figuur 1.1).

Toetsproces = Doelbepaling + Meten + Cijfergeven + Beslissen

Deze ontleding van het toetsproces geldt voor alle toetsen met een summatief doel. De grenzen tussen de fasen zijn niet precies te trekken omdat bij het concretiseren van een fase rekening wordt gehouden met de andere fase, bewust of onbewust. Maar deze vierdeling is als analytisch denkkader goed hanteerbaar. Elke fase is onder te verdelen in stappen, die in de volgende paragraaf worden uitgewerkt.

## 1.1 Fasen in het toetsproces

---

### ■ Doelbepaling

In de fase waarin het doel van de toets wordt bepaald, gaat het om:

- formuleren van de bedoeling(en) van de toets
- formuleren van de doelstellingen
- concretiseren van de doelen
- opstellen specificatietabel

De doelbepalingfase hoeft niet bij iedere toets te worden doorlopen omdat de doelstellingen van een toets meestal voor een aantal jaren vastliggen. Bovendien is niet iedere docent zelf verantwoordelijk voor de doelen. Het is een gedeelde verantwoordelijkheid op opleidingsniveau. In dit hoofdstuk worden het formuleren en concretiseren van de doelstellingen uitgebreid beschreven.

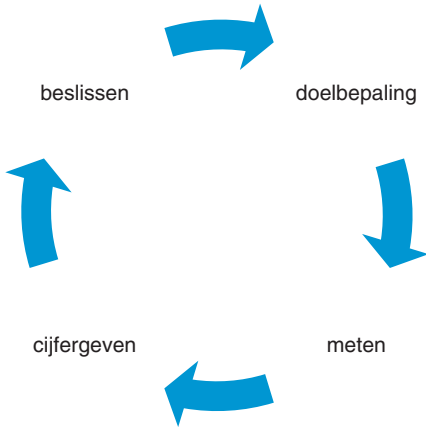
### ■ Meten

De meetfase omvat alle stappen die te maken hebben met het vormgeven en afnemen van een toets, en wel in een vaste volgorde:

- kiezen van de toetsvorm op basis van de doelen
- construeren van de toetstaken/toetsvragen
- vooraf controleren van de toetstaken/toetsvragen
- samenstellen van de toets
- afnemen van de toets
- nakijken van de toets
- kwaliteitscontrole

Deze stappen zijn op vrijwel alle toetsvormen toepasbaar. Bij een schriftelijke toets gaat het bijvoorbeeld om het construeren van toetsvragen, terwijl het bij een performance assessment gaat om het construeren van toetstaken. Beide moeten voldoen aan vergelijkbare eisen. De combinatie van toetsvragen/toetstaken vormt de toets. De stappen staan in de





■ **Figuur 1.1** Fasen in het toetsproces.

volgende paragraaf verder uitgewerkt; aan de kwaliteitscontrole van de toets is een afzonderlijk hoofdstuk gewijd (► hoofdstuk 2).

#### ■ **Cijfergeven**

De derde fase betreft het waarderingsaspect:

- bepalen van de zak/slaaggrens
- toekennen van cijfers aan de scores

De fase van het cijfergeven is wellicht voor studenten de belangrijkste, in ieder geval de spannendste. Maar ook voor docenten is deze fase belangrijk. Het bepalen van de zak/slaaggrens heeft grote gevolgen voor de studenten. In deze fase moeten docenten op basis van een score bepalen wat voor cijfer een student krijgt. Het geven van cijfers lijkt simpel, maar vergt grote aandacht van docenten. ► Hoofdstuk 7 gaat hier uitvoerig op in.

#### ■ **Beslissen**

In de laatste fase worden beslissingen genomen op basis van de in de voorafgaande fase toegekende cijfers:

- nemen en verantwoorden van beslissingen op grond van de toegekende cijfers
- rapporteren van de toetsuitslag

In de meest eenvoudige vorm bestaat de toetsuitslag uit niet meer dan het cijfer plus de consequentie voor de studenten, meestal gezakt of (voorwaardelijk) geslaagd, zonder verdere toelichting. Een cijfer alleen geeft echter geen informatie over wat er al wel of nog niet wordt beheerst. Deze extra informatie toevoegen maakt een toetsuitslag ook geschikt als feedbackinstrument voor zowel studenten als docenten. Bijvoorbeeld: het aangeven van het hoofdstuk waarop de vraag betrekking heeft of het achterhalen van de antwoorden op

**Tabel 1.1** Fasen met stappen in het toetsproces.

<i>Doelbepaling</i>	
1	Waarom beoordelen en toetsen: het formuleren van de bedoeling(en)
2	Het formuleren en concretiseren van de doelen
<i>Meten</i>	
3	Het kiezen van de toetsvorm
4	Het construeren en vooraf controleren van de toetsvragen
5	Het samenstellen van de toets
6	Het afnemen van de toets
7	Het nakijken van de toets, inclusief een kwaliteitscontrole
<i>Cijfergeven</i>	
8	Het bepalen van de zak/slaaggrens
9	Het toekennen van cijfers aan de scores
<i>Beslissen</i>	
10	Het kunnen verantwoorden van de beslissing
11	Het rapporteren van de toetsuitslag voor nabespreking, evaluatie en bijstelling

extreem moeilijke vragen. Rapportage aan de opleidings- c.q. examencommissie over de cijfers per student en over de kwaliteit van de toets (met behulp van een toetsanalyse na afname) is nodig. In ► hoofdstuk 2 staat vermeld welke gegevens van belang zijn voor de kwaliteitscontrole en hoe deze worden verkregen.

In ► tabel 1.1 is een overzicht opgenomen van de fasen in het toetsproces met bij elke fase de stappen die in ► paragraaf 1.2 verder worden uitgelegd.

## 1.2 Stappen in het toetsproces

De bovengenoemde fasen van het toetsproces kunnen worden uiteengehaald in elf concrete stappen (zie ► tabel 1.1). Hierna volgt een toelichting.

### 1.2.1 Stap 1: Waarom beoordelen en toetsen: formuleren van de bedoeling(en)

Algemeen gesteld: er zijn toetsen met een summatief doel om te komen tot beslissingen over studenten, bijvoorbeeld zakken of slagen, toelaten of afwijzen, aannemen of afwijzen, en er zijn toetsen met een formatief doel, bijvoorbeeld om aan te geven over welk onderwerp de student nog te weinig weet. De inhoud van dit hoofdstuk gaat over summatief toetsen, ► hoofdstuk 8 gaat over formatieve toetsing.

Er zijn globaal gesproken drie tijdvakken waarin uiteenlopende beslissingen over studenten worden genomen (zie ► tabel 1.2).

■ **Tabel 1.2** Tijdvak, doel en functie van toetsen.

Tijdvak	Doel	Functie
vóór de start	selectie	toelating
tijdens	leermiddel	sturing
tijdens	voortgangscntrole	bijsturing
aan het eind	kwalificatie	niveaubepaling

### ■ Toetsen als selectie-instrument

Wanneer, voorafgaand aan het doorlopen van het onderwijsprogramma, moet worden nagegaan of een student geschikt is voor de opleiding, heeft de toets (ingangs)*selectie* tot doel. Deze functie van toetsen valt buiten het kader van dit boek.

### ■ Toetsen als leermiddel en voortgangscntrole

Toetsen kunnen tijdens de onderwijsperiode worden gebruikt om studenten en docenten inzicht te geven in de mate waarin studenten zijn gevorderd in de bestudering van het studieonderdeel of in de mate van verkregen competentie. Wanneer toetsen op een zorgvuldige wijze in het onderwijs zijn geprogrammeerd, vormen ze een middel om studenten te activeren en hun leerproces te sturen. Toetsen die bij wijze van voortgangscntrole zijn afgenomen, beogen het leren te sturen en hebben een formatieve functie. Het belang van dit soort toetsen is de *feedback*functie. De studenten kunnen aan de hand van de toetsuitslag en de feedback bepalen waar hun hiaten in de leerstof zitten en welke vaardigheden zij onvoldoende beheersen. Hiertoe is het uiteraard wel nodig dat de feedback voldoende specifiek en gedetailleerd is.

### ■ Toetsen als kwalificeringinstrument

De meeste opleidingen hanteren toetsen aan het eind van een onderwijsperiode om na te gaan of studenten het studieonderdeel voldoende beheersen. Deze functie van toetsen wordt aangeduid met de *summatieve* functie. Een studietoets levert doorgaans een score op die aangeeft in welke mate studenten ‘iets’ beheersen. Met andere woorden, de score weerspiegelt het *niveau* van hun kennis en vaardigheden. De hoogte van het niveau bepaalt de uitkomst van beslissingen: slagen of zakken, voldoende of onvoldoende, toelaten of weigeren enzovoort. Een voldoende score impliceert dat de betrokken studenten voldoen aan de eisen. In dat geval verkrijgen studenten daarvan een officieel bewijs en zijn zij gekwalificeerd. Een summatieve toets is een proeve van bekwaamheid en sluit als het ware iets af.

## 1.2.2 Stap 2: Het formuleren en concretiseren van de doelen

Het formuleren van doelstellingen is belangrijk om te voorkomen dat onderwijs een kwestie van willekeur gaat worden. In dit boek is het begrip ‘doelstellingen’ – of kortweg ‘doelen’

– gekoppeld aan een gedeelte van de opleiding, een studieonderdeel, dat wordt getoetst. De definitie van doelstellingen aan de hand waarvan docenten een toets maken, luidt:

» omschreven kwaliteiten op het gebied van kennis, inzicht en vaardigheden en in voorkomende gevallen (beroeps)houding, waarover degene die het studieonderdeel voltooit, dient te beschikken. «

Doelstellingen vormen het startpunt om een toets te construeren.

#### ■ Categorisering en classificatie van doelen

Doelstellingen geven aan wat studenten van het onderwijs moeten gaan opsteken. Een voorbeeld is feitenkennis die vrijwel uitsluitend een beroep doet op het reproduceren van kennis. Maar van studenten mag ook worden verwacht dat ze deze feitenkennis in bepaalde situaties kunnen toepassen of er anderszins mee kunnen omgaan. Deze doelen, in de praktijk aangeduid met reproducerende kennis, inzicht en toepassen, vallen onder de noemer *cognitieve* doelstellingen. Daarnaast zijn opleidingen ook gericht op het bijbrengen van vaardigheden en beroepshoudingen.

De classificatie van Bloom (1972) is nog steeds de bekendste en meest gehanteerde. Doelstellingen zijn door hem onderverdeeld in drie (hoofd)categorieën:

1. Cognitieve doelstellingen: alles wat te maken heeft met verstandelijk functioneren.
2. Affectieve doelstellingen: doelen die te maken hebben met het ontvangen en reageren op signalen.
3. Psychomotorische doelstellingen: doelen gericht op perceptie en het handelen.

Iedere categorie is vervolgens verder onderverdeeld aan de hand van een nadere classificatie, ook wel aangeduid met *taxonomie*. Voor cognitieve doelstellingen hanteert Bloom de volgende classificatie:

- feitenkennis
- begrip
- toepassen
- analyseren
- synthetiseren
- evalueren

Andere veelgebruikte taxonomieën zijn die van Anderson & Kratwohl (2001; dit is een herziening van Bloom), Miller (1990) en Romiszowski (1999).

De rangordening is niet willekeurig. Het doorlopen van de verschillende klassen vergt een steeds hogere intellectuele activiteit. Een dergelijke onderverdeling is daarom een belangrijk hulpmiddel bij het inhoud geven van de toets om ervoor te zorgen dat de vragen in de toets een beroep doen op een breed scala aan intellectuele vaardigheden, en niet bijvoorbeeld uitsluitend op het reproduceren van kennis. In ► tabel 11.2 in ► hoofdstuk 11 staan per niveau voorbeelden van werkwoorden die kunnen worden gehanteerd bij het stellen van vragen van een bepaald niveau.

■ **Tabel 1.3** Voorbeeld van een specificatietabel.

Aantal vragen per onderwerp	Kennisreproductie	Inzicht	Toepassen
overzicht van onderwijssystemen	2	1	2
plaats van onderwijs in de maatschappij	0	3	2
onderwijsvormen in relatie tot: motivatie, leereffect, inzet, betrokkenheid, eigen verantwoordelijkheid, zelfstandigheid	3	3	4
rol van motivatie	2	2	3
geheugen	2	0	1
scholing en diploma's	2	0	2
duaal leren	1	2	2
attitudes	1	1	1
evalueren	4	0	4

■ **Specificatietabel = blauwdruk + niveau**

Om ervoor te zorgen dat een toets een goede spreiding van vragen heeft over onderwerpen en gewenst niveau, is het hanteren van een specificatietabel, ook wel aangeduid met toetsmatrijs, nuttig. De simpelste vorm is een lijst met onderwerpen die in de toets aan de orde moeten komen. Deze onderwerpen zijn doorgaans uit de doelen af te leiden. Een inzichtelijke vertaling van de doelen naar onderwerpen maakt duidelijk of alle doelen inderdaad in de toets aan de orde komen. Het is een vereiste de blauwdruk te maken voordat de constructie van vragen begint en de geconstrueerde vragen vervolgens in te passen in de blauwdruk. Dan pas kan de toets representatief en evenwichtig zijn. Een toets moet representatief zijn ten opzichte van de doelen van het studieonderdeel. Immers met behulp van de toetsuitslag worden belangrijke uitspraken gedaan over het studieonderdeel. Niet alleen ten aanzien van zakken en slagen, maar ook over waar precies de hiaten van studenten zitten zodat studenten zich gericht kunnen bijspijkeren. De toets moet dus voldoende representatief zijn om er valide gevolgtrekkingen uit te kunnen maken. Een term die synoniem is aan representativiteit is *inhoudsvaliditeit*.

Een specificatietabel dient er verder voor te zorgen dat in de uiteindelijke toets, naast een spreiding van de vragen over belangrijke onderwerpen, ook op alle beoogde (cognitieve) niveaus wordt getoetst. Dit hangt nauw samen met de doelen. Daarom is het nodig dat docenten per onderwerp aangeven op welk niveau (of welke niveaus) zij vragen over het betreffende onderwerp gaan stellen.

Het voorbeeld in ■ tabel 1.3 gaat uit van de situatie dat de toets tien onderwerpen bevraagt op drie cognitieve niveaus: kennisreproductie, begrip en toepassing. De gehele toets bestaat uit vijftig vragen.

### 1.2.3 Stap 3: Het kiezen van de toetsvorm

---

De keuze voor een toetsvorm dient aan te sluiten bij de doelstellingen. Met behulp van een toets moet de beoogde leeropbrengst kunnen worden gemeten. Zo is het bijvoorbeeld niet geloofwaardig om met een meerkeuzevraag te toetsen of een student een onderzoek kan presenteren. Daar zijn andere toetsvormen voor, zoals performance assessment (zie ► hoofdstuk 16).

De schriftelijke toetsvormen met open of gesloten vragen zijn (nog steeds) de meest voorkomende. Onder open toetsvormen vallen essay-, korte-antwoord- en aanvultoe- sen (► hoofdstuk 12), maar ook mondeling toetsen (► hoofdstuk 14). Gesloten toetsvormen zijn: ja/nee-, juist/onjuist-, meerkeuze- en matchingtoetsen (Dousma e.a., 1997, zie ► hoofdstuk 11). Bij open toetsvragen moeten studenten zelf een antwoord formuleren ('vrije antwoorden'), terwijl zij bij gesloten toetsvragen een van de alternatieven moeten aankruisen ('geprecodeerd'). Deze toetsen kunnen worden gekoppeld aan casusmateriaal. Bij medische opleidingen is een 'casus' dan een patiënt of een beschrijving ervan, en de toetsen vinden plaats met gestructureerde open vragen (► hoofdstuk 13).

Voor een uitvoerige beschrijving van de verschillende toetsvormen met hun voor- en nadelen, wordt verwezen naar de betreffende hoofdstukken in dit boek.

### 1.2.4 Stap 4: Het construeren en vooraf controleren van toetsvragen of toetstaken

---

#### ■ Het construeren van toetsvragen: inspiratiebronnen

Op basis van de gekozen toetsvorm moeten vragen/toetstaken worden geconstrueerd. Omdat veel docenten huiverig zijn om reeds gebruikte toetsvragen in de toets op te nemen, is het construeren van vragen een steeds terugkerende, grote klus. Het is denkbaar dat de inspiratie hiervoor na verloop van tijd opraakt. Toch bestaan er meestal meer bronnen dan verwacht, waaruit docenten kunnen kiezen (zie ► tabel 1.4).

De beste methode voor het construeren van toetsvragen is uit te gaan van de doelstellingen en de beheersingsniveaus zoals ze vastgelegd zijn in de specificatietabel en die vervolgens te concretiseren in de vorm van toetsvragen. Op die wijze geconstrueerde toetsvragen dragen de beste garantie in zich dat de toets representatief is voor de doelen, en dus inhoudsvalide is.

Vooraf in opleidingen waar vaardigheden belangrijk zijn, is de zogenaamde piramide van Miller (1990) handig bij het formuleren van toetsvragen (zie ► figuur 1.2). In de piramide symboliseren de lagen de complexiteit van competenties. Weten is noodzakelijk om te kunnen toepassen. Toepassen is weer noodzakelijk om iets te kunnen tonen. Om een verrichting in praktijk daadwerkelijk uit te kunnen uitvoeren, moet je als student de verrichting eerst hebben getoond.

In de tweede piramide staan de toetsvormen toegevoegd aan de verschillende niveaus (zie ► figuur 1.3). Het onderste niveau, weten, is te meten met relatief eenvoudige toetsen als meerkeuze en open toetsvormen. De tweede laag, toepassen, vereist wat meer complexe toetsen. Niet zozeer wat de toetsvorm betreft, die kunnen gelijk zijn aan die voor het

**Tabel 1.4** Inspiratiebronnen voor het construeren van vragen.

*Doelstellingen*

Toetsen dienen na te gaan in hoeverre de student de doelen beheerst en moeten dus een inhoudelijke relatie hebben met de doelen. Nu zijn deze vaak globaal van aard. Nodig zijn daarom concretisering in de vorm van nadere beschrijvingen van wat studenten moeten kennen en kunnen nadat zij het studieonderdeel hebben doorlopen. Een toetsvraag is in feite een ultieme operationalisatie van een doelstelling.

*Leerstof*

De leerstof dient een goede afspiegeling te zijn van de doelen; daarmee is de leerstof een goede inspiratiebron.

*Vervolgopleiding*

De inhoud van eventuele vervolgopleidingen kan ideeën opleveren voor het maken van toetsvragen. Het gaat om niet meer dan ideeën. Het kan niet de bedoeling zijn dat in de vooropleiding datgene wordt getoetst wat in de vervolgopleiding aan de orde komt. Primair uitgangspunt vormen de doelen van het 'eigen' onderwijs.

*Werkveld*

De meeste opleidingen zullen zich richten op een bepaalde beroepsuitoefening. De werkzaamheden in het toekomstige beroep kunnen ideeën opleveren voor het maken van vragen, mits men uiteraard die facetten uit het beroep neemt die in het te toetsen studieonderdeel worden behandeld. Ook hier moet men voorzichtig zijn. In het bijzonder moet men letten op het beheersingsniveau waarop de toets zich richt.

*Analyse van vaardigheid*

Er zijn onderwijssituaties waarin, naast een theoriemodule, practica of vaardigheidstrainingen zijn geprogrammeerd die inhoudelijke overeenkomst hebben met de theoriemodule. Bijvoorbeeld: binnen een verpleegkundige opleiding bestaat naast de theoriemodule 'Vervoer in het lichaam' de training 'Bloeddrukmeten'. Vragen over de kennis die studenten nodig hebben om bloeddruk te kunnen meten, zijn in de kennistoets op hun plaats.

*Antwoorden op gebruikte vragen*

De uiteenlopende antwoorden van studenten op open toetsvragen geven inspiratie voor nieuwe vragen of voor het formuleren van afleiders (dit zijn de niet-correcte alternatieven). Het komt regelmatig voor dat een studieonderdeel wordt getoetst met een combinatietoets, bestaande uit een gesloten en een open deel. Bij het construeren van de meerkeuzevragen is het zoeken naar aantrekkelijke afleiders vaak een probleem (zie ► hoofdstuk 11). Inspiratie voor het formuleren van een afleider kan men dan halen uit de (foute) antwoorden op open toetsvragen van een vergelijkbare toets.

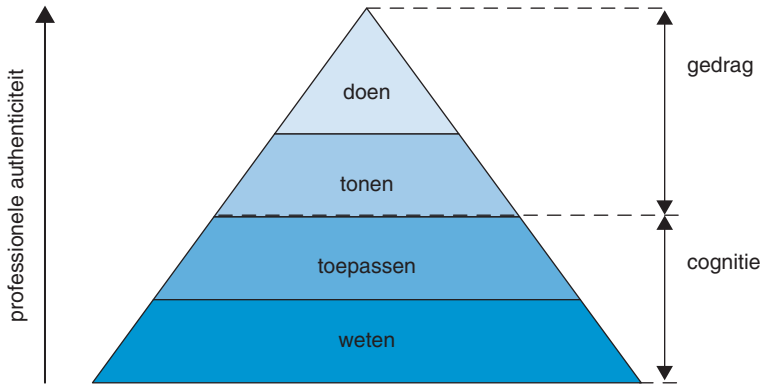
*Studenten*

Uit discussies tussen studenten tijdens het onderwijs kan de docent beluisteren waar studenten de nadruk leggen bij de bestudering van de leerstof. Kennelijk zijn dat aanstekelijke en voor studenten relevante onderwerpen geweest. Indien docenten hiervan korte aantekeningen maken, kunnen de notities later worden omgezet in toetsvragen. De mogelijke tegenwerping 'maar omdat ze er zo veel over gediscussieerd hebben, weten ze allemaal het antwoord!', is niet ter zake doende. Er is toch niets mis mee wanneer studenten bijna allemaal de vraag correct beantwoorden indien de vraag betrekking heeft op een relevant onderwerp? Kennelijk is het onderwijs op dit punt goed aangeslagen.

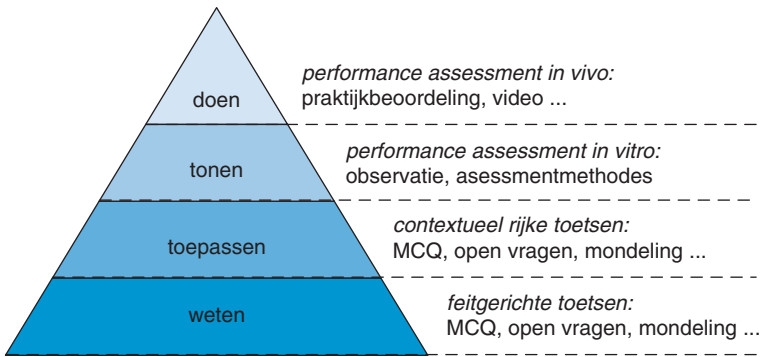
*Tussentijdse gedachten*

Gedurende de periode dat het onderwijs wordt voorbereid en gegeven, komen dikwijls gedachten op bij de docent over mogelijke toetsvragen. Het is nuttig deze gedachten meteen op te schrijven zodat ze later kunnen worden omgezet in toetsvragen.

toetsen van het laagste niveau. Maar wel dienen de toetsen 'rijk' te zijn. Dat wil zeggen, de vragen dienen plaatjes van praktische situaties te bevatten waarover studenten vragen beantwoorden. Bij computerafname kunnen dat ook filmpjes zijn. De hoogste twee niveaus, waar gedragingen moeten worden getoetst, zijn niet schriftelijke te toetsen (zie ook ► hoofdstuk 19). Docenten moeten daarin studenten observeren terwijl ze, bijvoorbeeld



■ **Figuur 1.2** De assessmentpiramide (Miller, 1990).



■ **Figuur 1.3** Assessmentpiramide en toetsvormen.

in de onderwijssituatie, een proef uitvoeren of een gesprek voeren. In de hoogste laag, het doen, gaat het om te toetsen of studenten in een 'echte' praktijksituatie, een werkplek (zie ► hoofdstuk 20) in staat zijn een vaardigheid uit te voeren.

Wanneer de docent eenmaal voldoende inspiratie heeft voor het construeren van toetsvragen, kan worden gestart met de reconstructie. De verschillende hoofdstukken in dit boek bevatten vele tips over hoe dit in zijn werk gaat. Voor specifieke constructieregels bij het formuleren van open en gesloten vragen wordt verwezen naar de ► hoofdstukken 11 en 12.

#### ■ Het controleren van de toetsvragen

Wanneer de toetsvragen zijn geconstrueerd, moet – voorafgaand aan de toetsafname – een indicatie worden verkregen over de kwaliteit van elke vraag door de vragen te controleren. Alle toetsvragen moeten voldoen aan inhoudelijke en vormtechnische eisen.

Over een aantal eigenschappen van een toetsvraag, zoals relevantie, eenduidigheid, objectiviteit en specificiteit, is vooraf een oordeel te geven. Een dergelijke controle kan bijvoorbeeld gebeuren door de vragen aan collega's voor te leggen en te bespreken. Over aspecten als de moeilijkheid en het onderscheidingsvermogen van een vraag is het lastig om vooraf een oordeel te geven – dit moet na afname worden gecontroleerd – zie onder stap 7 (en ► hoofdstuk 2).



### 1.2.5 Stap 5: Het samenstellen van de toets

---

Een toets is meer dan een verzameling vragen. De toets als geheel zal inhoudsvalide moeten zijn. Dat wil zeggen: de toets moet meten wat men wil meten. Om dit te bereiken is het in ieder geval noodzakelijk dat de doelstellingen evenwichtig in de toets aan de orde komen. Wanneer bijvoorbeeld ‘het verkrijgen van inzicht in ...’ een belangrijk doel is, dan is een toets die slechts bestaat uit vragen die een beroep doen op kennisreproductie beslist niet representatief. Het belangrijkste hulpmiddel om een representatieve toets te verkrijgen is het vooraf opstellen van een specificatietabel (zie stap 2). Met behulp van een specificatietabel kunnen docenten de toetsvragen verdelen over de belangrijkste onderwerpen van het studieonderdeel en over het gewenste niveau (kennisreproductie, toepassing, inzicht/begrip).

Tevens moet bij het samenstellen van de toets ook worden gedacht aan de studentinstructie, beoordelingscriteria en een antwoord- en scoremodel (bij open vragen).

### 1.2.6 Stap 6: Het afnemen van de toets

---

De condities waaronder een docent een toets afneemt, zoals de zaalruimte of temperatuur, kunnen een belangrijke invloed hebben op het toetsresultaat. De studenten moeten in elk geval voldoende tijd krijgen om de toets af te leggen – het is bijvoorbeeld in Nederland niet gebruikelijk een toets het karakter van een *speed test* te geven. Verder moet het de studenten vooraf bekend en duidelijk zijn wat zij precies kunnen verwachten. Wat zijn de onderwerpen van de vragen? Welke vragen zijn het belangrijkste en krijgen een hoger gewicht bij de beoordeling? Wat is de vraagvorm? Wanneer komt de uitslag? Ten behoeve van de transparantie is het goed om studenten dit soort informatie op voorhand te geven, bijvoorbeeld in de studiegids of in de vorm van een toetsinstructie. In de bijlage van dit hoofdstuk staat een uitgebreide toetsinstructie als voorbeeld.

### 1.2.7 Stap 7: Het nakijken van de toets, inclusief een kwaliteitscontrole

---

#### ■ Het nakijken van de toets

Het nakijken van de toets hoeft bij gesloten toetsen weinig tijd te kosten als dit per computer gebeurt. Met behulp van de antwoordsleutel is de uitslag in korte tijd bekend. Voor dat de uitslag bekend wordt gemaakt, moet duidelijk zijn dat de vragen kwalitatief goed zijn. Het nakijken van de vragen uit een open toets is veel tijdrovender omdat dit meestal handwerk is (zie ► paragraaf 9.4 voor het digitaal nakijken van open vragen). Het hanteren van een antwoordmodel bij het nakijken is onontbeerlijk (zie ► hoofdstuk 12). Voor het beoordelen van de toetstaken bij vaardigheidstoetsen of bij het uitvoeren van opdrachten dienen beoordelingsmodellen beschikbaar te zijn. Rubrics zijn hiervoor ook zeer geschikt (zie ► hoofdstuk 17).

### ■ Kwaliteitscontrole

Achteraf kan blijken dat er – ondanks alle zorgvuldigheid – kwalitatief toch iets is aan te merken op bepaalde vragen, toetstaken en/of de toets als geheel. Een groot voordeel van het computertijdperk is dat men op een betrekkelijk eenvoudige en snelle wijze inzicht verkrijgt in een aantal kwaliteitsaspecten van de vragen en van de toets als geheel: de moeilijkheid, het onderscheidingsvermogen van de vraag (beantwoordt de groep studenten die de stof beheerst de vraag correcter dan de groep die de stof minder beheerst?), de (on)aantrekkelijkheid van de afleiders en de betrouwbaarheid van de toets als geheel. Dit soort informatie is belangrijk op het moment dat een docent de toetsuitslag interpreteert. Wanneer de toets bijvoorbeeld zeer onbetrouwbaar is geweest (zie ► hoofdstuk 2 voor de normen hiervoor), is het niet eerlijk tegenover studenten om hun een cijfer te geven. Immers, een zeer onbetrouwbare toets betekent niets anders dan dat men net zo goed met een dobbelsteen de uitslag had kunnen bepalen. Bij het beoordelen van werkstukken kan de kwaliteitscontrole uitwijzen dat bepaalde docenten soepeler beoordelen dan anderen of dat beoordelingscriteria niet helder genoeg geweest zijn.

Ook kan de kwaliteitsanalyse docenten op het spoor zetten van inhoudelijke tekortkomingen van de vraag. Wanneer blijkt dat bijvoorbeeld slechts weinig studenten een bepaalde vraag correct beantwoorden, kan dit zijn veroorzaakt door de slechte voorbereiding van studenten, maar ook door de (toevallig) moeilijk uitgevallen vraag of door een onduidelijke vraagstelling. Als de analyse aantoont dat veel studenten een ander alternatief dan het correcte hebben gekozen, kan – bij nadere bestudering van de vraaginhoud – blijken dat dit alternatief óók correct is.

## 1.2.8 Stappen 8 en 9: Cijfergeven

---

Het bepalen van de zak/slaaggrens (de cesuur), het toekennen van cijfers en het geven van waarderingen aan de cijfers zijn sterk met elkaar verbonden. Rapporten en cijferlijsten drukken de waardering uit die de docenten hebben voor de leerprestaties van de studenten. Het waarderingscijfer voor deze leerprestaties wordt ontleend aan een cijferschaal die – althans theoretisch – loopt van 0, of 1, tot 10. Voor iedere stap op deze schaal bestaat een verbale kwalificatie die in alle gevallen een waarderende betekenis heeft. Zo is een acht ‘goed’ en een vijf ‘bijna voldoende’. Dergelijke kwalificaties impliceren dat er ergens grenzen zijn getrokken. Dat is echter niet eenvoudig omdat er geen hanteerbare methode is die in elke situatie toepasbaar is. Daarom resteert docenten niets anders dan uit het beschikbare aanbod die methode te kiezen die het beste is in de gegeven situatie. Daarvoor zijn wel enkele criteria te geven (zie ► hoofdstuk 7).

## 1.2.9 Stappen 10 en 11: Beslissen en rapporteren

---

Beslissingen binnen het onderwijs hebben een zeer bijzondere plaats gekregen. De telkens terugkerende spanning in een examenperiode maakt dit duidelijk. Vreemd is dit overigens niet want het zijn juist de beslissingen die een belangrijke betekenis hebben voor de

toekomst van studenten. De verlenging van een studieperiode met een jaar of meer, het niet toegankelijk zijn van de beoogde vervolgopleiding, ingedeeld worden bij een andere groep, in de knel raken met de nog beschikbare studieduur, een streep moeten zetten door uitgestippelde toekomstplannen enzovoort, zijn voorbeelden van belangrijke gevolgen van een toetsuitslag. Hoewel beslissingen die worden genomen op basis van hoge of lage cijfers onbetwistbaar zullen zijn, is het verschil tussen de nog net positieve en de nog net negatieve beslissing bijna altijd heel erg klein. Alleen al op die grond moet een zo groot mogelijke zorgvuldigheid met betrekking tot het nemen van beslissingen worden bepleit (Wijnen, 1993).

#### ■ Het rapporteren van de toetsuitslag

De toetsuitslag kan bestaan uit het behaalde cijfer plus eventueel de vermelding van de consequentie voor de studenten, bijvoorbeeld gezakt of (voorwaardelijk) geslaagd, en het aantal behaalde studiepunten. Een toetsuitslag kan ook functioneren als een feedbackinstrument wanneer studenten hieruit bijvoorbeeld tevens kunnen opmaken hoe zij hebben gescoord op onderdelen van de toets, op de afzonderlijke vragen of op de clustering hiervan tot bepaalde onderwerpen.

Naast een rapportage aan studenten dient de verantwoordelijke docent ook aan de opleiding c.q. de examencommissie te rapporteren. Uiteraard bestaat deze rapportage uit de cijfers per student ten behoeve van de administratie, maar de rapportage moet tevens enkele meest voor de hand liggende resultaten van de toets bevatten die een indicatie geven voor de kwaliteit van de toets, zoals: het gemiddelde cijfer, de verdeling van de cijfers, het aantal geslaagden en de betrouwbaarheid.

Deze rapportage kan dan ook gebruikt worden om de kwaliteit van de toets met collega's te bespreken en op basis hiervan het ontwerp van de toetsing voor een komende periode aan te passen.

#### Voorbeeld van een toetsinstructie

Toetsinstructie bij de toets *Inleiding in de Onderwijskunde*

Zo meteen krijg je de toets met de vragen uitgedeeld. Lees, voordat je begint, het onderstaande zorgvuldig. Je hebt 2,5 uur om de toets te maken; tegen het einde van de toets word je drie keer gewaarschuwd dat het inlevertijdstip nadert.

Invulinstructie

- Omcirkel bij de vijftig meerkeuzevragen het alternatief (A, B of C) dat naar jouw idee het beste antwoord bevat op de vraag. Breng vervolgens je keuze over naar het optisch leesbare formulier. Op dat formulier tref je een kolom aan met cijfers. De cijfers corresponderen met de nummers van de meerkeuzevragen. In een van de drie kolommen achter het cijfer zet je je keuze. Indien je 'A' antwoordt, zet dan een streepje tussen de lijntjes in de eerste kolom, met 'B' antwoordt je in de tweede kolom en kies je 'C', dan een streepje in de derde kolom.
- Geef in ieder geval een antwoord; fout gokken leidt niet tot strafpunten!
- Zet het streepje met een potlood, hardheid HB. Indien je het antwoord wilt veranderen, gum het dan helemaal weg en plaats het op een andere plaats opnieuw.

Mocht je je potlood en gum zijn vergeten, ga er dan *NU* een halen bij de surveillant.

- Het tweede gedeelte van de toets bestaat uit twaalf open vragen. Formuleer je antwoord direct in de ruimte onder de vraag. De vragen zijn zo gesteld dat de ruimte voor het antwoord voldoende is. Schrijf *NIET* in de kantlijn of op de achterzijde.

#### Scoring en uitslag

- Voor ieder correct antwoord op een meerkeuzevraag krijg je één punt. De twaalf open vragen leveren samen zestig punten op, iedere vraag vijf punten. Je kunt dus in totaal 110 punten behalen.
- De cijfertoekenning houdt rekening met hoe de toets is gemaakt door je medestudenten. Maar je kunt ervan op aan dat als je zestig punten hebt behaald, je in ieder geval een voldoende hebt.
- Indien je schrijft buiten de toegestane ruimte, wordt dit door de beoordelaar niet gelezen. Deze tekst maakt dus geen deel uit van het antwoord.
- Foute beweringen in je antwoord leiden tot puntenaftrek, ook als het goede antwoord er wel staat. De minimale score op de vraag is echter nooit negatief.
- In principe hangt over drie weken de uitslag van de toets op het mededelingenbord bij Bureau Onderwijs.

#### Algemeen

- Het is *NIET* toegestaan tijdens de toets hulpbronnen te raadplegen als een rekenmachine, woordenboek, aantekeningen, telefoon, iPod, iPad, enzovoort.
- Het is *NIET* toegestaan te communiceren met een medestudent, ook niet via het uitwisselen van potloden, papier, gum en/of pen.
- Vermeld allereerst op *IEDERE* bladzijde van de toets én op het optisch leesbare formulier je naam en studentnummer.
- Controleer de paginanummering van je toetsboekje. Indien een bladzijde ontbreekt, steek dan *NU* je hand op en vraag een nieuw toetsboekje.
- De examencommissie maakt je er op attent dat tijdens deze toets het fraudereglement zoals vermeld in het examenreglement van toepassing is.
- Indien je een vraag wilt stellen, steek dan je hand op.
- De eerste dertig minuten mag je de zaal niet verlaten, ook niet voor het bezoeken van het toilet.
- Indien je het toilet wilt bezoeken, steek dan je hand op en vraag toestemming aan de surveillant.
- Als je klaar bent met de toets, lever je de toets in bij de surveillant, legitimeer je je met je collegekaart, en verlaat de zaal.
- Na afloop van de toets word je verzocht een kort evaluatieformulier buiten de toetszaal in te vullen en in te leveren bij de surveillant.

(Uit: Van Berkel, 1999.)

# Toetsen: toetssteen of dobbelsteen

*Henk van Berkel en Anneke Bax*

## Inleiding

Hoe kunnen docenten rechtvaardig toetsen?

Hoe is onderscheid te maken tussen studenten die de stof wel en niet beheersen?

Hoe kan een toets representatief worden samengesteld?

Wat betekenen de begrippen validiteit, betrouwbaarheid, objectiviteit en transparantie?

Wat is een psychometrische analyse?

Hoe is de betrouwbaarheid van de toets te berekenen? En hoe hoog moet deze zijn?

Hoe is een toets na afname nog te verbeteren?

## 2.1 Kwaliteitseisen aan toetsen – 16

## 2.2 Kwaliteit van toets en toetsvragen – 16

## 2.3 Rechtvaardigheidseisen – 17

## 2.4 Psychometrische analyse: het berekenen van indicatoren – 19

2.4.1 Vraagmoeilijkheid (p-waarde) en normwaarden – 20

2.4.2 Onderscheidingsvermogen ( $R_{it}$ ) en normwaarden – 22

2.4.3 Betrouwbaarheid (coëfficiënt alfa) en normwaarden – 23

## 2.5 Reparatiemogelijkheden – 24

2.5.1 Veel vragen verwijderen verlaagt de representativiteit – 27

## 2.6 Ten slotte – 27

## 2.1 Kwaliteitseisen aan toetsen

De meeste toetsen functioneren in het hoger onderwijs als scherprechter. Hierbij gaat het om de situatie waarin de docenten ‘het kaf van het koren willen scheiden’, dat wil zeggen dat de toetsuitslag moet kunnen bepalen welke studenten de stof beheersen (competent zijn) en welke niet. Dit is de *summatieve* functie van een toets. Naast deze is er nog een andere belangrijke functie van studietoetsen, namelijk de *formatieve* functie. Wanneer een toets deze functie vervult, is het mogelijk aan de hand van de toetsprestaties te beoordelen welke onderwerpen studenten (nog) niet beheersen, en dus bijsturing behoeven.

► Hoofdstuk 8 gaat over formatieve toetsing.

Op basis van toetsuitslagen verkrijgen studenten hun diploma, of niet. Dat betekent nogal wat voor studenten. Vandaar dat een toetsuitslag een eerlijke en rechtvaardige uitspraak moet zijn over het kennen en kunnen van de student. Dit impliceert dat de kwaliteit van de toetsvragen van hoog niveau moet zijn en dat de toets daadwerkelijk moet meten wat bedoeld is te meten.

Dit hoofdstuk gaat over de kwaliteit van toetsen met open en/of gesloten vragen. Welke eisen mogen, of eigenlijk *moeten* worden gesteld aan deze toetsen en hoe kan hieraan zo goed mogelijk worden voldaan? De eisen die worden besproken (validiteit, betrouwbaarheid, objectiviteit, transparantie en normering) zijn samen te brengen onder de noemer ‘rechtvaardigheidseisen’. Vervolgens wordt uiteengezet hoe de kwaliteit van de afgenomen toets zichtbaar kan worden gemaakt – en zo nodig verbeterd – met behulp van een psychometrische analyse.

## 2.2 Kwaliteit van toets en toetsvragen

Er zijn verschillende fasen in het toetsproces waarin de kwaliteit van de toets bijzondere aandacht dient te verkrijgen. In de *constructiefase* moeten de toetsvragen voldoen aan inhoudelijke en vormtechnische eisen: heeft iedere toetsvraag betrekking op een doelstelling, is iedere toetsvraag relevant, is de formulering zorgvuldig? Bij de *samenstelling* van de toets zorgt een specificatietabel voor een goede vraagverdeling wat betreft de onderwerpen en het cognitieve niveau: is de toets representatief en valide, wordt inderdaad inzicht gevraagd waar dat is bedoeld? *Na* afname dient ook een kwaliteitscontrole plaats te vinden. Want ondanks alle zorgvuldigheid kan blijken dat er op bepaalde punten kwalitatief toch iets op de toets is aan te merken. Als een analyse bijvoorbeeld aantoont dat veel studenten een ander alternatief dan het correcte hebben gekozen, kan – bij nadere bestudering van de vraaginhoud – blijken dat dit alternatief óók correct is. Soms zijn er dan reparatiemogelijkheden die er aan bijdragen dat de toetsuitslag zo eerlijk mogelijk is. Aan het eind van dit hoofdstuk staan deze reparatiemogelijkheden.

Door de scores, toegekend aan de antwoorden van de studenten, psychometrisch te analyseren, zijn de moeilijkheid en het onderscheidingsvermogen van elke toetsvraag te berekenen en ook de betrouwbaarheid van de toets als geheel. Psychometrisch wil zeggen ‘getalsmatig’, dus niet inhoudelijk. Aan de getallen zijn normen gekoppeld die een indicatie over de kwaliteit van de toets(vraag) geven. Als getallen afwijken van de norm, is het van belang dat de docent de inhoud van de vraag opnieuw bestudeert. Als blijkt dat er inhou-

delijk ook wat is aan te merken op de vraag, kan de docent bijvoorbeeld besluiten om op basis van de inhoudelijke analyse de vraag al dan niet achteraf uit de toets te verwijderen, de score aan te passen of het modelantwoord te wijzigen.

Psychometrische waarden zijn dus vooral bedoeld om signalen af te geven. Het is niet aan te raden om uitsluitend op basis van een psychometrisch onderzoek kwalitatieve uitspraken over de toets(vragen) te doen. Inhoud is belangrijker.

## 2.3 Rechtvaardigheidseisen

Hiervoor is al aangegeven welke begrippen die betrekking hebben op toetsen onder de noemer 'rechtvaardigheid' zijn te categoriseren: validiteit, betrouwbaarheid, objectiviteit, transparantie en normering. In deze paragraaf staan hun definities en worden ze omschreven.

### ■ Validiteit

De validiteit van een toets heeft te maken met de betekenis, de bruikbaarheid en de juistheid van de conclusies die uit de toetsscores worden getrokken. In hoeverre meet de toets wat bedoeld is te meten? Al naargelang de aard van de conclusies kunnen drie typen validiteit worden onderscheiden: inhouds-, begrips- en criterium- (ook wel genoemd 'voorspelende' of 'predictieve') validiteit:

#### *Inhoudsvaliditeit:*

- De mate waarin de toets als geheel representatief is voor datgene wat men met de toets wil nagaan. Dit betekent onder andere dat ieder belangrijk onderwerp uit de leerstof in de toets moet worden bevraagd.

#### *Begripsvaliditeit:*

- De mate waarin de toetsscore kan worden beschouwd als een maat voor het bedoelde theoretische begrip. Zo is het niet valide om inzicht te toetsen door te vragen naar een uit het hoofd geleerde definitie.

#### *Criteriumvaliditeit:*

- De mate waarin de toetsscore samenhangt met soortgelijke metingen buiten de toets. Zo dienen bijvoorbeeld scores op toetsen over hetzelfde onderwerp aan elkaar te zijn gerelateerd.

Een hulpmiddel om ervoor te zorgen dat de toets inhouds- en begripsvalide is, is de specificatietabel. Een andere benaming hiervoor is 'toetsmatrijs'. Hierin staan de belangrijkste onderwerpen die uit de doelen zijn af te leiden en wordt het beheersingsniveau waarop de onderwerpen moeten worden getoetst vastgesteld (zie ► hoofdstuk 1). De criteriumvaliditeit is in de praktijk zelden zinvol te berekenen omdat inhoudelijk vergelijkbare toetsen om doelmatigheidsredenen niet worden afgenomen, met uitzondering van de voortgangstoets (zie ► hoofdstuk 15).

### ■ **Betrouwbaarheid**

De betrouwbaarheid van een toets geeft aan in hoeverre er vertrouwen kan zijn in de toets-score als meting, ongeacht de inhoud van de toets. Meet de toets echt iets of kan er net zo goed worden gedobbeld om uitspraken te doen over studenten? Als een weegschaal de ene keer 85 kilo aangeeft en vijf minuten later, met dezelfde persoon erop, 90, is de weegschaal niet betrouwbaar. Zo ook met toetsen. Als na een tweede afname, direct na de eerste, blijkt dat een vergelijkbare toets bij dezelfde deelnemers leidt tot geheel andere toetscores, is de toets niet betrouwbaar. Wanneer bijvoorbeeld een student bij de eerste toets als beste uit de bus komt en bij de tweede in de middenmoot eindigt, is dat merkwaardig. De correlatie tussen de scores op de beide toetsen is een aanwijzing voor de betrouwbaarheid van de toets, de zogenaamde *test-hertest* betrouwbaarheid. Deze correlatie is een belangrijke indicatie voor de kwaliteit van de toets als geheel. Omdat een hertest in het onderwijs niet voorkomt, zijn alternatieven bedacht om een indicatie van de betrouwbaarheid te verkrijgen met behulp van één toets. De meest bekende is Cronbach's coëfficiënt alfa.

De relatie tussen betrouwbaarheid en validiteit is vergelijkbaar met de relatie tussen een fundering en het huis dat erop staat. Een toeschouwer ziet alleen datgene wat zichtbaar is, de toets en de vragen. De fundering (de betrouwbaarheid) en alles wat daarop van invloed is als materiaal, waterstand, diepte, stevigheid, blijft verborgen. Slechts na een bodemonderzoek (toetsanalyse) zal de kwaliteit van de fundering blijken. Als de ondergrond drijfzand is (de toets is onbetrouwbaar), heeft het geen zin een huis (de toets) te bouwen, hoezeer de architecten ook hun best hebben gedaan om een stijlvol huis te ontwerpen. Je kunt ook andersom redeneren. Namelijk een stevig fundament leggen, dus een zeer hoge betrouwbaarheid, met daarop een gedrocht van een huis, bijvoorbeeld een huis met alleen maar badkamers. Dat 'huis' staat er over honderd jaar nog, maar mag geen huis heten omdat het niet voldoet aan de criteria voor een huis.

### ■ **Objectiviteit**

Over de eis van objectiviteit voor toetsen bestaan uiteenlopende opvattingen. In de meest gangbare wordt objectiviteit in verband gebracht met intersubjectiviteit. Dat wil zeggen dat de toetsuitkomst onafhankelijk moet zijn van storende invloeden van beoordelaars. Het mag er bijvoorbeeld niet toe doen wie de toets nakijkt. Intersubjectiviteit is niet alleen aan de orde in de fase van het nakijken, maar ook op het moment dat de vragen worden geconstrueerd en de sleutel (of het beoordelingsvoorschrift) door de docent wordt bepaald. Immers, het is mogelijk dat een docent al dan niet per abuis, een verkeerd antwoordmodel maakt of een verkeerd alternatief als correct aanmerkt.

### ■ **Transparantie**

In een toetssituatie is sprake van transparantie indien alle informatie voorhanden is die de best denkbare voorbereidings- en antwoordstrategie voor studenten mogelijk maakt. Het moet de toetsdeelnemers bijvoorbeeld duidelijk zijn op welke onderwerpen de docent de nadruk gaat leggen in de toets. Zijn ze alle even belangrijk? Of ligt het accent op een bepaald onderwerp? Tot welk niveau moet worden geleerd? Wat zijn de consequenties van een onvoldoende? Hoeveel tijd is beschikbaar voor hoeveel vragen? Hoe is de normering?



Over het algemeen kan aan de transparantie-eis worden voldaan door uitgebreid voorlichting te geven in het voorafgaande onderwijs en/of via een toetsinstructie bij de toets. Een voorbeeld van een toetsinstructie is opgenomen als bijlage van ► hoofdstuk 1.

#### ■ Normering

Normeren is het relateren van een toetsscore aan een criterium. Het normeren van toetsen is vooral belangrijk in grensgevallen waar tegengestelde beslissingen worden genomen. Bijvoorbeeld, daar waar twee dicht bij elkaar liggende scores (een '5+' en een '6-') tot twee verschillende beslissingen leiden. In dat geval moeten degenen die verantwoordelijk zijn voor de toets, hun beslissing 'hard kunnen maken'. Deze verantwoording is moeilijker dan wellicht in eerste instantie lijkt (zie De Groot, 1978). De Groot onderscheidt twee typen rechtvaardiging voor beslissingen. Eén die uitgaat van het voorspellend vermogen van een toets en een ander die een inhoudelijke rechtvaardiging tot basis heeft. De toets te zien als een voorspeller geeft zo veel problemen (Van Berkel, 2011), dat vaak niets anders rest dan de normeringeis te bezien in het kader van de relevantie of representativiteit van de toets. De docent zal moeten kunnen aantonen dat wat wordt getoetst, representatief is voor de doelstellingen van het onderwijs, maar de docent moet ook kunnen beargumenteren waarom deze doelen relevant zijn. Pas dan, in combinatie met een transparant toetssysteem, zijn beslissingen rondom kritieke zones te rechtvaardigen.

## 2.4 Psychometrische analyse: het berekenen van indicatoren

---

Wanneer de toets is afgenomen en de ruwe scores van de afzonderlijke vragen zijn berekend, kan de psychometrische analyse plaatsvinden. Het doel ervan is docenten op basis van de toetsscores inzicht te geven in de kwaliteit van de toetsvragen en van de toets als geheel. Bij de psychometrische analyse om de kwaliteit van de toets te analyseren zijn drie hoofdvragen van belang (Breukers e.a., 2011). Meten we consistent, of: is de toets betrouwbaar? Wat is de kwaliteit van de afzonderlijke items? En hoe hebben de studenten gescoord? Psychometrische waarden zijn – zoals al eerder opgemerkt – nuttig om signalen te geven. Het is beslist niet aan te raden om uitsluitend op basis van de uitkomsten van psychometrisch onderzoek *kwalitatieve uitspraken* te doen over de toets, omdat een psychometrisch onderzoek uitsluitend een kwantitatieve en geen inhoudelijke analyse betreft. Bovendien kleven er aan iedere kwantitatieve analyse onzekerheidsmarges die eventuele normen nuanceren. Desondanks is het van belang om een dergelijke analyse uit te voeren. Indien een toets bestaat uit twee onderdelen, bijvoorbeeld een open en een gesloten gedeelte, is het noodzakelijk op beide delen een analyse uit te voeren.

Om een psychometrische analyse te kunnen uitvoeren, is het vereist een elektronisch bestand aan te maken, bijvoorbeeld in Excel. De opbouw van een dergelijk bestand is simpel: op de rijen staan de scores op de toetsvragen van een individuele student. Bij meerkeuzevragen vormen de keuzes van studenten voor een van de alternatieven de scores, maar het kunnen ook nullen en enen zijn waarbij een nul staat voor een fout gekozen alternatief en een een voor het aanstrepen van het correcte alternatief, terwijl bij open toetsvragen de per vraag behaalde score staat.

Het is duidelijk dat een psychometrische analyse, zoals in dit hoofdstuk beschreven, uitsluitend van toepassing is op de toetsvormen met meerkeuzevragen en/of met open vragen. Voor andere toetsvormen, zoals *co-*, *peer-* en *self-*assessment, zijn (nog) geen hanteerbare psychometrische analysetechnieken ontwikkeld.

### ■ Indicatoren

Drie kwaliteitsindicatoren zijn minimaal noodzakelijk om zicht te verkrijgen op de kwaliteit van de toets – tussen haakjes staat de verkorte aanduiding:

1. de moeilijkheid van de toetsvraag (*p-waarde*);
2. het onderscheidingsvermogen van de toetsvraag tussen studenten die de stof wel en studenten die de stof niet beheersen ( $R_{ij}$ );
3. de betrouwbaarheid van de toets (*coëfficiënt alfa*).

Het berekenen van de indicatoren met computers is geen probleem. Op vrijwel iedere computer is tegenwoordig een statistisch pakket te installeren dat docenten in staat stelt enkele eenvoudige, maar noodzakelijke berekeningen uit te voeren. Er zijn veel kant-en-klare analyseprogramma's, maar een standaard statistisch pakket als SPSS voldoet, evenals Excel.

## 2.4.1 Vraagmoeilijkheid (p-waarde) en normwaarden

### ■ Gesloten toetsvraag

De vraagmoeilijkheid van een gesloten toetsvraag is de proportie studenten dat de vraag correct heeft beantwoord. Deze waarde wordt aangeduid met de p-waarde, waarbij de 'p' staat voor *probability*. Als de vraag bijvoorbeeld door zestig van de tachtig studenten correct is beantwoord, is de p-waarde  $60/80=0,75$ .

Het spreekt voor zich dat een gedeelte van de p-waarde van gesloten toetsvragen is veroorzaakt door het feit dat studenten door puur gokken al een kans hebben de vraag correct te beantwoorden. Dit is de raadkans. Zo is de raadkans bij een vierkeuzevraag 0,25, oftewel: er is een kans van 1 op 4 dat een student die niets van de inhoud af weet, de vraag toch correct beantwoordt. De p-waarde is hiervoor eenvoudig te corrigeren met behulp van de correctieformule:

$$P_c = P - (1-P) / (K-1)$$

$P_c$  = gecorrigeerde p-waarde

$P$  = oorspronkelijke p-waarde

$K$  = aantal alternatieven

Een voorbeeld: De p-waarde van een toetsvraag bestaande uit vier alternatieven is 0,83. De gecorrigeerde p-waarde is dan:

$$P_c = 0,83 - (1 - 0,83)/(4 - 1) = 0,83 - 0,17/3 = 0,83 - 0,06 = 0,77.$$

■ Tabel 2.1 Normen voor p-waarden bij summatieve toetsen.

	Optimale p-waarde	Ondergrens	Bovengrens
bij tweekeuzevragen	0,75	0,61	0,90
bij driekeuzevragen	0,67	0,50	0,90
bij vierkeuzevragen	0,62	0,44	0,90
bij open vragen	0,50	0,25	0,90

De gecorrigeerde p-waarde is een indicatie voor de proportie studenten dat de vraag correct weet te beantwoorden zonder te gokken.

#### ■ Open toetsvraag

De berekening van de vraagmoeilijkheid van open toetsvragen is wat bewerklijker omdat doorgaans – naast goed of fout – ook tussenliggende waarden mogelijk zijn. Indien studenten op een open toetsvraag bijvoorbeeld maximaal vijf punten kunnen behalen, zal de score waarden kunnen aannemen die lopen van 0 tot en met 5. De p-waarde van een open vraag is dan (computermatig) te berekenen door de score van iedere student op de betreffende vraag op te tellen en te delen door het aantal toetsdeelnemers, vermenigvuldigd met de maximale score. Deze berekening komt neer op het simpel berekenen van het gemiddelde op de toetsvraag, gerelateerd aan de maximale score. Als de gemiddelde score op een vraag 2,7 is en de maximale score 5, dan is de p-waarde:  $2,7/5 = 0,54$ . Een correctie voor raden is bij open vragen niet aan de orde omdat op dergelijke vragen de invloed van raden te verwaarlozen is.

#### ■ Normwaarden voor de vraagmoeilijkheid

De wenselijke hoogte van de p-waarde is afhankelijk van het doel waarvoor men toetst. Als het om een selectieve of summatieve toets gaat, moet iedere vraag uit de toets een bijdrage leveren aan deze summatieve functie. Statistisch is aan te tonen dat de maximale bijdrage wordt geleverd wanneer een toetsvraag een gemiddelde moeilijkheid heeft. Gemiddeld wil zeggen: precies op de helft tussen de maximale en minimale moeilijkheid.

Voor een *open toetsvraag* is deze waarde eenvoudig uit te rekenen. Stel, een toetsvraag levert een student maximaal vijf punten op. Dan is de wenselijke p-waarde  $2,5/5 = 0,50$ . Dit betekent dat, indien de gemiddelde score op deze toetsvraag 2,5 bedraagt, de vraag een maximale bijdrage levert aan de summatieve functie van de toets.

Het berekenen van de ideale p-waarde voor een *gesloten toetsvraag* is iets ingewikkelder. De raadkans speelt namelijk een rol. Een p-waarde die gelijk of lager is aan de raadkans, geeft geen zinvolle informatie over het kennisniveau van studenten. De beste p-waarde ligt in het midden tussen de maximale p-waarde (1,00) en de raadkans. Indien het een vijfkeuzevraag betreft, waar de raadkans dus  $1,00/5 = 0,20$  is, is de beste p-waarde  $(1,00 + 0,20)/2 = 0,60$ .

In **■** tabel 2.1 staan de p-waarden die bij een gegeven toetsvraag een maximale bijdrage leveren aan de summatieve functie van een toets (optimale p-waarde) en de waarden waarbuiten de vragen aan een inhoudelijke inspectie moeten worden onderworpen.

De stabiliteit van de p-waarde is afhankelijk van het aantal deelnemers aan de toets. Indien de groep toetsdeelnemers klein is ( $< 25$ ), zal men minder waarde kunnen hechten aan de gevonden p-waarde dan indien meer dan honderd studenten de toets hebben gemaakt.

## 2.4.2 Onderscheidingsvermogen ( $R_{it}$ ) en normwaarden

Elke toetsvraag in een summatieve toets moet onderscheid maken – *discrimineren* – tussen studenten die de doelstellingen van het studieonderdeel dat wordt getoetst wel of niet beheersen. Dus tussen studenten die hoog en laag scoren op de toets. Dit wordt het *onderscheidingsvermogen* van de vraag genoemd. Andere aanduidingen zijn: het discriminatievermogen of de discriminatie-index.

Het onderscheidingsvermogen wordt bepaald door de score op de betreffende vraag te relateren aan de eindscore op de toets. Toetsvragen die meer door hoog scorende studenten correct worden beantwoord in vergelijking tot studenten die laag scoren, zijn kenmerkend vragen die een onderscheid tussen beide groepen maken. Vragen die in gelijke mate door hoogscorers en laagscorers correct worden gemaakt, maken geen onderscheid. Vragen die laagscorers beter beantwoorden dan hoogscorers, onderscheiden weliswaar de beide groepen, maar doen dat op een verkeerde manier. Dit laatste vereist altijd nadere aandacht van de docent want meestal is er dan iets aan de hand, bijvoorbeeld een verkeerde sleutel.

### ■ Berekening van het onderscheidingsvermogen

De samenhang tussen de vraagscore en de eindscores is te bepalen door hun correlatie uit te rekenen. Dit kan met elk statistisch softwareprogramma. Deze correlatie wordt vaak kortheidshalve aangeduid met  $R_{it}$ , uit te spreken als R-I-T. De 'R' is een aanduiding voor de correlatie, terwijl 'it' aangeeft dat het hier de correlatie betreft tussen de vraag (item) en de eindscore (totale score), ook wel *item-testcorrelatie* genoemd. De eindscore is de optelling van de afzonderlijke vraagscores per toetsdeelnemer. Omdat de vraag waarvan men de  $R_{it}$  wil berekenen deel uitmaakt van de eindscore, is de correlatie eigenlijk te hoog. Nauwkeuriger is het om de eindscore te corrigeren door de score van de vraag waarvan men de  $R_{it}$  wil berekenen ervan af te trekken. In dat geval spreekt men van de  $R_{ir}$ , waarbij de laatste 'r' voor 'rest' staat: de *item-restcorrelatie*. Overigens, indien de toets uit 25 vragen of meer bestaat, is het verschil tussen  $R_{it}$  en  $R_{ir}$  minimaal; een correctie is dan niet nodig.

### ■ Normwaarden voor het onderscheidingsvermogen ( $R_{it}$ )

De  $R_{it}$  is een correlatiecoëfficiënt en heeft een maximale waarde van +1,00 en een minimale waarde van -1,00. Een  $R_{it}$ -waarde van +1,00 betekent dat alle studenten die hoog op de toets scoorden, de betreffende vraag correct hebben beantwoord en de studenten die laag op de toets scoorden, de betreffende vraag fout hebben beantwoord. Een waarde van -1,00 betekent dat alle studenten die laag op de toets scoorden, de vraag correct hebben beantwoord. In de toetspraktijk zullen deze uiterste waarden niet voorkomen. Door allerlei onvolkomenheden van een toetsvraag zal een vraag nooit door alle hoogscorers correct worden beantwoord en door alle laagscorers foutief. De waarde voor de  $R_{it}$  zal daarom ergens liggen tussen -1,00 en +1,00.

■ **Tabel 2.2** Normen voor het onderscheidingsvermogen ( $R_{it}$ ).

Gevonden waarden	Kwalificatie
0,35 en hoger	goed/zeer goed
0,25 - 0,35	voldoende/goed
0,15 - 0,25	middelmatig/voldoende
minder dan 0,15	slecht/middelmatig

De  $R_{it}$  is afhankelijk van verschillen in kennisniveaus onder de toetsdeelnemers. De schakering aan verschillen is bij een kleine groep toetsdeelnemers meer toevallig dan bij een grote groep. Daarom is de betekenis van de  $R_{it}$  bij weinig toetsdeelnemers ( $< 50$ ) niet hoog. Men moet dan de waarde voorzichtig interpreteren. In het algemeen geldt echter: hoe hoger het onderscheidingsvermogen van een toetsvraag, des te hoger de psychometrische kwaliteit. In ■ tabel 2.2 staat de kwalitatieve betekenis van  $R_{it}$ -waarden, ontleend aan Ebel en Frisbie (1991).

Een correct discriminerende vraag heeft een positieve  $R_{it}$ -waarde. Dit houdt in dat de vraag door de hoogscoorders beter is beantwoord dan door de laagscorders. Dat is ook de bedoeling van toetsvragen. Wanneer de  $R_{it}$  een negatieve waarde heeft, hebben vooral de hoogscoorders deze vraag foutief beantwoord. De meest voor de hand liggende verklaring hiervoor is bij gesloten toetsvragen een verkeerde sleuteling. Dat wil zeggen dat aan de computer bijvoorbeeld is opgegeven dat alternatief A correct is, terwijl dit alternatief B had moeten zijn. Dit vereist uiteraard meteen correctie, waarna de toetsanalyse opnieuw moet worden uitgevoerd, maar nu met de gecorrigeerde sleutel. Hetzelfde geldt voor de situatie waarin naast alternatief A alternatief B óók goed is. Daarnaast is het denkbaar dat er bij zo'n vraag een 'addertje onder het gras zit', waardoor de hoogscoorders op het verkeerde been zijn gezet. In elk geval vereist dit nadere aandacht van de docent.

Bij open vragen zijn verschillende verklaringen mogelijk voor een negatieve  $R_{it}$ . Misschien is er een tikfout in de vraag geslopen die studenten in verwarring heeft gebracht, of is de vraag anderszins onduidelijk geformuleerd.

### 2.4.3 Betrouwbaarheid (coëfficiënt alfa) en normwaarden

Bij de betrouwbaarheid van een toets gaat het om de stabiliteit van de toetsscores. Om te weten of de toets betrouwbaar is en of de meting iets te betekenen heeft, zou eigenlijk direct na de toetsafname een vergelijkbare toets moeten worden afgenomen onder dezelfde studenten. Dan pas kun je de zogenoemde *test-hertest* betrouwbaarheid berekenen. Een hertoets is in de praktijk niet haalbaar. Daarom zijn er alternatieven ontwikkeld waarvan onderzoeksmatig is aangetoond dat ze weinig verschillen van de test-hertestbetrouwbaarheid. De meest gangbare is coëfficiënt alfa. Een vergelijkbare index is de KR-20 die uitsluitend bij meerkeuzevragen toepasbaar is. De stabiliteit van een toets is daarmee te berekenen op basis van één enkele afname van de toets. Hierbij speelt de lengte van de toets een belangrijke rol. Doorgaans geldt: hoe meer vragen worden gesteld, des te betrouwbaarder

■ **Tabel 2.3** Normen voor de betrouwbaarheid (coëfficiënt alfa) van een summatieve toets.

Gevonden waarden	Kwalificatie
0,90 en hoger	goed/zeer goed
0,80 - 0,90	voldoende/goed
0,70 - 0,80	middelmatig/voldoende
minder dan 0,70	slecht/middelmatig

is de toets. Met zestig vragen in een meerkeuzetoets is een goede betrouwbaarheid te bereiken.

Coëfficiënt alfa gaat uit van het opsplitsen van de toets in deelttoetsen: alle vragen worden ieder voor zich opgevat als een deelttoets. De meeste statistische softwareprogramma's bevatten een mogelijkheid om deze coëfficiënt te berekenen.

#### ■ Normwaarden voor coëfficiënt alfa

De mogelijke waarden die alfa kan aannemen, liggen tussen de 0 (niet betrouwbaar) en 1 (maximaal betrouwbaar). Wanneer een toets niet betrouwbaar is, kan men geen enkele betekenis hechten aan de toetsuitslag. Wanneer een toets honderd procent betrouwbaar is, heeft men in ieder geval de zekerheid dat de toetsuitslag iets heeft te betekenen. Of de toetsuitslag ook *zinnvolle* betekenis heeft, staat los van de betrouwbaarheid. Dat moet blijken uit de bestudering van de inhoudsvaliditeit van de toets die voorafgaande aan de toets heeft moeten plaatsvinden. Wel is betrouwbaarheid een voorwaarde voor inhoudsvaliditeit. Wanneer een toets onbetrouwbaar is, is het niet mogelijk dat de toets wel inhoudsvalid is. De kwalitatieve betekenis van de waarden tussen 0 en 1 is afhankelijk van de functie die de toets heeft. Normwaarden voor summatieve toetsen staan in ■ tabel 2.3.


Bovenstaande normen gelden voor toetsen met een summatieve functie, zonder compensatiemogelijkheden. In sommige gevallen zijn lagere betrouwbaarheden dan 0,80 acceptabel. Dat geldt bijvoorbeeld wanneer de verkregen toetsscore kan worden gecompenseerd door scores op een andere toets, zoals bij veel studietoetsen het geval is; een betrouwbaarheid van 0,60 en hoger is dan acceptabel. In die gevallen mag worden aangenomen dat de somscore van meerdere toetsen samen betrouwbaarder zijn dan de afzonderlijke betrouwbaarheden.

## 2.5 Reparatiemogelijkheden

Het zwaartepunt van de kwaliteitscontrole dient voorafgaande aan de toetsafname plaats te vinden. Ondanks een zorgvuldige toetsconstructie kan de toets echter na de afname toch op bepaalde punten kwalitatief tekort blijken te schieten. Wanneer (na afname van de toets) enkele kwaliteitsindicatoren onder de maat blijken te zijn, bestaan er een paar mogelijkheden om de kwaliteit op te krikken, zij het beperkt. Hieronder staat een aantal uitkomsten van de drie besproken kwaliteitsindicatoren. Steeds wordt besproken of de toetsverantwoordelijke maatregelen moet nemen en welke dat moeten zijn.

*P-waarde is lager dan of gelijk aan de raadkans.* Wanneer de  $p$ -waarde lager dan of gelijk is aan de raadkans, is de meest waarschijnlijke oorzaak dat de vraag foutief is gesleuteld: een van de afleiders is per abuis tot het correcte antwoord bestempeld. Als dat na een inhoudelijke controle inderdaad het geval blijkt te zijn, is de remedie makkelijk: sleutel de vraag correct en voer de psychometrische analyse opnieuw uit.

De vraag kan ook een zogenaamd ‘instinkertje’ zijn. Studenten zijn op het verkeerde been gezet en kiezen *en masse* een verkeerd antwoord. De verantwoordelijke docentervaart zoiets wellicht als een overwinning (op wat?), maar studenten zien dat niet zo. En daar hebben ze gelijk in. Het kan toch niet de bedoeling van een toets zijn studenten ergens ‘in te luizen’, tenzij hiervoor zwaarwegende inhoudelijke argumenten zijn. Zo’n vraag dient uit de toets te worden verwijderd en de analyse moet opnieuw worden uitgevoerd.

*P-waarde is hoger dan de raadkans, maar beduidend lager dan de wenselijke waarde* (zie  tabel 2.1). Een lage  $p$ -waarde is een indicatie voor een zeer moeilijke vraag. Op zichzelf zijn moeilijke vragen toegestaan in toetsen. Zeer goede studenten kunnen er hun ‘10’ mee behalen. Studenten die minder goed zijn, zullen de vraag foutief beantwoorden. Of deze laatste redenering klopt, moet blijken uit de  $R_{it}$ -waarde van de vraag. Als deze ruim positief is, is er niets op tegen om de vraag te handhaven. Toch moeten docenten ervoor waken dat niet te veel van dergelijke moeilijke vragen deel uitmaken van de toets. Een toets met veel moeilijke vragen kent meestal een hoog zakpercentage en dat is niet altijd verdedigbaar.

Is de  $R_{it}$  sterk negatief, dan blijkt daaruit dat niet de hoogscorende studenten de moeilijke vraag correct hebben beantwoord, maar de laagscorende. In dat geval is het zaak om de vraag inhoudelijk opnieuw te beschouwen, het antwoord zo nodig aan te passen of de vraag uit de toets te verwijderen en de analyse opnieuw uit te voeren.

*P-waarde is (bijna) één.* Een  $p$ -waarde die (bijna) de maximale waarde van één aanneemt, betekent dat (praktisch) iedere student de vraag correct heeft beantwoord, een zeer makkelijke vraag dus. Een dergelijke vraag discrimineert daardoor niet tussen goede en slechte studenten. Toch is er geen doorslaggevende reden de vraag te verwijderen uit de toets. Kennelijk is het onderwijs over het onderwerp waar de vraag betrekking op heeft, goed bij studenten overgekomen. Studenten moeten dan niet worden ‘gestraft’ door hen geen punten te geven voor de vraag.

*$R_{it}$  is negatief of 0.* Een negatieve  $R_{it}$ -waarde betekent dat meer goede studenten de vraag fout hebben beantwoord dan slechte studenten. Bij meerkeuzevragen duidt dit over het algemeen op een foutieve sleuteling van de vraag. Blijkt dit na een inhoudelijke controle inderdaad het geval, dan is dit te verhelpen door de vraag correct te sleutelen en de analyse opnieuw uit te voeren. Wanneer blijkt dat de sleuteling toch juist is geweest, is een negatieve  $R_{it}$  moeilijk verdedigbaar. Het is in dat geval aan te bevelen de betreffende vraag (achteraf) uit de toets te verwijderen en de analyse opnieuw te verrichten. Dit geldt ook voor negatieve  $R_{it}$ -waarden bij open toetsen.

*$R_{it}$  is positief, maar lager dan 0,15.* Een vraag met een lage  $R_{it}$  maakt weinig onderscheid tussen hoog- en laagscorende studenten. In dat opzicht is het geen goede vraag, maar op zichzelf is dat onvoldoende reden om te adviseren de vraag uit de toets te verwijderen. Wanneer de  $p$ -waarde van de vraag ook discutabel is, bijvoorbeeld in de buurt van de kansscore, vormen deze twee gegevens in samenhang voldoende reden om de vraag uit de toets te verwijderen en de analyse opnieuw uit te voeren.

■ **Tabel 2.4** Reparatiemogelijkheden op basis van een psychometrische analyse.

P-waarde	Reparaties
p-waarde is lager dan of gelijk aan de raatkans	Is de vraag correct gesleuteld? Zo niet, sleutel wijzigen en analyse opnieuw uitvoeren. Zo ja, vraag uit de toets verwijderen en analyse opnieuw uitvoeren.
p-waarde is hoger dan de raatkans, maar beduidend lager dan de wenselijke waarde	Indien $R_{it}$ positief is, de vraag handhaven. Indien $R_{it}$ negatief is, de vraag verwijderen en de analyse opnieuw uitvoeren.
p-waarde is (bijna) 1	De vraag handhaven.
$R_{it}$ is negatief of 0	Is de vraag correct gesleuteld? Zo niet, sleutel wijzigen en analyse opnieuw uitvoeren. Zo ja, vraag uit de toets verwijderen en de analyse opnieuw uitvoeren.
$R_{it}$ is positief, maar lager dan 0,15	Indien de p-waarde eveneens laag is, de vraag uit de toets verwijderen en de analyse opnieuw uitvoeren.
alfa is lager dan 0,60	Toetsuitslag uitsluitend als formatief beschouwen.
alfa is hoger dan 0,60 maar lager dan 0,80	Indien de toets kan worden gecompenseerd met andere toetsen, is deze uitkomst niet verontrustend. Anders vragen met een lage of negatieve $R_{it}$ uit de toets verwijderen, indien dit niet ten koste gaat van de representativiteit, en de analyse opnieuw uitvoeren.

*Coëfficiënt alfa lager dan 0,60.* Een betrouwbaarheidswaarde van lager dan 0,60 is te laag om doorslaggevende beslissingen op te baseren. Er zullen in dat geval te veel onjuiste zak/slaagbeslissingen worden genomen. Een lage betrouwbaarheid impliceert dat er rond iedere toetsuitslag een relatief breed onzekerheidsgebied ligt. Daarom is het onverstandig een toets met een te lage alfa voor summatieve doelen te gebruiken. Indien een toets een dergelijke lage betrouwbaarheid heeft, zal blijken dat relatief veel vragen een negatieve  $R_{it}$ -waarde zullen hebben. Over het algemeen wordt aangeraden die vragen uit de toets te verwijderen, wat de betrouwbaarheid zal verhogen. Dan zal de betrouwbaarheid wellicht een meer acceptabele waarde gaan aannemen. Hier is echter een waarschuwing op zijn plaats die staat verwoord in de volgende paragraaf.

*Coëfficiënt alfa hoger dan 0,60 maar lager dan 0,80.* Een betrouwbaarheidswaarde tussen 0,60 en 0,80 is middelmatig en daarmee discutabel. Eigenlijk is de waarde te laag om er doorslaggevende beslissingen op te baseren. Indien de toetsuitslag wordt gecompenseerd met een andere toetsscore, is een alfa-waarde tussen 0,60 en 0,80 acceptabel. Overigens geldt ook hier dat allereerst de vragen met een negatieve  $R_{it}$  uit de toets moeten worden verwijderd, mits dit om inhoudelijke redenen is te verantwoorden.

■ Tabel 2.4 geeft een samenvatting van het hiervoor besprokene.



### 2.5.1 Veel vragen verwijderen verlaagt de representativiteit

---

In dit hoofdstuk is op sommige plaatsen geadviseerd vragen uit de toets te verwijderen en de analyse opnieuw uit te voeren. Toch kan men dat niet ongelimiteerd doen. Indien de toets op een correcte wijze is samengesteld, dat wil zeggen dat de vragen op een correcte manier zijn verdeeld met behulp van een specificatietabel, en er dus sprake is van een representatieve toets, doet het achteraf verwijderen van vragen afbreuk aan de representativiteit. Als bijvoorbeeld blijkt dat het merendeel van de vragen over een bepaald onderwerp wordt verwijderd, is de toets niet meer representatief. Daarom zullen docenten er bij de besluitvorming over het verwijderen van vragen altijd voor moeten waken dat de representativiteit niet in het gedrang komt. Zij moeten in dat geval dus kiezen uit twee kwaden.

## 2.6 Ten slotte

---

Elke toets met open of gesloten vragen moet voldoen aan een aantal kwaliteitseisen. Voorafgaande aan de toetsafname is het bereiken van kwaliteit het best te garanderen door de vragen zorgvuldig te construeren, door de toets samen te stellen met gebruikmaking van een specificatietabel en door voldoende vragen in de toets op te nemen. Een aantal van die eisen is, na de toetsafname, in een getal uit te drukken. De belangrijkste indicatoren voor de toetskwaliteit zijn, naast de inhoudsvaliditeit/representativiteit, de p-waarde (vraagmoeilijkheid), de  $R_{it}$  (onderscheidingsvermogen) en de coëfficiënt alfa (betrouwbaarheid). Voor deze indicatoren bestaan normwaarden. Indien die normwaarden niet worden bereikt, zullen – zo mogelijk – maatregelen moeten worden genomen om de kwaliteit van de toets te verhogen. Hiermee is bereikt dat de toets geen dobbelsteen maar een toetssteen is.

# Ontwikkelen van toetsbeleid

*Riet Martens en George Moerkerke*

## Inleiding

Wat is toetsbeleid en wat is de rol ervan bij de kwaliteit van toetsing?

Wat is de relatie met accreditatie?

Wat is de samenhang tussen toetskaders, toetsbeleidsplannen en toetsplannen?

Wat is de rol van examencommissies?

Welke rollen kun je als docent hebben bij het bevorderen, bewaken en borgen van toetskwaliteit?

## 3.1 De aandacht voor toetsbeleid – 30

## 3.2 Integraal toetsbeleid: een samenhangend stelsel – 31

## 3.3 Toetsbeleid en kwaliteit van toetsing op de verschillende organisatieniveaus – 33

3.3.1 Toetsbeleid op instellingsniveau: functie en inhoud toetskader – 33

3.3.2 Toetsbeleid op niveau van de faculteit: functie en inhoud toetsbeleidsplan – 34

3.3.3 Toetsbeleid op niveau van een opleiding: functie en inhoud toetsplan – 35

## 3.4 Borging van kwaliteit van tentamens en examens door de examencommissie – 37

3.4.1 De rol van de examencommissie – 37

3.4.2 Regels en richtlijnen van de examencommissie – 38

## 3.5 De docent en het toetsbeleid – 38

3.5.1 De docent als examiner – 38

3.5.2 De docent als lid van een examencommissie – 39

3.5.3 De docent als lid van een toetscommissie – 39

## 3.6 Ten slotte – 40

Het is voor hoger onderwijsinstellingen, maar ook voor toekomstige werkgevers, van cruciaal belang dat de instelling garant kan staan voor het eindniveau van de student. De garantie dat afgestudeerden voldoen aan de eindkwalificaties van een opleiding is er alleen als kan worden aangetoond dat er degelijk wordt getoetst. Aandacht voor toetsing en toetsbeleid is er daarom volop.

‘Wie betaalt, die bepaalt’ is een uitdrukking die beslist van toepassing is op het Nederlandse hoger onderwijs waarvan de belangrijkste financier de rijksoverheid is. Met de Wet op het Hoger Onderwijs en Wetenschappelijk onderzoek (WHW) regelt en stuurt de overheid de plichten en verantwoordelijkheden van de onderwijsinstellingen in het hoger onderwijs. De overheid bepaalt dus in hoge mate het handelen van de besturen van de hogescholen en de universiteiten. De – sinds de tweede druk van dit boek – toegenomen aandacht voor toetsbeleid is rechtstreeks te herleiden tot veranderingen in de WHW in 2010. De Inspectie van het Onderwijs, een uitvoerende dienst van het ministerie van Onderwijs, Cultuur en Wetenschappen (OCW), geeft de volgende veelgebruikte definitie (2003, p. 55):

» Integraal toetsbeleid is het samenhangende stelsel van kwaliteitszorg waarin maatregelen en voorzieningen zijn getroffen om de kwaliteit van toetsing en examinering te bewaken en te bevorderen. «

Via de Inspectie houdt OCW toezicht op het accreditatiestelsel voor het hoger onderwijs en op het functioneren van de Nederlands-Vlaamse Accreditatie Organisatie (NVAO). De NVAO ziet toe op de beoordeling van de kwaliteit van de opleidingen in het hoger onderwijs in Nederland. De NVAO stelt daartoe beoordelingskaders vast die in lijn zijn met de WHW. Onder gezag van de NVAO worden de opleidingen (en instellingen) in het hoger onderwijs beoordeeld.

### 3.1 De aandacht voor toetsbeleid

De overheid heeft wat betreft het hoger onderwijs geen bemoeienis met inhoud en vorm van toetsing en examinering. Wel heeft de overheid een aantal regels ingesteld om een goede gang van zaken te garanderen. Zo dienen instellingen voor elke opleiding of groep van opleidingen een examencommissie in te stellen, is er een accreditatiestelsel voor opleidingen en houdt de Inspectie van het Onderwijs toezicht op de kwaliteit van het onderwijs. De afgelopen jaren (vanaf ongeveer 2010) heeft de overheid een aantal veranderingen doorgevoerd om niveau en kwaliteit van opleidingen en de toetsing daarbinnen te verhogen.

- Examencommissies kregen een meer expliciete rol. De examencommissie is het orgaan dat op objectieve en deskundige wijze vaststelt of een student voldoet aan de voorwaarden die de onderwijs- en examenregeling stelt ten aanzien van kennis, inzicht en vaardigheden die nodig zijn voor het verkrijgen van een graad (WHW; 7.12). Met de publicatie van de Wet Versterking Besturing in 2010 is beoogd de examencommissie een meer onafhankelijke positie te geven ten opzichte van het instellingsbestuur en het opleidingsmanagement. De examencommissie heeft onder meer als belangrijke taak ‘het borgen van de kwaliteit van tentamens en examens’ (...) (WHW art 7.12 b lid 1a).

- De onderwijsinstellingen kunnen de NVAO verzoeken tot een instellingstoets kwaliteitszorg. Instellingsbesturen kunnen zo laten zien dat ze ‘in control’ zijn voor wat betreft de kwaliteit van de opleidingen, dus ook de kwaliteit van de tentamens en examens.
- De NVAO stelt hogere minimumeisen aan de kwaliteit van tentamens en examens bij accreditaties van opleidingen. Het nieuwe beoordelingskader voor bestaande opleidingen (waarvan de instelling de instellingstoets heeft behaald) kent drie standaarden waaronder: De opleiding beschikt over een adequaat systeem van toetsing en toont aan dat de beoogde eindkwalificaties worden gerealiseerd (NVAO kader, november 2011). Een onvoldoende op deze standaard voor toetsing leidt direct tot een eindoordeel dat onvoldoende is. Compensatie door een goede beoordeling op een andere standaard is niet meer mogelijk.

Sinds de aandacht voor toetsbeleid is vertaald naar wetswijzigingen, hebben kenniscentra voor het onderwijs en verenigingen van onderwijsinstellingen publicaties uitgegeven met aanwijzingen om toetsbeleid te ontwikkelen en te implementeren. De Vereniging Hogescholen (voorheen HBO-raad) bijvoorbeeld heeft een handreiking examencommissies uitgegeven (2011), met daarin aanbevelingen voor het implementeren van de veranderingen in de WHW van 2010.

Dus: de overheid en de samenleving vinden dat examinering een belangrijk onderdeel vormt van het onderwijs aan hogescholen en universiteiten. Dat heeft geleid tot wetgeving waarin de verantwoordelijkheden voor het leveren van kwaliteit (instellingsbestuur), het borgen van kwaliteit (examencommissie) en het toezien op kwaliteit (NVAO) scherper zijn vastgelegd. Een gevolg hiervan is dat het opzetten en uitvoeren van toetsbeleid – het samenhangend stelsel van kwaliteitszorg – hoog op de agenda is komen te staan.

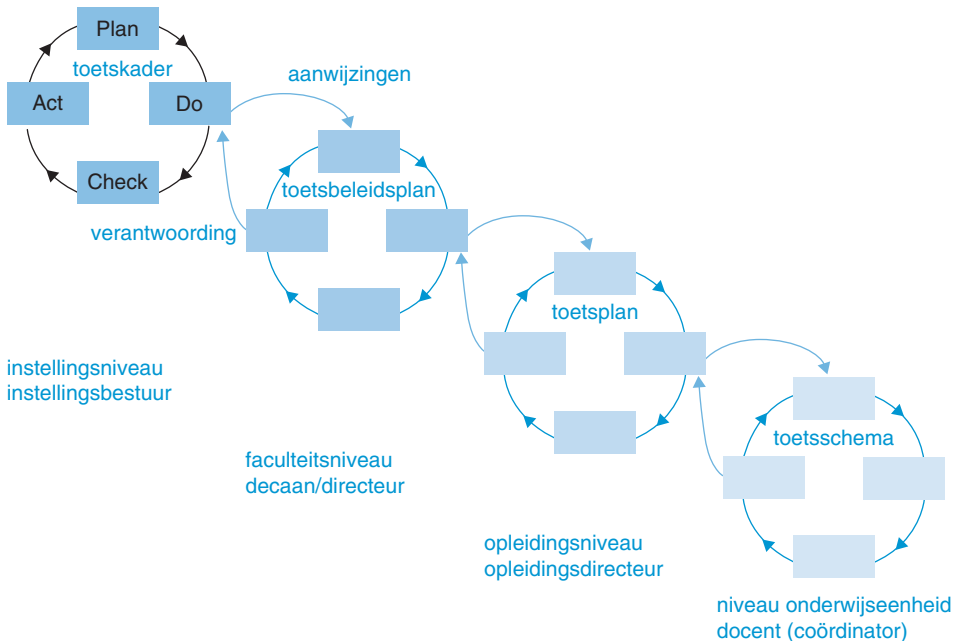
### 3.2 Integraal toetsbeleid: een samenhangend stelsel

---

Volgens de definitie van de onderwijsinspectie is integraal toetsbeleid een samenhangend stelsel maatregelen en voorzieningen met als specifiek doel de kwaliteit van toetsing en examinering te bewaken en te bevorderen. Dat samenhangend stelsel kan worden beschreven aan de hand van verantwoordelijkheden van de belangrijkste spelers in de organisatie, hun activiteiten en de documenten waaruit hun activiteiten blijkt.

Op hogescholen en universiteiten is de PDCA-cyclus (Plan-Do-Check-Act) een veel gebruikt hulpmiddel in de kwaliteitszorg om kwaliteitsverbetering te bereiken. De cyclus werkt als volgt. In de plan-fase worden (beleids)plannen vastgesteld. In de do-fase worden de plannen uitgevoerd en wordt de uitvoering ervan gemeten. In de check-fase worden de metingen afgezet tegen de plannen. In de act-fase worden op grond van de analyse nieuwe plannen gemaakt en daarmee start weer een nieuwe cyclus.

Ten aanzien van toetsing wordt binnen onderwijsinstellingen een aantal van die PDCA-cycli uitgevoerd op verschillende organisatieniveaus: niveau van de instelling, niveau van



■ **Figuur 3.1** Integraal toetsbeleid: een samenhangend stelsel van maatregelen op verschillende niveaus.

een faculteit, niveau van een opleiding en niveau van een onderwijseenheid. Zo ontstaat het cyclische plaatje in ■ figuur 3.1

Uiteraard moet er samenhang zijn tussen de verschillende niveaus en uitwisseling van informatie (zie ■ figuur 3.1). Het instellingsbestuur stelt kaders waarbinnen het management van een faculteit of school dient te opereren (toetskader van de instelling). Een faculteitsbestuur of decaan werkt dit kader uit in concrete aanwijzingen voor de opleidingen die onder de faculteit vallen (facultair toetsbeleidsplan). Deze aanwijzingen hebben betrekking op de opleiding als geheel en (op hoofdlijn) op onderwijseenheden binnen de opleiding. De opleidingsbesturen of -directeuren stellen voor iedere opleiding een toetsplan op waarin onderwerpen worden geregeld als de wijze waarop de opleiding ervoor zorgt dat studenten de eindkwalificaties van de opleiding bereiken en de samenhang van de toetsen in de opleiding. Een docent van een onderwijseenheid (cursus of module) maakt in lijn met het toetsplan van de opleiding een toetsschema waarin onder andere wordt vastgelegd wat de inhoud van de toetsen is, en hoe deze worden afgenomen en beoordeeld.

Tussen de organisatieniveaus is niet alleen een top-downcommunicatie van strategisch beleid naar operationele uitvoering. Er is ook een bottom-uprichting in de communicatie: de 'lagere' organisatieniveaus moeten verantwoording afleggen ten opzicht van de 'hogere', en kunnen daarbij ook verzoeken doen tot heroverweging van het beleid (zie ■ figuur 3.1). Opleidingsdirecteuren zullen in de act-fase van de PDCA-cyclus op opleidingsniveau bijvoorbeeld verslag doen aan de decaan van de faculteit. Zij zullen daarbij rapporteren in hoeverre het toetsbeleid van de faculteit is gerealiseerd, maar ook zullen zij opvallende zaken rapporteren (in de managementliteratuur ook wel exep-tierapportage genoemd). Op

grond daarvan zou een opleidingsdirecteur de decaan bijvoorbeeld kunnen vragen om een pilot te starten met een instapcursus wiskunde.

De documenten die op de verschillende niveaus rond toetsing worden vastgesteld, hebben niet alleen een rol bij de communicatie en afstemming (= samenhang) tussen de verschillende organisatieniveaus, maar vooral ook in de kwaliteitsbewaking en -verbetering op elk niveau.

### 3.3 Toetsbeleid en kwaliteit van toetsing op de verschillende organisatieniveaus

---

Het is van belang dat er op de verschillende organisatieniveaus, naast het plannen van toetsbeleid, ook sprake is van jaarlijkse meting met betrekking tot de beleidsplannen en dat de resultaten worden gebruikt voor verbetering van de kwaliteit (PDCA-cyclus). Welke informatie moeten de documenten in ieder geval bevatten of op welke vragen moeten zij antwoord geven om tot kwaliteitsverbetering te kunnen komen? Deze paragraaf gaat in op het toetsbeleid en de bewaking en bevordering van toetskwaliteit op instellingsniveau (3.3.1), faculteitsniveau (3.3.2), en opleidingsniveau (3.3.3).

#### 3.3.1 Toetsbeleid op instellingsniveau: functie en inhoud toetskader

---

Het instellingsbestuur is de meest gezaghebbende beleidsmaker binnen een instelling. Het bestuur analyseert de externe ontwikkelingen, brengt deze in verband met de eigen visie op besturen en op opleiden, onderzoekt verschillende maatregelen voor de aansturing van de interne processen en maakt daarin keuzen. Deze keuzen legt het instellingsbestuur vast in het kader voor facultair toetsbeleid, kortweg het toetskader. Het toetskader houdt rekening met de wettelijke kaders en politieke en maatschappelijke vragen. Met het toetskader bepaalt het instellingsbestuur de ruimte in de faculteiten.

Een instellingsbreed toetskader moet zich beperken tot de hoofdlijnen; bijvoorbeeld een opmerking dat de opleiding ervoor zorgt dat toetsvormen passend zijn, dat de gehanteerde cesuur passend is en dat dit wordt onderbouwd en vastgelegd in het toetsplan dat voor elke opleiding aanwezig moet zijn. Volgens Molkenboer (2011) is het logisch dat elke faculteit een eigen toetsbeleidsplan heeft en elke opleiding een toetsplan. 'Een opleiding binnen een faculteit bètawetenschappen zal anders willen toetsen dan binnen letteren en anders dan binnen geneeskunde. Kennis over de juiste opleidings- en toetsvormen moet worden over gelaten aan zij die weten wat van de student in de praktijk wordt verwacht.' (p. 14)

In het toetskader zal het instellingsbestuur minimaal op hoofdlijn de onderwerpen moeten en willen regelen die te maken hebben met:

- de eisen die de wettelijke kaders en maatschappelijke en politieke context (van de minister) stellen ten aanzien van de toetsing in het hoger onderwijs; bijvoorbeeld 'elke toets voldoet aan de kwaliteitseisen validiteit, betrouwbaarheid en transparantie'

- (accreditatiekader NVAO) of ‘in een examencommissie moet minimaal één extern lid worden benoemd’ (in de WHW is sinds 2010 die mogelijkheid geschapen);
- de onderwijsvisie van de instelling en aansluitend de visie op toetsing;
  - beleid dat op instellingsniveau wordt vastgesteld zoals personeelsbeleid; bijvoorbeeld alle docenten moeten een basiskwalificatie onderwijs (BKO) hebben (toetsing is daar een onderdeel van);
  - maatregelen die het instellingsbestuur voor de hele instelling hetzelfde wil hebben geregeld om ‘in controle’ te kunnen zijn wat betreft de kwaliteit van toetsing, bijvoorbeeld dat elke opleiding een toetsplan heeft;
  - zaken die een andere laag in de organisatie (formeel) niet kan regelen; bijvoorbeeld centraal overleg tussen de examencommissies of de wijze waarop decanen de examencommissie faciliteert bij de uitvoering van de wettelijke taken;
  - punten die beter op hoger organisatieniveau kunnen worden geregeld omdat iedereen anders het wiel moet uitvinden, zoals (in grote lijn) taken, bevoegdheden en verantwoordelijkheden met betrekking tot (de uitvoering) van toetsing.

#### ■ **Kwaliteitsbewaking en -verbetering op instellingsniveau**

Welke rol speelt het toetskader bij de kwaliteitsbewaking en -verbetering ten aanzien van toetsing op instellingsniveau? Kwaliteitsverbetering indiceert dat het toetskader jaarlijks wordt geëvalueerd en op grond van de bevindingen aangepast (PDCA-cyclus). Hoe kan het CvB op instellingsniveau de kwaliteit bewaken?

In de plan-fase geeft het CvB aan welke aspecten met betrekking tot de kwaliteit van toetsing zij de komende periode wil realiseren of wil dat de faculteiten realiseren. Het instellingsbestuur kan maatregelen uitzetten ten behoeve van uitvoering van beleid, bijvoorbeeld door middelen beschikbaar te stellen voor de uitvoering (of ondersteuning). In de do-fase werken de faculteiten de plannen uit in toetsbeleid voor de faculteit en meten de daadwerkelijke gang van zaken. In de check-fase bespreekt het instellingsbestuur (meestal in een jaarlijkse gespreksronde waar ook andere aspecten van kwaliteitszorg worden besproken) met de faculteitsbesturen in hoeverre het uitgezette beleid is gerealiseerd. In de act-fase analyseert het instellingsbestuur de resultaten van de verschillende faculteiten en plant op grond van deze analyse nieuwe acties. Het toetskader wordt aangepast en gecommuniceerd en de cyclus start opnieuw.

### **3.3.2 Toetsbeleid op niveau van de faculteit: functie en inhoud toetsbeleidsplan**

Het faculteitsbestuur maakt in lijn met het toetskader van de instelling een toetsbeleid voor de faculteit. Het facultaire toetsbeleidsplan geldt voor alle opleidingen die onder verantwoordelijkheid van de faculteit worden verzorgd. Over het algemeen zijn dat opleidingen die redelijk ‘bij elkaar passen’. Wat het faculteitsbestuur in het toetsbeleid zal willen en moeten regelen, is deels afhankelijk van de mate waarin het instellingsbestuur een en ander in het toetskader heeft geregeld en van de mate waarin de faculteit eigen aanvullend beleid heeft. Het is raadzaam om het facultaire toetsbeleid af te stemmen met de examencommissie(s) van de betreffende opleidingen; immers, die borgen de kwaliteit van

de tentamens en examens (zie paragraaf ‘borging van kwaliteit van examens en tentamens door de examencommissie’).

In het facultaire toetsbeleidsplan zal het faculteitsbestuur in ieder geval willen opnemen:

- de visie op toetsing (aansluitend op de onderwijsvisie);
- richtlijnen voor het format van het toetsplan van een opleiding (veelal in afstemming met de examencommissies), bijvoorbeeld over het opnemen van de samenhang tussen leerlijnen en eventuele leerwegen;
- manier waarop de toetsplannen van opleidingen moeten worden ingebed in de kwaliteitszorg in het algemeen;
- het minimum aan (toets)deskundigheid voor docenten (bijvoorbeeld BKO met toetsmodule) en het minimum aan bijscholingsverplichting (bijvoorbeeld jaarlijks één dag);
- de wijze waarop het niveau van de eindwerkstukken wordt bewaakt (bijvoorbeeld elke opleiding heeft een beoordelingsformulier voor beoordeling van alle eindopdrachten);
- minimumeisen die aan alle tentamens en toetsen worden gesteld (bijvoorbeeld scoringsformulieren);
- de wijze waarop de opleidingsdirecteuren de examencommissie faciliteren in het uitvoeren van haar wettelijke taken;
- eventueel de uitvoering van interne onderwijsaudits volgens vast protocol.

#### ■ **Kwaliteitsbewaking en -verbetering op facultair niveau**

Welke rol speelt het toetsbeleidsplan bij de kwaliteitsbewaking en -verbetering ten aanzien van toetsing op faculteitsniveau? In het kader van de kwaliteitsverbetering zal het toetsbeleid jaarlijks worden geëvalueerd en op grond van de bevindingen.

In de plan-fase geeft het faculteitsbestuur aan welke aspecten met betrekking tot de kwaliteit van toetsing zij de komende periode wil realiseren/wil dat de opleidingen realiseren. Het faculteitsbestuur kan maatregelen uitzetten ten behoeve van uitvoering van beleid, bijvoorbeeld door middelen beschikbaar te stellen voor de uitvoering (of ondersteuning). In de do-fase werken de opleidingsdirecteuren/-besturen de plannen uit in het toetsplan van de opleiding en meten de actuele stand van zaken. In de check-fase bespreekt het faculteitsbestuur in een gespreksronde met de opleidingsdirecteuren in hoeverre het uitgezette beleid is bereikt, zetten zij de actuele stand van zaken af tegen het uitgezette beleid en gaan na in hoeverre het beleid is gerealiseerd. In de act-fase analyseert het faculteitsbestuur de resultaten en plant op grond van deze analyse nieuwe acties. Faculteitsbesturen rapporteren en bespreken resultaten naar/met instellingsbestuur. Het toetsbeleidsplan wordt aangepast en de cyclus start opnieuw.

### **3.3.3 Toetsbeleid op niveau van een opleiding: functie en inhoud toetsplan**

De opleidingsdirecteur stelt in lijn met het facultair toetsbeleidsplan een toetsplan op. Het toetsplan van een opleiding heeft als doel:



- transparant te maken vanuit welke visie en op welke wijze wordt getoetst; liefst zodanig beschreven, dat het bij (internationale) accreditatie rechtstreeks bruikbaar is;
- duidelijk te maken op welke wijze de opleiding er zorg voor draagt dat studenten de eindkwalificaties bereiken;
- de kwaliteitsverbetering en -bewaking op het gebied van toetsing inzichtelijk te maken.

Wat moet er in een toetsplan zijn opgenomen om aan deze doelen te voldoen? Dat hangt af van hoever een opleiding is met het bevorderen van de toetskwaliteit. Niet alles kan tegelijk worden aangepakt en er zal altijd een evenwicht moeten zijn met de middelen die beschikbaar zijn. In het toetsplan van de opleiding zal een opleidingsdirecteur in ieder geval willen opnemen:

- de onderwijsvisie en daarvan afgeleide toetsvisie van de opleiding;
- hoe het geheel van toetsen in het toetsprogramma past bij de onderwijsvisie (bijvoorbeeld: soort toetsvormen, spreiding, variatie);
- de wijze waarop de eindkwalificaties van de opleiding worden geoperationaliseerd in leerdoelen van de onderwijseenheden/wijze waarop de leerdoelen van de verschillende onderwijseenheden bijdragen aan de eindkwalificaties van de opleiding. Het is zinvol om bij jaarlijkse wijzigingen van (de leerdoelen van) onderwijseenheden de examencommissie hiermee in te laten stemmen;
- hoe de eindkwalificaties aansluiten op de Dublin-descriptoren in verband met niveau van de opleiding;
- hoe de groei binnen een leerlijn wordt geoperationaliseerd (bijvoorbeeld in een voortschrijdende scoring op dezelfde rubrics; zie ► hoofdstuk 17);
- eisen die worden gesteld aan docenten (bijvoorbeeld: elke docent moet BKO hebben of begeleiders van eindwerkstukken moeten gepromoveerd zijn);
- hoe de opleiding de minimumkwaliteit van toetsen (validiteit, betrouwbaarheid en transparantie) bewaakt (bijvoorbeeld: een toets moet altijd worden voorgelegd aan collega's, beoordelingen met het cijfer 5 moeten altijd door een tweede beoordelaar worden bekeken);
- de maatregelen die de opleiding neemt om bijvoorbeeld fraude, plagiaat en meeliftgedag te voorkomen.

#### ■ **Kwaliteitsbewaking en -verbetering op opleidingsniveau**

Door het hanteren van de PDCA-cyclus, ook voor de toetsing, kan jaarlijks worden nagegaan in hoeverre de geplande resultaten ook daadwerkelijk zijn bereikt en op welke aspecten verbetering mogelijk is. Welke rol speelt het toetsplan bij de kwaliteitsbewaking en -verbetering ten aanzien van toetsing op opleidingsniveau? In het kader van de kwaliteitsverbetering zal het toetsplan jaarlijks worden geëvalueerd en op grond van de bevindingen aangepast (PDCA-cyclus).

In de plan-fase geeft de opleidingsdirecteur aan welke aspecten met betrekking tot de kwaliteit van toetsing hij of zij de komende periode wil realiseren. Een en ander wordt vastgelegd in het toetsplan van de opleiding. De opleidingsdirecteur kan maatregelen

uitzetten ten behoeve van uitvoering van beleid, bijvoorbeeld door middelen beschikbaar te stellen voor de uitvoering (of ondersteuning). In de do-fase wordt het toetsplan ingevuld. Docententeams voeren de toetscyclus uit bij de onderwijseenheden, coördinerende docenten bewaken de samenhang in de opleiding, er worden trainingen voor professionalisering gevolgd, en de eindwerkstukken leiden tot afsluiting van de opleiding. In de check-fase verzamelt de opleidingsdirecteur de nodige informatie en bespreekt deze met de betrokkenen (docentteams, coördinatoren). Dit leidt tot een jaarrapportage (dit kan een specifiek tentamenjaarverslag zijn of een onderdeel van het opleidingsjaarverslag). In de plan-fase analyseert de opleidingsdirecteur de resultaten en plant op grond van deze analyse nieuwe acties. Het toetsplan wordt aangepast en de cyclus start opnieuw.

### **3.4 Borging van kwaliteit van tentamens en examens door de examencommissie**

---

Hogescholen en universiteiten zijn verantwoordelijk voor de kwaliteit van hun onderwijs en de kwaliteit van de toetsing daarvan. Een aantal geldende procedures en rechten en plichten met betrekking tot het onderwijs en de examens moeten bij wet (WHW; 7.13) voor elke (groep van) opleiding(en) al zijn opgenomen in de onderwijs- en examenregeling (OER). Andere onderwerpen worden geregeld in een toetskader, toetsbeleidsplannen en toetsplannen. Vanwege het belang van toetsing heeft de overheid ook – in diezelfde WHW – vastgesteld dat voor elke opleiding of groep van opleidingen een examencommissie moet worden ingesteld.

#### **3.4.1 De rol van de examencommissie**

---

De examencommissie is het orgaan dat op objectieve en deskundige wijze vaststelt of een student voldoet aan de voorwaarden die de onderwijs- en examenregeling stelt ten aanzien van kennis, inzicht en vaardigheden die nodig zijn voor het verkrijgen van een graad (WHW; 7.12). Het instellingsbestuur verleent de graad en de examencommissie geeft het getuigschrift af. Ook heeft de examencommissie onder andere de taak om de kwaliteit van examens en tentamens te borgen: onafhankelijk nagaan of de kwaliteit voldoet aan de eisen. Borgen betekent zekerheid geven. De examencommissie heeft dus als taak zekerheid te geven over de kwaliteit van de tentamens en examens van de opleiding of groep van opleidingen. Dit kan zij onder andere doen door een onafhankelijk oordeel te geven over de kwaliteit van de toetsing, bijvoorbeeld aan de hand van vooraf gestelde kwaliteitscriteria. Een eerste stap voor de examencommissie is vast te leggen wat toetskwaliteit impliceert. Zo heeft bijvoorbeeld de examencommissie van de OU voor de borging van toetskwaliteit de volgende perspectieven onderscheiden: opleiding, onderwijseenheid, eindwerkstukken, examinatoren. Voor elk perspectief zijn indicatoren vastgelegd waar de examencommissie op zal letten (Moerkerke, 2012).

### 3.4.2 Regels en richtlijnen van de examencommissie

---

De examencommissie legt regels vast over de uitvoering van haar taken (WHW; 7.12b) en dus ook over hoe zij de kwaliteit van tentamens en examens borgt. Daarmee is zij transparant over de uitvoering. De examencommissie zou dit kunnen doen aan de hand van protocollen die zij gebruikt bij het onafhankelijk meten/beoordelen van de kwaliteit. Het ligt voor de hand dat het zinvol is dat de maatregelen en voorzieningen die het management ten aanzien van (kwaliteit van) toetsing heeft vastgelegd in haar documenten en de protocollen van de examencommissie op elkaar zijn afgesteld.

De examencommissie zal ten aanzien van de borging van de kwaliteit in ieder geval nagaan of:

- examinatoren die zij voornemens is aan te wijzen, toetsbekwaam zijn;
- in het toetsplan van de opleiding transparant is aangegeven hoe de eindkwalificaties van de opleiding zijn geoperationaliseerd in de leerdoelen van onderwijseenheden;
- alle leerdoelen adequaat worden getoetst;
- het geheel aan toetsen er zorg voor draagt dat alle individuele studenten de eindkwalificaties bereiken;
- alle toetsen voldoen aan kwaliteitseisen;
- beoordelingen voldoen aan kwaliteitseisen.

De examencommissie zal ook richtlijnen en aanwijzingen vastleggen om de uitslag van tentamens en examens te beoordelen en vast te stellen (WHW; 7.12b lid b). De examencommissie kan hiermee haar taak ten aanzien van de borging van kwaliteit vormgeven, bijvoorbeeld door een richtlijn te geven dat eindwerkstukken altijd moeten worden beoordeeld door twee examinatoren.

## 3.5 De docent en het toetsbeleid

---

Wat heeft nu een docent te maken met al dat toetsbeleid?

Het is, mede in verband met de opleidingsaccreditatie, van belang dat opleidingen beschikken over een adequaat systeem van toetsing en een goede kwaliteit van tentamens en examens. De docenten die het onderwijs van een opleiding verzorgen, kunnen in sterke mate bijdragen aan de kwaliteit. Een docent kan daarbij verschillende rollen toebedeeld krijgen: als examinator, als lid van een examencommissie (van de eigen of van een andere opleiding) of als lid van een toetscommissie.

### 3.5.1 De docent als examinator

---

Een docent kan als examinator worden aangewezen. In de WHW staat daar het volgende over:

- Voor het afnemen van tentamens en het vaststellen van de uitslag daarvan wijst de examencommissie examinatoren aan.
- De examinatoren verstrekken de examencommissie de gevraagde inlichtingen.

De examencommissie zal vooral examinatoren aanwijzen die bewezen kundig zijn op het gebied van toetsing. Docenten die examiner worden, mogen verwachten dat zij goed worden geïnstrueerd door de examencommissie of opleidingsdirecteur. Een examiner mag ook verwachten dat de faculteit zorgt voor de toetsdeskundigheid en gerichte ondersteuning, bijvoorbeeld in de vorm van handleidingen, cursussen, of job aids.

De examiner is als eerste verantwoordelijk voor de kwaliteit van het tentamen bij een onderwijseenheid. Een examiner dient er voor te zorgen dat de toetsing aan alle eisen voldoet zoals die zijn vastgelegd in het toetsplan. Als er bij een onderwijseenheid meer docenten betrokken zijn bij de toetsing (bijvoorbeeld als beoordelaar of als toetsconstructeur), dan dient de examiner er voor te zorgen dat deze goed worden geïnstrueerd. De inlichtingen die de examencommissie kan vragen, hebben vaak betrekking op onderwerpen die te maken hebben met de kwaliteit van toetsen bijvoorbeeld slaagcijfers, betrouwbaarheid, toetschema's. Meer en meer gaan examencommissies ook jaarlijks gesprekken aan met een aantal examinatoren om na te gaan of de instructies voldoende worden nageleefd.

### 3.5.2 De docent als lid van een examencommissie

---

Een docent kan worden gevraagd als lid van een examencommissie. Een docent die in de examencommissie bij de eigen opleiding zit, krijgt een andere relatie met de opleidingsdirecteur. Enerzijds valt de docent in de hiërarchische lijn onder de opleidingsdirecteur, anderzijds dient de docent als lid van de examencommissie wel in volledige onafhankelijkheid de kwaliteit van de opleiding te borgen. Een lid van de examencommissie die ook docent is, mag verwachten dat het instellingsbestuur voldoende *checks and balances* heeft ingesteld om in conflictsituaties beschermd te zijn.

Een trend is om externe leden te benoemen in de examencommissies. Vóór 2010 bepaalde de WHW dat een examencommissie samengesteld diende te zijn uit de docenten van de opleiding. De huidige WHW biedt ruimte voor de benoeming van externe leden. In het maatschappelijk debat is er discussie of dergelijke benoemingen niet moeten worden verplicht gesteld. Externe leden kunnen een positieve invloed hebben op de onafhankelijkheid van de examencommissie, de toetsdeskundigheid van de examencommissie en de praktijkrelevantie van de tentamens en examens.

### 3.5.3 De docent als lid van een toetscommissie

---

Vooraf in het hbo ontstaan steeds vaker toetscommissies die opleidingsbestuur en collega-docenten ondersteunen. Deze toetscommissies zijn groepen van getrainde/gespecialiseerde docenten en onderwijskundigen. Zij ondersteunen onder andere het opleidingsbestuur bij het maken van toetsplannen, en docenten bij het opstellen van toetschema's en specificatietabellen. In universiteiten worden deze ondersteunende taken vaak uitgevoerd door opleidingscoördinatoren en onderwijskundigen. De opleidingsdirecteuren sturen toetscommissies aan. In wezen voeren zij grote delen uit van de PDCA-cyclus op opleidingsniveau. Voorbeelden van taken die toetscommissies uitvoeren:

- uitwerken van toetsbeleidsplan naar een concreet toetsplan voor een opleiding
- implementeren en bewaken van de uitvoering van het toetsplan per opleiding
- bewaken van de kwaliteit van de toetsen en het geven van feedback
- adviseren van docenten over toetsing
- evalueren en bijstellen van het toetsplan per opleiding
- rapporteren over de kwaliteit van de uitvoering van het toetsplan
- doen van voorstellen en projecten ter verbetering van de kwaliteit van toetsen.

Toetscommissies kunnen ook worden ingesteld om de examencommissie te ondersteunen bij de steekproefsgewijze controles op de kwaliteit van toetsing. Het opleidingsbestuur en de examencommissie blijven uiteraard verantwoordelijk voor respectievelijk het zorgen voor kwaliteit en het borgen van kwaliteit. Het is daarbij niet wenselijk om de ondersteuning van deze twee taken bij dezelfde commissie neer te leggen.

### 3.6 Ten slotte

---

Er zullen medewerkers in het hoger onderwijs zijn die integraal toetsbeleid een bureaucratisch circus vinden. Het is echter wel de manier waarop instellingen voor hoger onderwijs kunnen laten zien dat zij zorgvuldig omgaan met hun maatschappelijke opdracht en de middelen die de overheid daarvoor beschikbaar stelt. Daarbij moet er, zoals eerder gesteld, een evenwicht zijn tussen de middelen die beschikbaar zijn voor het onderwijs en het kwaliteitsniveau dat moet en kan worden nagestreefd. Er zullen binnen het toetsbeleid altijd aspecten zijn die kunnen worden verbeterd, maar niet alles kan en hoeft tegelijk te worden aangepakt. Instellingen en opleidingen kunnen hiervoor meerjarenplannen maken. Een toetsplan van een opleiding bijvoorbeeld zal ook een beschrijving en analyse geven van de huidige stand van zaken met betrekking tot toetsing. Het kan als een soort werkdocument worden beschouwd, waarbij zwakke punten niet worden verdoezeld, maar aanleiding zijn om op een geëigend moment verbeteringen door te voeren.

# Toetsen met landelijke kennistoetsen

*Arnoud van Leuven en Wim Lansu*

## Inleiding

Waarom een landelijke kennistoets?

Wat is de rol van de docent?

Wat is te bereiken met een landelijke kennistoets?

Wat zijn de voor- en nadelen?

Welke voorwaarden gelden voor succesvolle toepassing?

- 4.1 **Waarom een landelijke kennistoets: de gemeenschappelijke kennisbasis – 42**
- 4.2 **Voorbeeld: de rol van de docent – 43**
- 4.3 **Mogelijkheden – 46**
- 4.4 **Voordelen – 47**
- 4.5 **Nadelen – 48**
- 4.6 **Voorwaarden voor een succesvolle toepassing – 49**

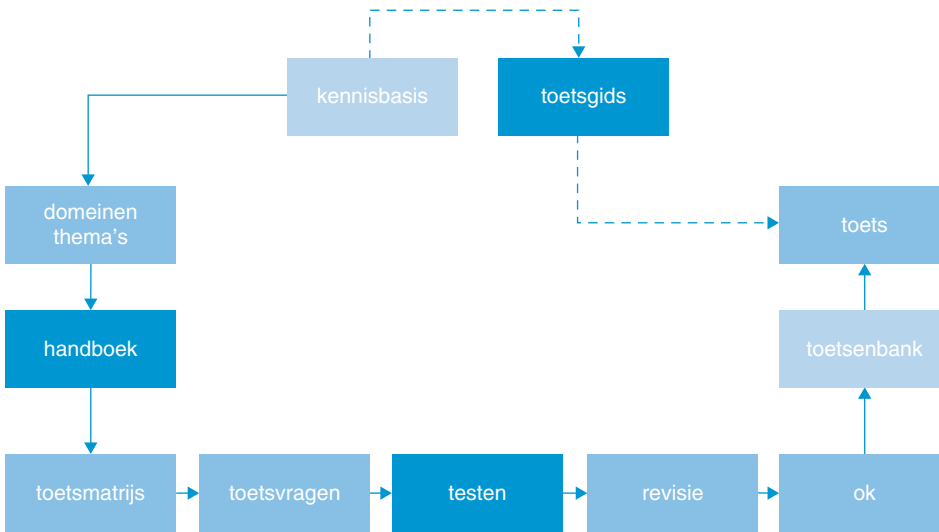
In 2011 stelt de toenmalige HBO-raad, nu de Vereniging Hogescholen, de Commissie externe validering examenkwaliteit hoger beroepsonderwijs in, ook wel Commissie Bruijn genoemd. Aanleiding voor het instellen van de commissie is de ongerustheid in de samenleving over de examenkwaliteit in het hbo. In haar rapport 'Vreemde ogen dwingen' stelt de commissie dat 'de samenleving het zich eenvoudigweg niet kan veroorloven dat er twijfels bestaan over de examinering binnen een voor Nederland zo belangrijke onderwijssector als het hbo' (Commissie Externe Validering, 2011, p. 12). De commissie komt met een breed palet aan aanbevelingen. De eerste is de externe validering van toetsing door middel van instellingsoverstijgende toetsing, ook wel leerwegonafhankelijke of landelijke toetsing genoemd.

Dit hoofdstuk gaat over landelijke kennistoetsen in het hoger onderwijs. Als voorbeeld dient daarbij de landelijke kennistoets voor de tweedegraads lerarenopleidingen Geschiedenis. Alle studenten die de opleiding volgen, of dat nu in Leeuwarden of in Sittard is, nemen deel aan een landelijke kennistoets. Daarmee voldoen studenten bij hun afstuderen in het hele land aan uniforme eisen wat betreft de vakinhoud. De toets is een direct gevolg van afspraken tussen het ministerie van OCW en de lerarenopleidingen over het verhogen en borgen van de kwaliteit van de kenniscomponent in de opleidingen (ministerie van OC&W, 2008). Bij de totstandkoming van de toets spelen docenten een belangrijke rol.

#### 4.1 **Waarom een landelijke kennistoets: de gemeenschappelijke kennisbasis**

---

Rond 2005 klinken in de politiek kritische geluiden ten aanzien van de kwaliteit van de lerarenopleidingen. Is het wel goed gesteld met de vakkennis van beginnende leraren? In reactie hierop formuleren de lerarenopleidingen per vak een kennisbasis: een systematische beschrijving van wat leraren aan vakkennis moeten bezitten. Om te voorkomen dat deze kennisbasis een vrijblijvend karakter zou hebben, besluiten de opleidingen (verenigd in de Vereniging Hogescholen) in overleg met het ministerie van OCW om na vaststelling van de gemeenschappelijke kennisbasis te starten met de ontwikkeling van een landelijke kennistoets. De docenten van de lerarenopleidingen houden het initiatief zelf in handen door vanuit hun midden een redactie samen te stellen om deze kennisbasis te schrijven. De docenten die in de redactie plaatsnemen, redeneren vaak in eerste instantie vanuit het curriculum zoals dat op hun eigen instelling bestaat. Echter, enerzijds dient er een middenweg gevonden te worden tussen de curricula van de verschillende lerarenopleidingen en anderzijds moet de kennisbasis op een robuuste en herkenbare wijze de kern van het vak weergeven. Voor vaststelling bekijkt een panel uit wetenschap en schoolpraktijk de voorgestelde kennisbasis en valideert – in het geval van Geschiedenis na ingrijpende aanpassingen – de voorgestelde tekst.



■ **Figuur 4.1** Het proces van kennisbasis tot landelijke kennistoets.

## 4.2 Voorbeeld: de rol van de docent

In deze paragraaf komen de verschillende rollen van docenten bij de totstandkoming van de landelijke kennistoets aan bod. Het gaat daarbij om de docent als: 1. Opsteller van een toetsmatrijs; 2. Ontwerper van toetsvragen; 3. Beoordelaar van toetsvragen (inhoudelijk en toetstechnisch). In ■ figuur 4.1 staat het proces vanaf de kennisbasis tot aan de definitieve toets.

De kennisbasis Geschiedenis (zie ► [www.10voordeleraar.nl](http://www.10voordeleraar.nl)) beschrijft de kennis die startende tweedegraads leraren Geschiedenis minimaal moeten bezitten en vormt het vertrekpunt voor de toetsredactie. Opnieuw, naar analogie van de kennisbasisredactie, vormen docenten van de lerarenopleidingen de redactie. Ze ontwikkelen samen digitale toetsen die toetsen of aanstaande leraren Geschiedenis in Nederland de kennisbasis beheersen. Om binnen de redactie tunnelvisie te vermijden en de kwaliteit te borgen, zijn twee commissies aan de toetsredactie gekoppeld.

- De vakcommissie bestaat uit docenten die deskundig zijn met betrekking tot de inhoud en de vorm van te ontwikkelen toetsvragen. De voorzitter is een extern deskundige, afkomstig van een expertisecentrum en/of universiteit, en beschikt over vakspecifieke wetenschappelijke expertise.
- De kwaliteitscommissie bestaat uit docenten met specifieke toetsexpertise, bijvoorbeeld op grond van hun lidmaatschap van een toets- of examencommissie. De voorzitter is een wetenschappelijk geschoolde toetsexpert.

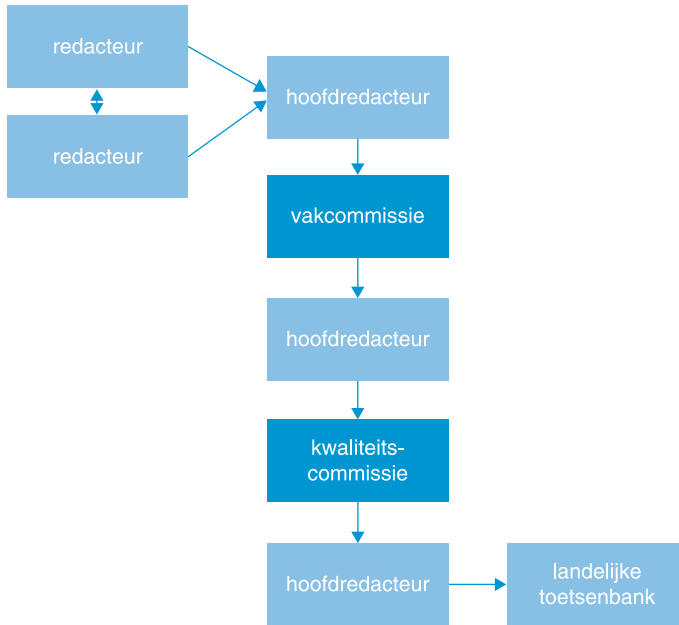
Alle stappen die de redactie neemt in de toetsconstructie, worden door één van deze commissies becommentarieerd. Op basis van de kennisbasis heeft de redactie met de



■ Tabel 4.1 Toetsmatrijs Geschiedenis.

Thema	Domein	Aantal
1: Jagers en boeren 5 vragen	1.1 Tijdvak	5
	1.2 Historisch redeneren	0
2: Grieken en Romeinen 7 vragen	2.1 Tijdvak	6
	2.2 Historisch redeneren	1
3: Monniken en ridders 8 vragen	3.1 Tijdvak	7
	3.2 Historisch redeneren	1
4: Steden en staten 10 vragen	4.1 Tijdvak	8
	4.2 Historisch redeneren	2
5: Ontdekkers en hervormers 12 vragen	5.1 Tijdvak	9
	5.2 Historisch redeneren	3
6: Regenten en vorsten 14 vragen	6.1 Tijdvak	11
	6.2 Historisch redeneren	4
7: Pruiken en revoluties 17 vragen	7.1 Tijdvak	12
	7.2 Historisch redeneren	5
8: Burgers en stoommachines 20 vragen	8.1 Tijdvak	13
	8.2 Historisch redeneren	7
9: Wereldoorlogen 22 vragen	9.1 Tijdvak	14
	9.2 Historisch redeneren	8
10: Televisie en computers 24 vragen	10.1 Tijdvak	15
	10.2 Historisch redeneren	9
	Totaal	140

vakcommissie een toetsmatrijs ontwikkeld. De toetsmatrijs is de blauwdruk voor het samenstellen van een toets en beschrijft de domeinen en hun relatieve gewicht in de toets (zie ook ► hoofdstuk 1). De vakcommissie heeft bepaald dat de gehele kennisbasis landelijk wordt getoetst met uitzondering van de vakdidactiek. De vakdidactiek leent zich niet voor digitale toetsing, zo is de mening, de instellingen gebruiken hiervoor eigen methoden. ■ Tabel 4.1 laat de vereenvoudigde versie van de toetsmatrijs zien. Er is een vaste verdeling van de in de toets gestelde vragen over de verschillende domeinen. Iedere student krijgt dezelfde vragen per toetsafname, deze vragen worden per blok gerandomiseerd gepresenteerd. Er zijn tien blokken vragen in de toets, met 140 vragen in totaal. Ieder blok beslaat één tijdvak. Vragen over het kennisbasisdomein ‘Historisch redeneren’ (verzamelen, ordenen, verklaren, beeldvormen) worden gesteld binnen de context van een tijdvak.



▣ **Figuur 4.2** De betrokken partijen in het ontwikkelproces.

De toetsmatrijs vormt voor de docenten in de toetsredactie de basis voor het schrijven van de vragen. Iedere vraag wordt door een andere redacteur gecontroleerd en verbeterd waar nodig. De hoofdredacteur keurt de vraag goed en stuurt de vraag door naar de vakcommissie (zie ▣ figuur 4.2). Deze commissie kijkt naar de inhoudelijke relevantie van de vraag in relatie tot de kennisbasis. De vragen die de vakcommissie afkeurt, gaan terug naar de redactie voor bijstelling of verwijdering. Goedgekeurde vragen gaan door naar de kwaliteitscommissie. Zij beoordeelt de vraag op toetskundige criteria zoals relevantie, objectiviteit, efficiëntie en moeilijkheidsniveau. De gehele werkwijze wordt in Question Mark Perception, het gebruikte toetsservicesysteem, gefaciliteerd door een workflow. Zie het kader hierna voor een voorbeeldvraag uit de oefentoets.

Om via psychometrische gegevens over de vragen en toetsen een beeld te krijgen van de kwaliteit van de toetsvragen zie ook ► hoofdstuk 2), zijn tussen 2010 en 2013 de vragen voorgelegd aan testpanels bestaande uit derdejaarsstudenten.

Binnen dit project moet steeds voldoende binding worden gehouden met alle lerarenopleidingen Geschiedenis in Nederland. Ook al zijn binnen de redactie en de vakcommissie zes van de tien lerarenopleidingen Geschiedenis met docenten vertegenwoordigd, alle opleidingen dienen op de hoogte te zijn van de ontwikkelingen ten aanzien van de landelijke toets. Daarom is vanuit het project het landelijk vakoverleg (LVO) geïnitieerd, waarin docenten van alle opleidingen zitting hebben.



### Voorbeeld

*Item uit de itembank Geschiedenis (deze wordt gebruikt voor oefentoetsen)*

#### Vraagstelling

Inwoners van Berlijn kijken gespannen toe hoe het ene na het andere Amerikaanse sportvliegtuig landt op het vliegveld van de stad.

Wat was de achtergrond van deze situatie?

#### Antwoorden

- De wegen naar het westelijk stadsdeel waren door de Russen geblokkeerd, zodat bevoorrading alleen nog door de lucht kon plaatsvinden.
- Door de bouw van de Berlijnse Muur moesten de West-Berlijners van voedsel worden voorzien door Amerikaanse vliegtuigen.
- Na de val van de Muur werd de stad eindelijk toegankelijk voor vliegtuigen uit het westen.
- De Berlijners verwelkomden de komst van Amerikaanse bevrijders aan het eind van de Tweede Wereldoorlog.

## 4.3 Mogelijkheden

Het invoeren van een landelijke kennistoets biedt opleidingen de mogelijkheid een antwoord te formuleren op twijfels over de kwaliteit van de examinering zoals gesignaleerd door de Commissie Bruijn. Zoals uit bovenstaand voorbeeld van de landelijke kennistoets Geschiedenis blijkt, biedt een landelijke kennistoets de mogelijkheid het vreemde ogen principe toe te passen en tunnelvisie te voorkomen. Dit geldt zowel voor het vaststellen van de te toetsen leerstof (in het voorbeeld: de kennisbasis) als voor de toets zelf. Het gesprek tussen docenten van verschillende hogescholen over de leerstof werkt als een vorm van collegiale consultatie en zorgt ervoor dat zij worden uitgedaagd opvat-

tingen over hun vak kritisch onder ogen te zien. De betrokkenheid van externen uit het afnemend werkveld (in het voorbeeld: docenten vo/mbo) zorgt ervoor dat de praktische toepasbaarheid van de leerstof niet uit het oog wordt verloren. Vakexperts uit de wetenschap dragen in belangrijke mate bij aan de objectiviteit en het actualiteitsgehalte van de gemeenschappelijk afgesproken leerstof. Kortom: de leerstof wordt breed gedragen. Ook de toets zelf komt tot stand in een proces van *checks and balances*. Iedere vraag is door meerdere vakexperts en een toetsexpert gezien en becommentarieerd. Op deze wijze is dat bij een toets die op een enkele hogeschool wordt afgenomen (ook wel: instellingsgebonden toets genoemd) niet of nauwelijks mogelijk. Een landelijke toets als aanvulling op instellingsgebonden toetsen verhoogt de betrouwbaarheid van de examinering. Daarmee is het niet alleen een instrument om de kwaliteit van de examinering te verhogen, maar ook om het vertrouwen van de samenleving in de betreffende sector te doen toenemen.

#### 4.4 Voordelen

---

Landelijke kennistoetsen zijn al lange tijd gemeengoed in Nederland. De Cito-toets in groep acht van het basisonderwijs en het centraal examen als afsluiting van het voortgezet onderwijs, zijn daar de bekendste voorbeelden van. Een landelijke kennistoets in de vorm van een dergelijk 'klassiek landelijk eindexamen' heeft een aantal evidente voordelen. De belangrijkste daarvan is de beheersbaarheid van de kwaliteit van de examens, zowel in de zin van het niveau van de toets zelf, als wat betreft de handhaving van een adequaat niveau van normering (Commissie Externe Validering, 2011, p. 5). Door deze kenmerken is het klassieke landelijk eindexamen een aantrekkelijke vorm van toetsing voor velen, onder andere voor politici en beleidsmakers. Hun belang is immers dat zij burgers uit kunnen leggen dat het onderwijs in het hele land en door de jaren heen van hetzelfde hoge niveau is. Dit hoge niveau wordt geborgd door een examen dat voor alle deelnemers aan de betreffende vorm van onderwijs in het hele land op hetzelfde moment wordt afgenomen en aan dezelfde normering wordt onderworpen. Ook voor de leerlingen (en hun ouders) biedt het gelijke speelveld dat het klassieke landelijke eindexamen creëert, voordelen. Bij toelating tot vervolgonderwijs zijn ze niet alleen afhankelijk van een schoolgebonden advies, maar is er ook een objectieve toetsuitslag die de prestatie van de leerling vergelijkbaar maakt met die van andere leerlingen die zich aanmelden voor dat vervolgonderwijs. Ook voor landelijke kennistoetsen in het hoger onderwijs gelden de genoemde voordelen. Zoals hierboven bepleit, biedt de landelijke kennistoets Geschiedenis een antwoord op de zorg van de politiek over het niveau van de vakkennis van startende leraren. Studenten die de toets moeten maken, beoordelen het feit dat de toets landelijk wordt afgenomen als positief. Voor schoolleiders in het voortgezet onderwijs en mbo geeft de toets de garantie dat de docent die ze in dienst nemen, ongeacht de opleidingsplaats, over de in de kennisbasis beschreven vakkennis beschikt.

## 4.5 Nadelen

---

Tegenover de genoemde voordelen van het klassieke landelijke eindexamen staat een aantal nadelen. Deze nadelen vloeien grotendeels voort uit het feit dat de toets tot stand komt buiten de organisatie van de deelnemende scholen en daarmee zonder betrokkenheid van docenten uit die scholen. Docenten hebben geen invloed op de toetsdoelen. Omdat toetsing het curriculum stuurt, wordt de verantwoordelijkheid voor het curriculum de individuele docent en de vaksectie grotendeels ontnomen. Het klassieke landelijke eindexamen beperkt kortom de professionele ruimte van de docent. Dit kan bij hen leiden tot een gebrek aan draagvlak voor de betreffende toets. Ook leidt het ontstaan van lijstjes met goed en slecht scorende scholen ertoe bij dat de aandacht in het onderwijs zich versmalt tot de leerdoelen van het eindexamen. Speciale eindexamentrainingen moeten de student helpen door middel van een eindsprint het felbegeerde papiertje te halen en de school helpen om in de concurrentiestrijd het hoofd boven water te houden. Dit onderwijzen en leren voor de toets (*teaching to the test*) kan ten koste gaan van alle andere doelen die met goed onderwijs worden beoogd en past hoe dan ook niet bij moderne onderwijsconcepten (zie ook ► hoofdstuk 15 over voortgangstoetsen). Een ander gevolg van het feit dat de toets buiten de organisatie van de deelnemende scholen tot stand komt, is dat de toetsexpertise zich concentreert buiten de school in plaats van bij docenten in de school. Voor een sector als het hoger beroepsonderwijs, waar in den brede twijfels over de kwaliteit van examinering bestaan, is dit in het bijzonder een groot nadeel. Er zijn meer nadelen aan het klassieke landelijke eindexamen die niet zozeer andere vormen van onderwijs maar des te meer het hoger beroepsonderwijs betreffen. Deelname aan het hoger beroepsonderwijs leidt tot een beroepskwalificatie waarin niet alleen kennis maar ook vaardigheden en attitude (de beroepshouding) van eminent belang zijn. De beheersing van de juiste kennis is voor het eindniveau van een hbo-opleiding uitermate belangrijk maar niet alleen zaligmakend. Om bij het voorbeeld van dit hoofdstuk te blijven: een goed historicus is niet per se een goede docent Geschiedenis. Daarvoor is meer nodig. Zo zal de docent onder andere het vak op een enthousiasmerende manier voor het voetlicht moeten kunnen brengen en het groepsproces in de klas in goede banen moeten kunnen leiden. Ook zal een goede docent zich moeten kunnen verplaatsen in de leefwereld van jongeren. In dit kader wordt wel gesproken over *beroepscompetenties*, als een samenspel van de voor het beroep benodigde kennis, vaardigheden en houding. Voor het meten van dergelijke beroepscompetenties is de kennistoets alleen, hetzij landelijk of instellingsgebonden, een te beperkt middel. Aanvullende informatie over de prestatie van de student op een beroepscompetentie dient verkregen te worden uit bijvoorbeeld een praktijktoets of *performance assessment* waarin kennis, vaardigheden en beroepshouding integraal tot uiting komen (zie ► hoofdstuk 16).

Met de beroeps- of praktijkgerichtheid van het hoger beroepsonderwijs hangt ook het volgende nadeel van het klassieke landelijk eindexamen voor deze sector samen. In een hbo-opleiding wordt veel geleerd in en met de praktijk: het afnemende beroepenveld. Dit beroepenveld kent sterk regionale verschillen. Onderwijs verzorgen op een school in de Schilderswijk in Den Haag vraagt andere vaardigheden dan het verzorgen van onderwijs op een school in de Achterhoek. De gevraagde vaardigheden kunnen verschillen op peda-

gogisch terrein (bijvoorbeeld vanwege specifieke grootstedelijke versus plattelandsproblematiek), maar ook op vakinhoudelijk en -didactisch terrein (bijvoorbeeld aandacht voor de leefomgeving in relatie tot het vak Geschiedenis of Biologie). De Commissie Toekomstigbestendig Hoger Onderwijs Stelsel (2010) pleit voor meer differentiatie tussen instellingen voor hoger onderwijs. Regionale betrokkenheid is één van de dimensies waarop instellingen zich kunnen profileren. De commissie spreekt in deze over 'couleur locale'. Vanzelfsprekend kan een klassiek landelijk eindexamen geen recht doen aan dergelijke regionale verschillen.

## 4.6 Voorwaarden voor een succesvolle toepassing

---

In de voorgaande paragrafen is duidelijk geworden dat aan landelijke kennistoetsen niet alleen voordelen maar, zeker ook voor het hoger onderwijs, belangrijke nadelen kleven. In deze paragraaf zal worden belicht welke voorwaarden gelden voor een succesvolle toepassing van een landelijke kennistoets in het hoger onderwijs, waarbij de nadelen zo veel mogelijk worden voorkomen. Als voorbeeld daarbij zal weer de landelijke kennistoets Geschiedenis worden gebruikt. Bij de totstandkoming van deze toets is veel aandacht besteed aan de rol van de docent en aan externe legitimering. Een nog niet genoemd aspect is de rol van de examencommissies.

Het beperken van de professionele ruimte van de docent kan worden voorkomen door docenten te betrekken bij de toetsontwikkeling. Nog beter is de landelijke kennistoets geheel in eigen beheer te ontwikkelen en niet uit te besteden aan een derde partij zoals het geval is bij het centraal examen voortgezet onderwijs. Het is hoe dan ook van cruciaal belang dat docenten invloed hebben op belangrijke inhoudelijke besluiten, zoals het vaststellen van de leerdoelen en de toetsmatrijs. Bij de landelijke kennistoets Geschiedenis zijn docenten zowel bij het een als het ander betrokken. Docenten schrijven de kennisbasis en nemen deel in de toetsredactie of vakcommissie. Zij werken intensief samen met collega opleiders van andere instituten aan de ontwikkeling van de toetsmatrijs en toetsvragen en ontvangen *training-on-the-job*. Hun toetsexpertise neemt toe en daarmee de kwaliteit van de landelijke kennistoets. Positief neveneffect is dat de opgedane expertise doorwerkt in de wijze waarop docenten instellingsgebonden toetsen ontwikkelen. In die toetsen is er ruimte voor profilering. Op grond van bijvoorbeeld regionale kenmerken zetten docenten sommige onderdelen van de kennisbasis in instellingsgebonden toetsen zwaarder aan dan in de landelijke kennistoets. (In die zin is de landelijke kennisbasis, geborgd door de landelijke toets, het minimale kennisniveau dat de student beheerst.) Ook kunnen zij in instellingsgebonden toetsen een breed palet aan toetsvormen inzetten, zoals een toets met open vragen, essay, portfolio-assessment of performance assessment. Zowel de betrokkenheid van docenten bij besluitvorming over en ontwikkeling van de landelijke kennistoets als de ruimte voor profilering dragen bij aan het draagvlak in de opleidingen voor een landelijke kennistoets.

Afhankelijk van het doel dat wordt nagestreefd, is de betrokkenheid van docenten alleen niet voldoende om van een landelijke kennistoets een succes te maken. De betrokkenheid van externen, zowel op vakinhoudelijk als toetstechnisch gebied, draagt bij aan de

validiteit en objectiviteit van de toets, en daarmee aan het vertrouwen dat de toets genereert. In dit verband is het opmerkelijk dat de Commissie externe validering onder 'extern' bij externe validering verstaat dat een andere *hogeschool* mede de toets ontwikkelt, en nergens refereert aan de betrokkenheid van experts van buiten het hbo. Is het dan niet zo dat de sector als geheel nog steeds het verwijt kan worden gemaakt dat 'de slager zijn eigen vlees keurt'? Zoals in ► paragraaf 4.3 al is betoogd, heeft betrokkenheid van wetenschap en afnemend werkveld een toegevoegde waarde die moeilijk kan worden overschat en dus ten minste het nastreven waard is.

Met betrekking tot de examen- of afnamecondities, inzage en bezwaar en beroep, is het van belang een gelijk landelijk speelveld te creëren. Het is noodzakelijk examencommissies in staat te stellen hun wettelijke taken in het kader van de Wet op het hoger onderwijs en wetenschappelijk onderzoek uit te oefenen. Nalatigheid op dit vlak zou de inspanningen die getroost zijn om een kwalitatief goede toets te ontwikkelen, teniet kunnen doen. Een gelijke behandeling van studenten wordt bevorderd door examencommissies in het kader van de landelijke kennistoets afspraken te laten maken over de wijze waarop ze hun taak op de eigen hogeschool uitoefenen. Een concreet voorbeeld kan dit duidelijk maken. Bij de landelijke kennistoets Geschiedenis hebben opleidingen afgesproken dat ze dezelfde voorwaarden voor deelname, dezelfde studielast en dezelfde herkansingsregeling opnemen in de Onderwijs- en Examenregeling. De examencommissies van de betrokken opleidingen zien toe op de naleving hiervan. Zij zijn georganiseerd in een landelijk overleg waarin ze hun bevindingen delen. Vertegenwoordigers van dit landelijk overleg zien toe op de kwaliteit van de landelijke kennistoets, spelen een rol bij de vaststelling van de cesuur, en informeren de betrokken examencommissies. Op deze wijze kunnen examencommissies doen wat ze moeten doen: toezicht uitoefenen op de toetscondities 'in eigen huis', toezicht uitoefenen op de kwaliteit van de landelijke kennistoets en met vertrouwen de landelijk vastgestelde cesuur overnemen.

Tot slot is een voorwaarde voor een succesvolle toepassing van landelijke kennistoetsen dat betrokken docenten, externen en leden van examencommissies voldoende worden gefaciliteerd. Hoewel het buiten het kader van dit hoofdstuk valt welke modellen hiervoor in aanmerking komen, zal uit het voorbeeld van de landelijke kennistoets Geschiedenis duidelijk zijn dat landelijke kennistoetsen een flinke tijdsinvestering vraagt van alle betrokkenen. Dit aspect dient daarom tijdig onder de aandacht van het betrokken management te worden gebracht. Het ligt daarbij voor de hand dat een landelijke kennistoets die aan bovengenoemde voorwaarden voldoet, zo veel bijdraagt aan de betrouwbaarheid van de examinering van een opleiding dat deze toets een deel van de instellingsgebonden toetsing kan vervangen. De landelijke kennistoets is daarmee niet een extra toets, maar een van de toetsen van het toetsprogramma van een opleiding. En dan zal weleens kunnen blijken dat de gedane investering zich (meer dan) terugverdient.

# Kwaliteit van toetsprogramma's

*Liesbeth Baartman, Raymond Kloppenburg en Frans Prins*

## Inleiding

Wat is een toetsprogramma?

Waar komt het idee van toetsprogramma's vandaan?

Wat zijn de achterliggende gedachten van een toetsprogramma?

Welke probleem lost het op?

Hoe werkt het in praktijk?

Hoe bepaal je de kwaliteit van een toetsprogramma?

- 5.1      Wat is een toetsprogramma? – 52**
- 5.2      Waarom toetsprogramma's? – 53**
  - 5.2.1    Dekking van de opleidingskwalificaties – 53
  - 5.2.2    Meer betrouwbare uitspraken over studenten – 55
  - 5.2.3    Stimuleren en volgen van de ontwikkeling van studenten – 56
- 5.3      Ontwerp van een toetsprogramma – 57**
- 5.4      De kwaliteit van toetsprogramma's – 58**
- 5.5      Zelfevaluatie-instrument KIT – 61**
- 5.6      Ten slotte – 62**




Toetsen moeten van kwalitatief hoog niveau zijn. Beslissingen over slagen of zakken worden genomen op basis van toetsen en studenten sturen hun leerproces op basis van wat zij verwachten te moeten laten zien in een toets. In dit hoofdstuk wordt de kwaliteit van toetsing niet op het niveau van een afzonderlijke toets bepaald, maar op het niveau van een toetsprogramma. De redenen hiervoor zijn dat een juiste beslissing over het toekennen van het diploma alleen mogelijk is als deze is gebaseerd op een samenhangend geheel van meerdere toetsen en dat alleen in een kwalitatief goed toetsprogramma de ontwikkeling van de student kan worden gevolgd en bijgestuurd. Dit betekent dat de validiteit en betrouwbaarheid niet alleen op het niveau van een afzonderlijke toets een rol spelen, maar ook op programmaniveau. Ofwel, dat niet alleen afzonderlijke toetsen zo valide en betrouwbaar mogelijk moeten zijn, maar ook het totaal aan toetsen. Voor docenten betekent dit afstemming en samenwerking binnen het docententeam, want kwaliteit op programmaniveau vereist een overzicht over alle cursussen en toetsen van de opleiding. Docenten moeten weten hoe hun toetsen zijn ingepast in het totaal aan toetsen en moeten hun keuze voor bepaalde toetsvormen afstemmen in relatie tot de andere cursussen en toetsen in het totale programma.

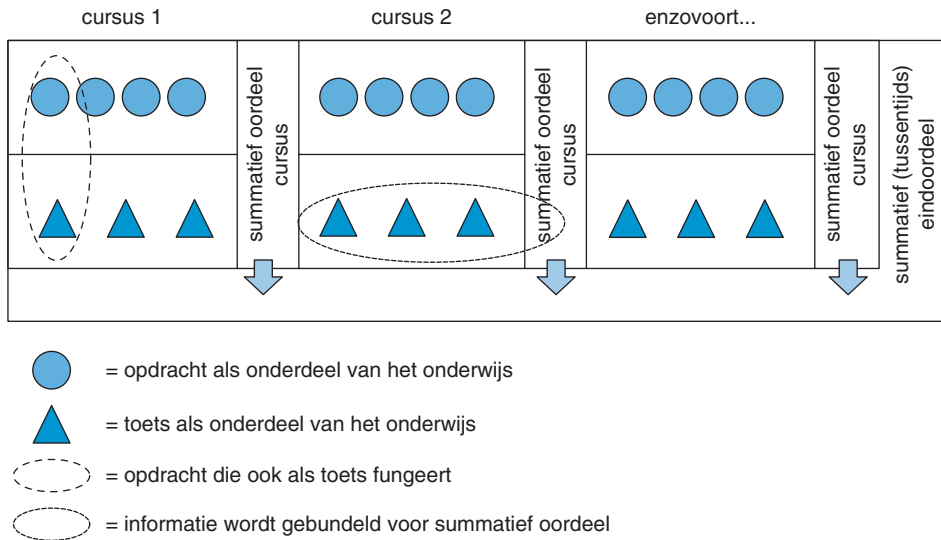
## 5.1 Wat is een toetsprogramma?

---

Een toetsprogramma is meer dan de optelsom van alle toetsen van een opleiding. Een toetsprogramma is ook meer dan een willekeurige mix van toetsvormen. Van een programma is pas sprake als de combinatie van toetsvormen bewust is samengesteld, passend bij de doelen, inhoud, opbouw en structuur van het curriculum. Vergelijk het met de combinatie van toetsvragen of items in een toets: een goede toets is meer dan een random set van toetsvragen, ook al is elke vraag op zich van goede kwaliteit; het gaat om een samenhangend geheel van het totaal. Wat betreft inhoud zijn de toetsen in een toetsprogramma complementair: ze vullen elkaar aan wat betreft de dekking van de opleidingskwalificaties. Wat betreft opbouw en structuur past een toetsprogramma bij het curriculum, bijvoorbeeld doordat er gedurende de studiejaren sprake is van een toenemend niveau, diepgang en complexiteit van toetsen die in overeenstemming is met de opbouw van leerlijnen in het curriculum. Of doordat de toetsen gaandeweg de opleiding meer integratief van karakter zijn en overeenstemmen met authentieke taaksituaties en eisen van de beroepspraktijk.

In een curriculum kunnen horizontaal en verticaal samenhangende toetsprogramma's worden onderscheiden. Voorbeelden van horizontaal samenhangende toetsprogramma's zijn: alle eerstejaarstoetsen die bepalend zijn voor het bindend studieadvies, een set van toetsen in het tweede jaar van de opleiding waarmee de stagewaardigheid van een student wordt bepaald, of de examenonderdelen in de afstudeerfase van de opleiding. Voorbeelden van verticaal samenhangende toetsprogramma's zijn alle toetsen gedurende alle jaren van de opleiding waarmee de kennisbasis van het beroep wordt beoordeeld (bijvoorbeeld met een voortgangstoets) of alle toetsen die gericht zijn op een cluster van opleidingskwalificaties (bijvoorbeeld met betrekking tot contact met de cliënt of onderzoek).

In het model van Van der Vleuten e.a. (2012) wordt getoond hoe de verschillende formatieve en summatieve toetsmomenten in een programma worden gebruikt om beslissingen te nemen over de studievoortgang en het leerproces van de studenten (zie  figuur 5.1).



■ **Figuur 5.1** Onderdeel van het model voor toetsprogramma's in een curriculum (Van der Vleuten e.a., 2012).

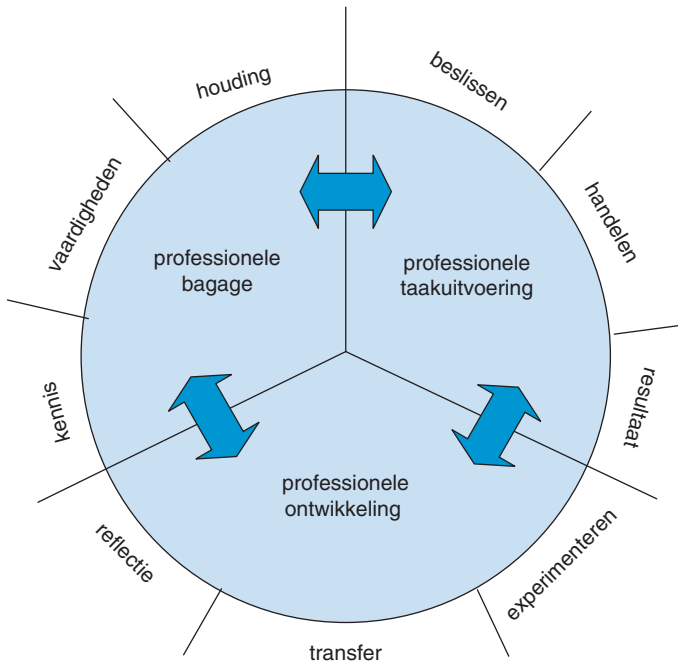
Formatieve toetsen zijn onderdeel van de dagelijkse lespraktijk (de driehoekjes in ■ figuur 5.1) en leveren betekenisvolle feedback op voor studenten. Deze toetsen zijn verweven met opdrachten (de bolletjes in ■ figuur 5.1): de studenten oefenen met de stof. Aan het eind van de cursus kunnen deze toetsen (en de opdrachten die ook als toets kunnen fungeren) summatief worden ingezet: er volgt een beslissing over het wel of niet behalen van de cursus. In ■ figuur 5.1 is dit weergegeven bij cursus 2: de informatie uit de toetsen (de driehoekjes) wordt samengevoegd om een summatief oordeel te vellen op cursusniveau. Naast beoordelingen op het niveau van een cursus, worden in het curriculum ook meer overkoepelende beslissingen genomen, zoals het bindend studieadvies, die bepalend zijn voor de studievoortgang. Dit zijn de summatieve (tussentijdse) eindoordelen in ■ figuur 5.1. Voor een summatief (tussentijds) eindoordeel wordt de informatie uit de toetsen van de verschillende cursussen samengenomen. Aan het eind van het hele toetsprogramma volgt de uiteindelijke beslissing: het wel of niet toekennen van het diploma aan de student.

## 5.2 Waarom toetsprogramma's?

Toetsprogramma's zorgen voor de dekking van de opleidingskwalificaties, voor het verhogen van een betrouwbare uitspraak over studenten en voor het stimuleren van de ontwikkeling van studenten.

### 5.2.1 Dekking van de opleidingskwalificaties

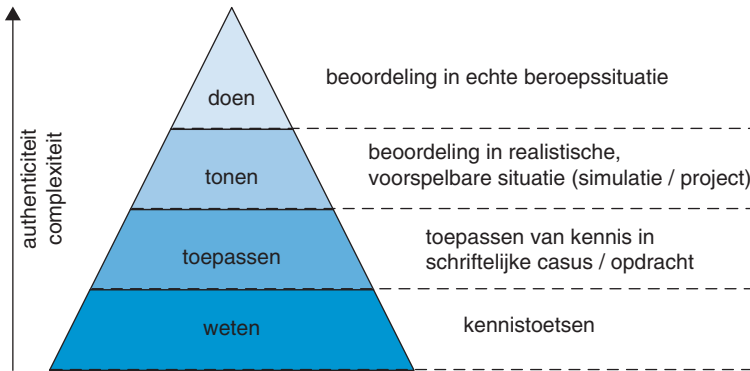
Studenten krijgen een diploma uitgereikt als duidelijk is dat ze de eindkwalificaties van de opleiding hebben behaald. Deze eindkwalificaties bestaan uit kennis-, vaardigheids- en



■ **Figuur 5.2** Dimensies van beroepsgerichte opleidingskwalificaties (Kloppenborg, 2011).

houdingscomponenten (de professionele bagage), het adequaat inzetten hiervan in professionele taaksituaties, en de persoonlijke professionele ontwikkeling (zie ■ figuur 5.2). De multidimensionele samenstelling van eindkwalificaties maakt duidelijk dat het onmogelijk is om alle dimensies op een adequate manier te toetsen met één toets. Er zijn meerdere en verschillende toetsen nodig om het geheel van dimensies van opleidingskwalificaties afzonderlijk en integratief te kunnen beoordelen (Baartman, Prins, Kirschner & Van der Vleuten, 2007). Een schriftelijke kennistoets is bijvoorbeeld een efficiënte manier om de kennisbasis van studenten te toetsen. Maar met een kennistoets alleen is het onmogelijk om te beoordelen of studenten hun kennis ook daadwerkelijk kunnen inzetten in een concrete beroepssituatie, daar afwegingen kunnen maken, de juiste beslissingen kunnen nemen en de juiste strategieën kunnen kiezen om het gewenste resultaat te bereiken (Roelofs, 2006).

Dit idee van het combineren van verschillende toetsvormen om een inhoudelijk dekkend toetsprogramma te bereiken, wordt onder meer weergegeven in de piramide van Miller (1990). Daarin wordt een onderscheid gemaakt tussen toetsen op het niveau van 'weten', 'toepassen', 'tonen' en 'doen' (zie ■ figuur 5.3). Het toetsen van opleidingskwalificaties die bestaan uit bovengenoemde dimensies, impliceert toetsvormen op zowel het niveau van 'weten' en 'toepassen' (de kennisbasis) als het niveau van 'tonen' en 'doen' (de professionele taakuitvoering). Dit betekent echter niet dat de piramide ook weergeeft wat opeenvolgend in verschillende studiejaar wordt getoetst, dus van 'weten' in het eerste jaar naar 'doen' in het laatste jaar. Het gaat er juist om dat deze aspecten goed in balans zijn: zowel over de jaren heen als per jaar. Wat de piramide van Miller laat zien, is dat je alleen dekkend kunt toetsen in een totaal toetsprogramma waarin de verschillende toetsvormen



■ **Figuur 5.3** piramide van Miller (1990).

complementair zijn en elk bijdragen aan de toetsing van verschillende dimensies van de opleidingskwalificaties. De validiteit van toetsing wordt dus niet bepaald door één of enkele toetsen, maar door het totale toetsprogramma van de opleiding. Validiteit is een kwaliteitskenmerk dat aansluit bij de dekking van de opleidingskwalificaties. Om de validiteit van het hele toetsprogramma te kunnen bepalen geldt de vraag: geven de toetsen samen de juiste en voldoende informatie om een gerechtvaardigd oordeel te kunnen vellen over de student?

## 5.2.2 Meer betrouwbare uitspraken over studenten

Onderzoek heeft aangetoond dat de professionele taakuitvoering in de beroepspraktijk taakspecifiek en contextspecifiek is: als een student één bepaalde beroepstaak goed uitvoert, heeft dit weinig voorspellende waarde voor een volgende beroepstaak (Ruiz-Primo & Shavelson, 1996; Van der Vleuten e.a., 2010). Voor de toetsing betekent dit dat het beoordelen van de professionele taakuitvoering ('tonen' en 'doen' in de piramide van Miller) op basis van één beroepstaak onbetrouwbaar is. Omdat het om een enkel meetmoment gaat, kan die beoordeling door de invloed van toevallige factoren vertekend zijn. Te denken valt aan de subjectiviteit van één beoordelaar (halo- of horn-effect; zie ook de bijlage achter in dit boek), afleidende omstandigheden, of bijvoorbeeld een slechte lichamelijke conditie van de student.

Een betrouwbare beoordeling of een student zich tot een aankomend professional ontwikkelt, vereist dus meerdere toetsen die in een zorgvuldig gekozen combinatie samen het toetsprogramma vormen. Door informatie uit meerdere 'meetmomenten' in het toetsprogramma samen te voegen, kan een meer betrouwbaar en generaliseerbaar oordeel worden geveld over de student. Deze meetmomenten kunnen op steeds andere aspecten van een eindkwalificatie zijn gericht, er kunnen verschillende soorten toetsvormen in wisselende contexten worden gebruikt en meerdere beoordelaars kunnen worden ingezet. Beoordelaars hebben verschillende perspectieven en daarmee verschillende informatie over de student. Door deze verschillen met elkaar te confronteren kan onderlinge afstemming plaatsvinden

om zo tot overeenstemming te komen over de uiteindelijke beoordeling. Dit wil niet zeggen dat betrouwbaarheid pas aanwezig is als de beoordelingen van verschillende beoordelaars volledig overeenkomen. Denk aan een stagebegeleider die de student in realistische praktijksituaties meemaakt en daardoor een ander beeld heeft van de student dan een docent die de student alleen op de opleiding meemaakt. Als deze beoordelaars in een overleg tot een oordeel komen, kan de informatie die zij beiden inbrengen bijdragen aan een completer beeld van de student. Deze rijkheid van informatie leidt uiteindelijk, in combinatie met andere toetsen, tot een meer betrouwbaar oordeel over de student. Dat wil zeggen: het oordeel is minder vatbaar voor meetfouten en heeft een hogere voorspellende waarde doordat het is gebaseerd op een grotere hoeveelheid verschillende informatie over de student.

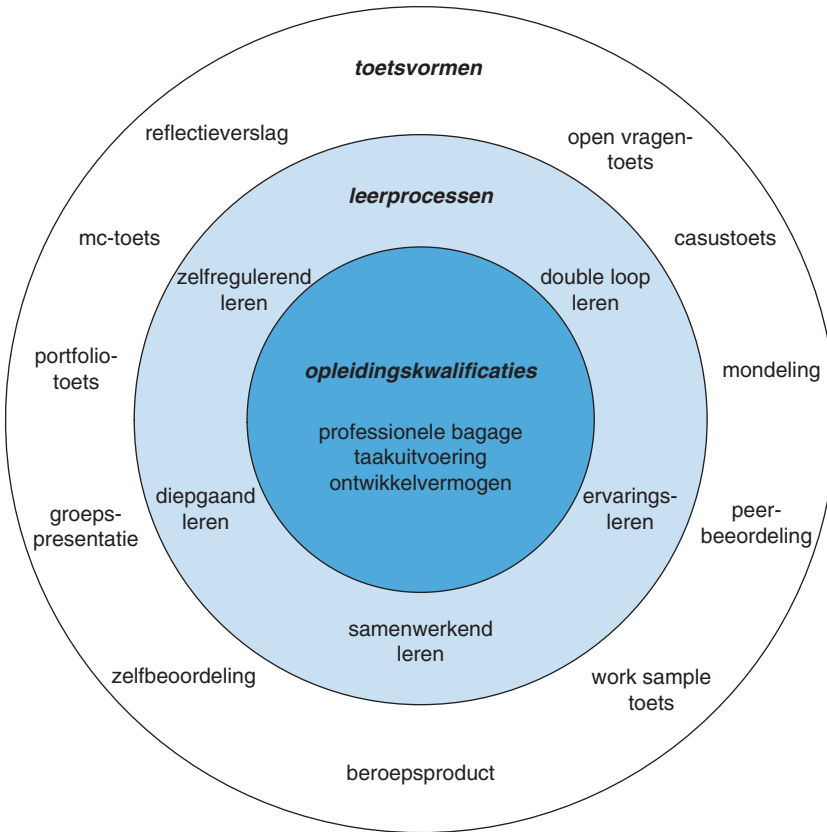
### 5.2.3 Stimuleren en volgen van de ontwikkeling van studenten

---

Binnen een curriculum zijn alle cursussen in een bepaalde volgorde en met een bepaalde opbouw georganiseerd en dragen allen bij aan de ontwikkeling van studenten. In het totale toetsprogramma worden de opleidingskwalificaties ook op verschillende niveaus en met een opbouw in complexiteit getoetst. Het denken in termen van een geheel toetsprogramma biedt voordelen als het gaat om het stimuleren en volgen van de ontwikkeling van studenten. Het leerproces van studenten wordt niet alleen beïnvloed door één afzonderlijke toets binnen een cursus, maar door het geheel aan toetsen binnen de opleiding. Toetsen zijn in het curriculum een hulpmiddel bij het stimuleren van wenselijke leerprocessen. Opleidingsgestuurd leren kan bijvoorbeeld gaandeweg de opleiding overgaan in studentgestuurd leren en samenwerkend leren. In een consistent toets- en onderwijsprogramma zijn de gekozen toetsvormen en inhoud een afspiegeling van deze leerprocessen (Gibbs & Simpson, 2004).

Ook biedt een toetsprogramma de mogelijkheid om het leerproces gedurende de hele opleiding – dus ook over cursussen heen – te sturen door middel van feedback en het opstellen van concrete leerdoelen. Doordat opleidingskwalificaties op meerdere niveaus worden getoetst in een toetsprogramma, kunnen studenten naar aanleiding van een toetsuitslag concrete doelen opstellen en zo aan hun ontwikkeling werken in de voorbereiding naar een volgend toetsmoment. De feedback die ze hebben gekregen bij een toets op een lager niveau wordt dan gebruikt voor een toets over dezelfde leerinhoud maar op een hoger niveau. Een toetsprogramma functioneert zo als een richtinggevend instrument om feedback te geven op de ontwikkeling van studenten door het hele curriculum heen.

Tot slot moeten in een toetsprogramma de formatieve en summatieve functies van toetsing in balans zijn. Studenten laten hun leerproces vooral leiden door wat zij verwachten te moeten laten zien in de toetsen met een summatieve functie, omdat deze toetsen meetellen. Een toetsprogramma moet beide functies vervullen: de toetsing moet zowel betekenisvolle feedback opleveren en gewenste leerprocessen stimuleren (formatief) als valide en betrouwbare beslissingen opleveren (summatief). In het totale toetsprogramma kunnen hierin afwegingen worden gemaakt: een deel van de toetsen is vooral efficiënt en levert minder feedback op, een (ander) deel wordt voornamelijk formatief ingezet en voorziet in rijke feedback voor de studenten.



■ **Figuur 5.4** Ontwerpcirkels van een toetsprogramma.

### 5.3 Ontwerp van een toetsprogramma

De opleidingskwalificaties zijn bepalend voor het ontwerp van een curriculum en daarmee samenhangend toetsprogramma. Het gaat dan zowel om de inhoud van de opleidingskwalificaties als om het stimuleren van de leerprocessen die daartoe leiden. Een veel gebruikte opzet is die van leerlijnen die zijn gericht op verschillende clusters van opleidingskwalificaties (bijvoorbeeld werken in het primaire proces, de organisatie, beroepsontwikkeling) of bepaalde specifieke leerprocessen (zoals het diepgaand bestuderen van kennis, het leren in de praktijk van werkervaringen, en het zelfregulerend leren). De wijze waarop het toetsprogramma is samengesteld, dat wil zeggen de verdeling van toetsinhouden, de opklimmende complexiteit, en de variatie van toetsvormen, is hiervan afhankelijk.

In **■** **figuur 5.4** wordt aan de hand van ontwerpcirkels zichtbaar gemaakt hoe de opleidingskwalificaties de inhoudelijke kern vormen bij het ontwerp van een curriculum en toetsprogramma. In de eerste cirkel om de kern staan de leerprocessen die leiden tot de ontwikkeling van deze opleidingskwalificaties. In de buitenste cirkel worden de toetsvormen gekozen die inhoudelijk en qua vorm in overeenstemming zijn met de binnenste cirkels.

## 5.4 De kwaliteit van toetsprogramma's

Zoals eerder aangegeven wordt de kwaliteit van toetsing niet alleen bepaald op het niveau van afzonderlijke toetsen, maar juist ook op het niveau van het toetsprogramma als geheel. De kwaliteit van toetsprogramma's kan in vier kwaliteitscategorieën worden verdeeld: validiteit, betrouwbaarheid, functies en condities. Onder elk van deze categorieën ligt een aantal kwaliteitscriteria ten grondslag (zie [▣](#) tabel 5.1). Deze kwaliteitscategorieën en -criteria zijn gebaseerd op een vergelijking en samenvoeging van de kwaliteitsinstrumenten uit twee onderzoeken: het onderzoek naar Competentie Assessment Programma's (CAP) van Baartman (2008) en het onderzoek naar kwaliteitscriteria voor competentiegerichte toetsing van Kloppenburg (2011).

### ▀ Validiteit

Validiteit van toetsing – meet je inderdaad wat je beoogt te meten – is misschien wel de meest omvattende kwaliteitscategorie waarover in de wetenschap en de praktijk stevige discussies zijn gevoerd die hebben geleid tot verschillende, elkaar opvolgende definities en toepassingen (Kane, 2006). Er zijn twee visies op validiteit die van belang zijn voor het (hoger beroeps)onderwijs. Ten eerste de meer smalle of toegespitste visie waarbij de essentie van validiteit van toetsing weergegeven wordt als de overeenkomst tussen de opleidingskwalificaties zoals deze zijn bedoeld en zoals deze in de toetsing worden bepaald (Borsboom, Mellenbergh & Van Heerden, 2004; De Groot, 1981). Ten tweede een meer brede visie waarin validiteit ruim wordt opgevat en niet alleen betrekking heeft op de beoordeling van de beroepskwalificaties, maar ook op de consequenties van de toetsing (consequentiële validiteit) (Messick, 1995). Bij de validiteit van toetsprogramma's gaat het om de smallere definitie van validiteit. De validiteit van een totaal toetsprogramma wordt bepaald aan de hand van vier kwaliteitscriteria: operationalisatie, dekking, complexiteit en vorm (zie [▣](#) tabel 5.1). Het gaat er steeds om of het totale toetsprogramma een adequate representatie is van de inhoud van de opleidingskwalificaties en of de gebruikte toetsvormen passen bij deze inhoud.

### ▀ Betrouwbaarheid

Betrouwbaarheid heeft betrekking op de accuraatheid of herhaalbaarheid van toetsuitslagen en op de factoren die bias kunnen introduceren, zoals beoordelaars of de beperkte generaliseerbaarheid over verschillende beroepstaken. Over de betrouwbaarheid van toetsing van beroepskwalificaties is de afgelopen decennia veel geschreven en gediscussieerd. Soms gaan deze discussies over de vraag of beroepskwalificaties überhaupt wel betrouwbaar kunnen worden beoordeeld (Dierick, Watering & Muijtjens, 2002; Luken, 2004) of over de vraag hoe betrouwbaar een enkele toets kan zijn (Van der Vleuten e.a., 2010).

Bij toetsing van beroepsfunctioneren en toetsen op het 'doen' niveau van Miller (1990) is altijd een menselijk oordeel nodig, dat wil zeggen het oordeel van iemand die bijvoorbeeld observeert in een praktijksituatie/video of een student bevrageet in een criteriumgericht interview of portfoliogesprek. Zoals al eerder besproken kan een toetsprogramma de betrouwbaarheid verhogen door het inzetten van meerdere beoordelaars en het zorgen

■ **Tabel 5.1** Kwaliteitscategorieën en kwaliteitscriteria voor toetsprogramma's.

Kwaliteitscategorie	Kwaliteitscriteria
validiteit	<p><i>Operationalisatie</i></p> <p>Het toetsprogramma is op een duidelijke en juiste wijze afgeleid van de opleidingskwalificaties. De opleidingskwalificaties zijn uitgewerkt in specifiekere beoogde leeruitkomsten per cursus en in beoordelingscriteria of correctiemodellen van de toetsen behorende bij deze cursussen.</p>
	<p><i>Dekking</i></p> <p>De inhoud van het toetsprogramma is een goede afspiegeling van de inhoud van de opleidingskwalificaties en aspecten daarvan. De kennisbasis, beroepsvaardigheden en de uitvoering van complexe beroepstaken worden voldoende getoetst.</p>
	<p><i>Complexiteit</i></p> <p>Het toetsprogramma is van het juiste niveau, passend bij de fase van de opleiding, en kent een opbouw in complexiteit.</p>
	<p><i>Vorm</i></p> <p>De combinatie van toetsvormen in het gehele toetsprogramma past bij de te beoordelen inhoud.</p>
betrouwbaarheid	<p><i>Vergelijkbaarheid</i></p> <p>Waar mogelijk zijn inhoud en omstandigheden van de toetsing vastgelegd (beoordelingscriteria, correctiemodellen en normering) en worden verschillen tussen beoordelaars geminimaliseerd door te zorgen voor gedeeld begrip.</p>
	<p><i>Triangulatie</i></p> <p>Het toetsprogramma levert voldoende rijkheid aan informatie om een betrouwbaar (integraal) oordeel te kunnen vellen. Beslissingen zijn niet gebaseerd op het (subjectieve) oordeel van één beoordelaar of op het (toevallige) oordeel in één situatie of beroepstaak, maar informatie wordt getrianguleerd over beoordelaars en toetsmomenten heen.</p>
functie	<p><i>Selectie</i></p> <p>Op basis van het toetsprogramma kunnen goede en zwakke studenten worden onderscheiden en geselecteerd (t.b.v. geschiktheid voor vervolg studie, uitoefening van het beroep). Deze selectie moet tijdig in de opleiding plaatsvinden, zodat studenten gemakkelijker kunnen overstappen naar een andere studie, of kunnen stoppen met de studie.</p>
	<p><i>Leereffect</i></p> <p>Het toetsprogramma heeft een wenselijk effect op het leerproces van de studenten. De verschillende toetsen samen stimuleren de gewenste leerprocessen en leveren rijke feedback op.</p>
	<p><i>Onderwijseffect</i></p> <p>Het toetsprogramma geeft rijke informatie over het ermee samenhangende onderwijs, zodat docenten het gegeven onderwijs kunnen evalueren en verbeteren.</p>
conditie	<p><i>Transparantie</i></p> <p>Studenten krijgen duidelijke informatie over alle toetsen, criteria, normering en uitslagen van de toetsing, zodat zij weten wat er van hen wordt verwacht.</p>



■ Tabel 5.1 Vervolg

Kwaliteitscategorie	Kwaliteitscriteria
	<p><i>Oordeelkundigheid</i> De beoordelaars beschikken over voldoende kennis en vaardigheden over de toetsinhoud, de manier van toetsen en het geven van feedback. Zij krijgen hier bijvoorbeeld training in.</p>
	<p><i>Organisatie</i> Het toetsprogramma is efficiënt en effectief georganiseerd en beoordelaars hebben voldoende tijd om tot een betrouwbaar oordeel te komen. Voorbeelden zijn: toetsroosters zijn op tijd gereed, surveillanten zijn geregeld, toetsuitslagen zijn op tijd bekend, feedbackbijeenkomsten worden georganiseerd.</p>

voor een grote steekproef uit het totale beroepsdomein. Hierdoor is het oordeel minder vatbaar voor meetfouten en kan het een hogere voorspellende waarde hebben.

Hiernaast kan de betrouwbaarheid worden verhoogd door een zekere mate van vergelijkbaarheid tussen toetsen in het toetsprogramma. Dat wil niet zeggen dat standaardisatie wordt nagestreefd op alle aspecten van toetsing. Dit is ook onmogelijk als het een beoordeling in de authentieke beroepspraktijk betreft, waar het juist gaat om het beroepsfunctioneren in complexe en onvoorspelbare situaties. Wel kunnen beoordelingscriteria en normering eenduidig worden vastgelegd en kunnen beoordelaars worden getraind, zodat alle betrokkenen de criteria zo veel mogelijk op dezelfde manier interpreteren. Het gaat hier om het bereiken van een zekere mate van een gedeeld begrip door bijvoorbeeld gezamenlijk te beoordelen, te vergelijken en te bespreken waardoor verschillen in interpretatie worden veroorzaakt.

#### ■ Functies

Toetsprogramma's moeten zo worden samengesteld dat het totale toetsprogramma voldoet aan de doelen die ermee worden nagestreefd (Van der Vleuten e.a., 2012). Summatief gaat het om het vaststellen of studenten als startbekwame professionals kunnen toetreden tot de arbeidsmarkt, en formatief om het stimuleren van de gewenste leerprocessen bij studenten. Wanneer studenten in de beginfase van de opleiding matige of slechte toetsuitslagen laten zien, biedt dit in een vroeg stadium de mogelijkheid om eventueel de keuze voor de opleiding te heroverwegen. Wat betreft het formatieve doel van toetsing is al eerder vermeld dat toetsen grote invloed hebben op het leerproces van studenten (Black & Wiliam, 1998; Gibbs & Simpson, 2004) en informatief zijn voor docenten. Analyses van toetsresultaten geven belangrijke informatie over de kwaliteit van het gegeven onderwijs, bijvoorbeeld over onderdelen van het onderwijs die meer aandacht behoeven.

#### ■ Condities

De kwaliteit van toetsprogramma's is tot slot in belangrijke mate afhankelijk van een aantal randvoorwaarden. Deze condities van toetsing zijn uitdrukkelijk ook opgenomen als kwaliteitscategorie. Immers, een toetsprogramma kan op papier nog zo goed in elkaar

zitten, maar de kwaliteit van toetsing wordt uiteindelijk bepaald door het gebruik ervan in de praktijk (Baartman, Prins e.a., 2007). Vergelijk het met het onderscheid dat Van den Akker (1999) maakt tussen het curriculum zoals het is beoogd, zoals het is uitgevoerd in de praktijk, en zoals blijkt uit de leerresultaten.

Ten eerste dienen de studenten tijdig informatie te krijgen over de toetsing, zodat zij weten wat er van hen wordt verwacht. Belangrijk is dat docenten deze informatie ook daadwerkelijk gebruiken in hun lessen, zodat beoordelingscriteria – gekoppeld aan beoogde leeropbrengsten van de cursus en de uiteindelijke opleidingskwalificaties (het criterium operationalisatie) – betekenis krijgen in het onderwijs. Ten tweede zijn goede docenten niet als vanzelfsprekend goede beoordelaars. Zoals eerder vermeld vereist beoordeling van beroepsfunctioneren het vertrouwen op een menselijk oordeel (van der Vleuten e.a., 2010). Een goed gebruik van een toetsprogramma vereist ervaring met de gebruikte toetsvormen, oefening in het gebruik van de beoordelingscriteria en ook het vermogen om goede feedback te geven aan de student. Oordeelkundigheid betekent ook dat docenten als professionals op het gebied van toetsing kunnen worden beschouwd. Beoordelaars moeten worden geschoold en gefaciliteerd om een valide en betrouwbaar oordeel te kunnen vellen. Tot slot de toetsorganisatie: de voorbereiding, uitvoering en administratie van toetsing. De toetsorganisatie is merkbaar op het meest concrete niveau en is in sterke mate van invloed op de tevredenheid over het verloop van toetsing van studenten en beoordelaars.

## 5.5 Zelfevaluatie-instrument KIT

---

Voor het evalueren van de kwaliteit van een toetsprogramma is door de auteurs het zelfevaluatie-instrument KIT (Kwaliteitsinstrument Toetsprogramma) ontwikkeld. Het instrument is een digitale tool waarmee opleidingsteams of toetscommissies een analyse kunnen maken van de kwaliteit van hun toetsprogramma. Het doel van deze analyse is op de eerste plaats formatief, dat wil zeggen gericht op de verbetering van de kwaliteit van het programma. De in ► paragraaf 5.4 beschreven kwaliteitscriteria zijn in het instrument voorzien van indicatoren. Bij elke indicator wordt aan de hand van een score aangegeven in hoeverre dit van toepassing is op het toetsprogramma en of dit ook voldoende is. Daarnaast wordt steeds gevraagd naar argumenten en bewijzen die aangeven waarop de score is gebaseerd (zie ■ figuur 5.5 voor een screendump). Voorbeelden hiervan zijn: informatie uit studiehandleidingen, evaluatiegegevens van studenten, beoordelingen van externen en toetsmatrijzen.

Het digitale KIT is onderdeel van een uitgebreidere procedure van zelfevaluatie. Voorafgaand aan het invullen wordt eerst het toetsprogramma afgebakend: hoe ziet het totale toetsprogramma eruit, welke toetsen zitten hierin, gaat om het een programma van een bepaald studiejaar (horizontaal) of van een leerlijn (verticaal)? Vervolgens vult een groep docenten, teamleiders, leden van de toetscommissie, en/of studenten het digitale KIT in. Naar analogie van het kwaliteitscriterium triangulatie gebeurt dit idealiter door meerdere mensen die (soms) verschillende informatie hebben over het toetsprogramma. Zo kunnen docenten bijvoorbeeld informatie inbrengen over de praktische afname van de toetsen, studenten over het effect dat de toetsen hebben op hun leerproces en toetscommissies over

**Kwaliteitsinstrument Toetsprogramma**

Start | Instructie | Overzichten | Exporteren

**Operationalisatie**

Het toetsprogramma is op een duidelijke en juiste wijze afgeleid van de opleidingskwalificaties.

Indicatoren	In welke mate is dit van toepassing op uw toetsprogramma?	Vindt u dit voldoende?	Dit blijkt uit ... Geef een voorbeeld / bewijs
De opleidingskwalificaties zijn helder omschreven, uitgewerkt en afgebakend.	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7 <input type="radio"/> 8 <input type="radio"/> 9 <input type="radio"/> 10 <input type="checkbox"/> onbekend	Ja <input type="radio"/> Nee <input checked="" type="radio"/>	Test: We gebruiken de dublin-descriptoren en op basis daarvan een uitwerking gemaakt van de kerntaken van het beroep en ook koppeling
De opleidingskwalificaties zijn duidelijk uitgewerkt in leerdoelen per cursus.	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7 <input type="radio"/> 8 <input type="radio"/> 9 <input type="radio"/> 10 <input type="checkbox"/> onbekend	Ja <input type="radio"/> Nee <input checked="" type="radio"/>	De kerntaken zijn de basis voor het curriculum. We moeten nog een duidelijke koppeling maken van welke k/v daar nu precies in
De leerdoelen per cursus zijn duidelijk uitgewerkt in beoordelingscriteria van de toetsen.	<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/> 6 <input type="radio"/> 7 <input type="radio"/> 8 <input type="radio"/> 9 <input type="radio"/> 10 <input type="checkbox"/> onbekend	Ja <input type="radio"/> Nee <input checked="" type="radio"/>	Dat kan nog niet, want vorige stap nog niet gedaan.

bewaar

■ **Figuur 5.5** Screenshot van het Kwaliteitsinstrument Toetsprogramma (KIT).

toetsmatrijzen en competentiematrices. Tot slot worden de uitkomsten van de invulling van het KIT gezamenlijk besproken. Het doel ervan is de uitkomsten gezamenlijk te interpreteren en hieraan conclusies te verbinden voor verbetering van het toetsprogramma. Voorbeelden van verbetermaatregelen zijn het verhogen van het leereffect van toetsen door studenten beter te informeren over inhoud en betekenis van de toetsuitslagen, het verhogen van de betrouwbaarheid door meerdere beoordelaars in het toetsprogramma te betrekken of het verhogen van de validiteit door ontbrekende inhoud van de opleidingskwalificaties op te nemen.

## 5.6 Ten slotte

Van een toetsprogramma is pas sprake als de combinatie van toetsvormen bewust is samengesteld, passend bij de doelen, inhoud, opbouw en structuur van het curriculum. Hiervoor is nodig dat alle betrokkenen binnen een opleiding zicht hebben op het hele programma en het totaal aan toetsen en toetsvormen. Als docenten hun keuze voor bepaalde toetsvormen afstemmen met docenten van andere cursussen en toetsen van het totale programma, ontstaat een samenhangend en consistent geheel van toetsen in een opleiding. Dit vergt afstemming en overleg met elkaar. Dat kan door het uitvoeren van de volgende activiteiten:

- Maak een overzicht van alle toetsen van alle cursussen van de opleiding.
- Bekijk als team of de inhoud van alle opleidingskwalificaties hierin terugkomen.
- Plaats de toetsen op de piramide van Miller: op welke laag zitten de toetsen van de opleiding.

Ook voor het bereiken van gedeeld begrip is een heel team nodig. Denk aan het inzetten van tweede beoordelaars bij het bespreken van eindwerkstukken van studenten om

gedeeld begrip te bereiken over de beoordelingscriteria en normering. Voor de formatieve functie van een toetsprogramma kunnen studentevaluaties gerichter worden ingezet. Gesprekken met studenten over hun leerproces en de invloed van verschillende toetsen hierop kunnen hier meer zicht op geven. Tot slot leidt het gezamenlijk vaststellen en evalueren van een toetsprogramma (zoals met het KIT) tot afstemming tussen docenten van verschillende cursussen en tot een gedeelde visie over het doel van de opleiding en de verschillende cursussen hierbinnen.

# Toetsen zonder plagiëren

## Twee detectiemethoden nader uitgewerkt

*Henk van Berkel*

### Inleiding

Wat is plagiaat?

Hoe kun je plagiaat voorkomen?

Welke methoden zijn er om plagiaat op te sporen?

Welke softwareprogramma's kun je hanteren?

Wanneer besluit je dat er sprake is van plagiaat?

- 6.1 Inleiding – 66
- 6.2 Voorkomen is beter dan genezen – 67
- 6.3 Soms zijn studenten onwetend – 67
- 6.4 Programma's voor het opsporen van plagiaat – 68
- 6.5 SafeAssign (Blackboard) – 69
- 6.6 WCopyfind – 70
- 6.7 Besluiten tot het oordeel plagiaat – 71

Het schrijven van een paper is vooral in latere studiejaar een goede manier om een studieonderdeel af te sluiten. Van studenten wordt verwacht dat zij de kennis die ze hebben vergaard op een systematische manier op papier kunnen zetten. Dat is voor velen geen gemakkelijke opdracht. Het schrijven zelf is vaak een moeizaam proces. Studenten kunnen in de verleiding komen 'leentjebuur' te spelen bij anderen. Ze zoeken op het internet, of schrijven stukken over van medestudenten. Dat kan leiden tot plagiaat. Het is moeilijk een scherpe scheidslijn te trekken tussen correct citeren en plagiëren. Plagiaat wordt hier gedefinieerd als een situatie waarin de studenten ten onrechte de indruk wekken dat een tekst(fragment) van henzelf afkomstig is.

## 6.1 Inleiding

---

Het schrijven van een paper wordt zowel door studenten als docenten gezien als een welkome afwisseling op de gebruikelijke schriftelijke toetsen. Een paper vereist van studenten verschillende vaardigheden. Studenten moeten zich taalkundig correct uitdrukken, zij moeten op een systematische en bondige manier hun kennis en gedachten op papier zetten en zij moeten de hoofd- en bijzaken in de gelezen literatuur kunnen scheiden. Dit alles is niet eenvoudig. Toch geven studenten aan dat het schrijven van papers in de meeste gevallen te prefereren is boven het afleggen van een schriftelijke toets. Dat is ook wel voorstelbaar. Een toets is een momentopname. In de twee of drie uur die een toets duurt, moeten de studenten 'pieken'. Aan een paper kunnen zij niet alleen langer en in hun eigen tijd werken, maar het schrijven van een paper is ook niet plaatsgebonden: dat kan thuis, of in een andere, niet door de opleiding gecontroleerde situatie. Dat laatste maakt een paper ontvankelijk voor het plegen van plagiaat. Plagiaat is sinds jaren een *hot topic*. Toenmalig staatssecretaris Rutte stelde in 2005 dat plagiaatplegers in het hoger onderwijs moesten worden verwijderd uit de opleiding. Zover is het niet gekomen. Maar het geeft wel aan hoe over de ernst ervan gedacht wordt.

Niet alles wat een student overschrijft van anderen is plagiaat. Als bijvoorbeeld een tekstfragment wordt overgenomen en de student zet daar een correcte bronvermelding bij, is er geen sprake van plagiaat in de hier gehanteerde definitie. Ook niet als het citaat bijvoorbeeld drie volle bladzijden beslaat. Dat de student in dat geval waarschijnlijk zondigt tegen een andere regel, bijvoorbeeld tegen de in de opleiding geldende regels ten aanzien van correct citeren, is in dit hoofdstuk niet aan de orde.

Apart aandacht verdient de situatie waarin studenten teksten van zichzelf overnemen, en niet van anderen. Deze situatie doet zich bijvoorbeeld voor wanneer studenten in een paper in het tweede studiejaar tekstfragmenten overnemen van een paper dat ze in het eerste jaar hebben geschreven, zonder dat te vermelden. Sommige frauderichtlijnen vermelden dat zoiets ook plagiaat is, omdat het de indruk wekt dat het tekstfragment specifiek is geschreven voor de paper in het tweede jaar terwijl dit niet het geval is. 'Het is toch van mijzelf afkomstig?', zo stellen studenten dan. Hier is duidelijk sprake van een grensgebied tussen plagiëren en niet-correct citeren (zie verder).

## 6.2 Voorkomen is beter dan genezen

---

Het zou natuurlijk beter zijn wanneer studenten niet in de verleiding komen om teksten van anderen over te nemen en net te doen alsof ze van henzelf zijn. Met andere woorden, is het te voorkomen dat studenten plagiaat plegen?

Helemaal voorkomen zal niet lukken. Wel is een aantal maatregelen te nemen die studenten minder in verleiding zullen brengen, zoals:

- Verstrek een duidelijke en specifieke opdracht, bijvoorbeeld: schrijf een individueel paper waarvan de inhoud moet zijn te herleiden naar de behandelde leerstof en waarvan een papieren én een elektronische versie moet worden ingeleverd.
- Wanneer studenten mogen kiezen uit verschillende opdrachten, zorg dan dat de keuzevrijheid beperkt is en dat de opdracht is te maken met behulp van de leerstof.
- Eis dat de paper specifieke onderdelen bevat, bijvoorbeeld: je moet van dat en dat artikel gebruikmaken, of: maak uitsluitend gebruik van bronnen van minder dan een jaar oud.
- Raak als docent betrokken bij het schrijfproces, laat bijvoorbeeld hoofdstuk voor hoofdstuk inleveren (en op het eind het geheel).
- Laat studenten hun papers mondeling presenteren en betrek deze presentatie bij de vaststelling van het toetscijfer.
- Laat een geannoteerde literatuurlijst toevoegen.
- Eis dat de paper wordt afgesloten met een soort reflectieverslag, met daarin onderwerpen als wat geleerd is, welke problemen er waren en hoe ze zijn opgelost, hoe de bronnen zijn gevonden. (Harris, 2004).

Deze maatregelen komen erop neer dat docenten zodanig specifieke opdrachten verstrekken aan studenten dat leentjebuurspelen niet kan volstaan.

## 6.3 Soms zijn studenten onwetend

---

Hoewel er ongetwijfeld kwaadwillende studenten zijn die willens en wetens van anderen overschrijven, is het beslist niet uitgesloten dat studenten vanuit een bepaalde onwetendheid handelen en op die manier plagiaat plegen. In het voortgezet onderwijs hebben studenten geleerd gebruik te maken van het internet bij het schrijven van werkstukken. Daarmee gaan ze in het hoger onderwijs door zonder te beseffen dat daar andere regels gelden. Bovendien is het internet een vrij toegankelijk medium. Daaraan verbinden ze de conclusie dat het publiek eigendom is, en dat zij het dús mogen gebruiken. Die redenering klopt ook wel, maar studenten verzuimen dan zich te realiseren dat ze met hun paper de indruk wekken dat ze de tekst zelf hebben geschreven. Met andere woorden, studenten weten niet goed wat plagiaat is. Ze weten niet wat het verschil is tussen correct citeren en onrechtmatig overnemen. Hoewel de regels daarvoor niet eensluidend zijn, is de grote lijn wel uit te leggen. Overigens verschillen de specifieke regels hiervoor per discipline. Voor een studie Rechten zijn er andere regels dan voor een studie Psychologie. Ook bij het geparafraseerd overnemen van een tekst hoort een bronvermelding. In veel opleidingen

krijgen studenten aan het begin van de opleiding een bibliotheekinstructie. Daar hoort deze scholing in thuis.

Waar het om gaat is dat studenten duidelijk wordt uitgelegd wat acceptabel is en wat niet. Een docent zal ook aan studenten moeten uitleggen dat verantwoord gebruik van bronnen iets toevoegt aan de kwaliteit van een paper. Het behoort tot de leerschool.

## 6.4 Programma's voor het opsporen van plagiaat

Het aantal softwareprogramma's om langs elektronische weg een paper op plagiaat te controleren, is groot en nog steeds groeiende. De internetsite ► <http://tlt.its.psu.edu/plagiarism/links/> bevat talloze rubrieken met links naar instanties die plagiaat kunnen opsporen, naar publicaties waarin softwareprogramma's worden vergeleken, naar artikelen over de mate waarin plagiaat voorkomt, naar manieren om plagiaat te voorkomen en naar artikelen over de redenen waarom studenten plagiaat plegen.

In essentie komen alle softwareprogramma's neer op eenzelfde systematiek: steeds worden tekstfragmenten uit de paper vergeleken met tekst uit andere documenten. Een tekstfragment kan bijvoorbeeld een zin zijn, of tien achtereenvolgende woorden. Hoe de programma's dat precies doen, wat de achterliggende (technische) mechanismen zijn, is niet altijd duidelijk. Uit commerciële overwegingen geven de meeste ondernemingen dit geheim niet prijs. Uit vergelijkend onderzoek blijkt echter dat ze wat de uitkomsten betreft niet voor elkaar onderdoen. Verschillen hebben te maken met de prijs (meestal betaal je als opleiding per ingediende paper en is het softwareprogramma zelf niet te koop; soms betaal je naar rato van het aantal studenten in de opleiding), snelheid, lay-out, leesbaarheid, wijze van presenteren van overeenkomsten, gebruiksvriendelijkheid (invoer én uitvoer) en technische ondersteuning.

### ■ Twee soorten softwareprogramma's

Er is een duidelijke tweedeling in de programma's die plagiaat opsporen:

- programma's die het internet afspeuren om te zoeken naar vergelijkbare tekstfragmenten;
- programma's die papers van een groep studenten die over een bepaald onderwerp hebben geschreven, met elkaar vergelijken.

Naar schatting zijn op het internet acht miljard documenten naspeurbaar; programma's die zoeken op het internet zijn daardoor onmiskenbaar in het voordeel. Er zijn echter ook toetsituaties waarin het niet zo noodzakelijk is papers te vergelijken met het hele internet. Bijvoorbeeld wanneer een groep studenten een individueel verslag moet schrijven over een experiment dat iedereen heeft uitgevoerd, of over het practicum dat ze allemaal hebben gevolgd. Dat zijn unieke situaties die moeten worden beschreven en waarvoor het internet geen hulp zal kunnen bieden. In deze situaties zal het echter wel voor de hand liggen om de papers van studenten *met elkaar* te vergelijken. Ze beschrijven immers situaties die gemeenschappelijk zijn geweest. Het is dan verleidelijk teksten van medestudenten over te nemen.



Zoals gezegd, het opsporen van plagiaat is een hot topic en het is daarom voortdurend in beweging. Daardoor is het in een boek zoals dit niet zinvol om de laatste stand van zaken weer te geven. Het is immers praktisch zeker dat de situatie over zes maanden is veranderd. Daarom wordt in dit hoofdstuk gekozen voor de uitleg van twee programma's waarvan mag worden verwacht dat ze redelijk stand zullen houden te midden van alle ontwikkelingen. Van elk van de twee soorten programma's wordt er één beschreven: *SafeAssign*, voor het vergelijken van papers met het internet, en *WCOPYfind* voor het vergelijken van papers met elkaar.

## 6.5 SafeAssign (Blackboard)

Het eerste programma is *SafeAssign*. Dat is een *buildingblock* binnen Blackboard. Blackboard is een elektronische leeromgeving (ELO) die in Nederland op veel instellingen voor hoger onderwijs wordt gehanteerd. SafeAssign is vooral geschikt om papers met internetdocumenten te vergelijken, maar het kan ook papers met elkaar vergelijken. Als dat laatste het doel is, is het aan te bevelen het programma *WCOPYfind* (zie verder) te gebruiken, omdat dat programma sneller werkt. Bij het hanteren van SafeAssign gaan alle papers eerst naar een server in de Verenigde Staten om van daaruit te worden vergeleken met internetdocumenten. Dat kan een paar uur duren. Overigens is dat in de meeste gevallen geen probleem. Studenten moeten hun paper vaak voor een bepaalde datum indienen. Stel dat studenten hun paper de avond voor de inleverdatum om 23.55 uur indienen. Dan wordt deze automatisch eerst via de computer doorgestuurd naar de Verenigde Staten om te worden vergeleken, maar docenten merken daar niet zo veel van omdat dit gedurende de nacht gebeurt. De volgende ochtend kunnen ze de ingediende papers op de eigen computer raadplegen, inclusief de uitkomst van de plagiaatcontrole.

SafeAssign vergelijkt tekstfragmenten via Microsoft Live Search met internetbronnen. Daarnaast raadpleegt Blackboard een aantal andere bronnen waarvan bekend is dat studenten ze hanteren bij het schrijven van papers. Dit zijn onder andere:

- digitale databases als Medline, PubMed
- de Findarticles-database van LookSmart (5,5 miljoen artikelen uit 900 periodieken)
- MyDropBox, met 300.000 documenten waarvan bekend is dat ze te koop worden aangeboden
- ProQuest ABI/Inform-database
- archief bestaande uit alle papers die door gebruikers van dezelfde Blackboard-omgeving in SafeAssign zijn geupload
- Global Reference Database: een archief bestaande uit alle papers die door de uploader zijn vrijgegeven als vergelijkingsmateriaal

Docenten moeten geautoriseerd worden voor toegang tot SafeAssign. SafeAssign kan zes typen tekst aan: Word (.doc en .docx), Open Document Format (.odt), Rich Text Format (.rtf), Portable Document Format (.pdf), HTML en platte tekst (.txt). Wanneer een student een paper indient bij SafeAssign, wordt deze paper automatisch toegevoegd aan het institutionele archief waarmee toekomstige ingeleverde papers binnen de instelling worden

vergeleken. Daarnaast heeft de student de optie de paper toe te voegen aan de zogenoemde Global Reference Database. Deze database bevat alle papers die ooit met SafeAssign zijn geüpload. Dit heeft als voordeel dat papers met elkaar worden vergeleken, zij het via de omweg van Blackboard (het hierna te bespreken WCopyfind is voor dat doel geschikter); een nadeel kan zijn dat de ingediende papers nu voor iedereen vrij toegankelijk worden. Immers, Blackboard toont twee fragmenten op het scherm: de oorspronkelijke paper die op plagiaat wordt gecontroleerd en het tekstfragment dat gevonden is in de grote database.

Nadat de noodzakelijke stappen zijn gedaan, bestaat de uitvoer van SafeAssign uit een aantal Matching Scores.

*Sentence Matching Scores:* deze geven in percentages de waarschijnlijkheid aan dat twee zinnen dezelfde betekenis hebben. Bijvoorbeeld: een score van 90% betekent dat het 90% waarschijnlijk is dat deze twee zinnen hetzelfde zijn en 10% waarschijnlijk dat ze bij toeval hetzelfde zijn.

*Overall Matching Scores:* deze is het gemiddelde van alle Sentence Scores, gewogen voor de lengte en de 'uniekheid' van de zin. Dit laatste is gebaseerd op de in de praktijk voorkomende frequentie van de woorden waaruit de zin is opgebouwd. De score moet worden opgevat als een indicator. Het is raadzaam alle papers met een hoge Overall Matching Score nader te bestuderen, waarbij rekening moet worden gehouden met de volgende interpretaties van de scores:

- Papers met een Overall Matching Score lager dan 15% bevatten meestal enkele citaten en/of slechts weinig zinsneden die uit andere documenten komen. Hierbij is doorgaans geen sprake van plagiaat.
- Waarschijnlijk bevatten papers met een score in de range 15%-40% (te) veel citaten uit andere documenten. Of er sprake is van plagiaat is niet bij voorbaat uit te sluiten. Nadere bestudering van de betreffende paper is wenselijk.
- Papers met een score hoger dan 40% bevatten zeer veel teksten uit andere documenten. Bestudering van de betreffende paper is noodzakelijk.

Nadat de noodzakelijke stappen zijn gedaan, komt het zogenaamde SafeAssign Report op het scherm. Daarin is gemakkelijk te vinden in welke bronnen bepaalde tekstfragmenten zijn te vinden. Iedere vindplaats heeft een andere kleur. Indien op de gekleurde tekst wordt geklikt, verschijnt de betreffende vindplaats en een tekstfragment uit de vindplaats waarmee de paper overeenkomt. De docent kan aan de hand van bestudering van de teksten besluiten of er sprake is van plagiaat.

## 6.6 WCopyfind

*WCopyfind* is een *free access* softwareprogramma (► <http://plagiarism.bloomfieldmedia.com/z-wordpress/software/wcopyfind/>) waarmee docenten elektronisch ingeleverde papers met elkaar kunnen vergelijken. Dit programma is bijzonder geschikt om papers over hetzelfde onderwerp met elkaar te vergelijken. Studenten dienen daartoe de papers elektronisch in te dienen. Docenten plaatsen de papers vervolgens in een bepaalde directory. De docent start het programma op, 'sleept' de documenten naar een venster en het pro-

gramma vergelijkt de teksten binnen enkele minuten. Het programma zoekt naar tekstfragmenten die identiek zijn. De lengte van de tekstfragmenten is in te stellen. Standaard zoekt het programma naar identieke tekstfragmenten die zes woorden beslaan.

Het programma kan verschillende tekst-formats aan, maar het is het overzichtelijkst wanneer studenten hun papers in hetzelfde format inleveren, bijvoorbeeld in Word. Ervaring leert dat studenten slordig zijn met het elektronisch aanleveren van teksten. Het is aan te bevelen duidelijke richtlijnen aan studenten te geven. Ook de naamgeving van de paper is belangrijk. Immers, via de paper moet de student traceerbaar zijn. De naam van de paper moet daarom minimaal de naam en het ID-nummer van de student bevatten. In een richtlijn voor studenten zouden bijvoorbeeld de volgende elementen kunnen staan:

- Lever de paper in als een attachment bij het volgende e-mailadres: ...
- De paper dient aangemaakt te zijn in Word.
- Zet alles in één document. Dus geen aparte documenten inleveren voor bijvoorbeeld tabellen en referentielijst.
- De Word-naam van het document moet zijn: (je eigen achternaam xxxxxx ... .doc) waarbij de xxxxxx je ID-nummer is (zonder de aanduiding ID).
- Indien het document niet aan deze richtlijnen voldoet, wordt het als niet-ingeleverd beschouwd.

Nadat de correcte procedure is gevolgd, verschijnt het File Comparison Report op het scherm. In de kolommen staat het aantal overeenkomstige woorden. Tussen de vierkante haken staat het aantal overeenkomstige woorden gerelateerd aan het totale aantal woorden in de paper.

Het programma geeft geen richtlijn voor signaalwaarden. De docent moet zelf bepalen of er sprake is van plagiaat. Het is daarbij handig steeds te beginnen bij dat paar papers waarvan het percentage overeenkomstige woorden het hoogst is. Als bij dit paar naar het oordeel van de docent geen sprake is van plagiaat, dan is de kans groot dat ook bij andere paren, met minder overeenkomstige woorden, geen sprake is van plagiaat. Voor de zekerheid kan dit worden gecontroleerd door ook andere paren te onderzoeken op plagiaat.

## 6.7 Besluiten tot het oordeel plagiaat

---

In dit hoofdstuk zijn twee softwareprogramma's voor het opsporen van plagiaat toegelicht. Beide programma's sporen identieke tekstfragmenten op en kwantificeren de mate van plagiaat. Maar overeenkomsten tussen twee tekstfragmenten zijn er altijd. Een docent moet op basis van de overeenkomsten besluiten: 'Dit is plagiaat en dát niet.' Blackboard/SafeAssign komt met suggesties: 'tussen 15% en 40% is er een redelijke kans dat er sprake is van plagiaat'. WCopyfind verbindt geen betekenis aan de gevonden percentages overlap. Met het hanteren van een plagiaatdetectieprogramma is de kous dus nog niet af. Een docent zal altijd moeten bekijken of studenten zich in de papers voldoende hebben verantwoord wat betreft de authenticiteit van hun tekst. Hiervoor zijn enkele richtlijnen te geven. Wanneer de programma's aangeven dat de overeenkomstige teksten op verschillende plaatsen en stukje bij beetje in de paper staan, zal er doorgaans geen sprake zijn

van plagiaat. Het loont niet kleine hoeveelheden tekst van anderen over te nemen. Ook overeenkomstige tekstfragmenten in de literatuurlijst duiden niet op plagiaat. Die horen juist met elkaar overeen te komen. Het begint pas verdacht te worden indien een bepaalde redenering van anderen is overgenomen zonder bronvermelding. Ook wanneer een docent een geheel overgeschreven alinea aantreft, is dat een reden om de betrokken student hierop aan te spreken. Bij het hanteren van WCopyfind ligt het iets gecompliceerder. Een docent weet niet bij voorbaat wie van wie heeft overgeschreven. Een gesprek met beide studenten ligt in de rede.

Het zal duidelijk zijn dat het plegen van plagiaat niet zonder sanctie zal kunnen blijven. Wanneer docenten plagiaat veronderstellen, is het raadzaam dat zij contact opnemen met de examencommissie van hun opleiding. Deze commissie heeft doorgaans meer ervaring met het omgaan met plagiaatgevalen dan individuele docenten. De sanctie, die uitsluitend de examencommissie mag opleggen, kan variëren tussen het ongeldig verklaren van de paper en/of het geven van een waarschuwing tot het uitsluiten van de student van examens voor de periode van een jaar. Dit laatste is conform wettelijke bepalingen de maximale straf. Een sanctie dient evenredig te zijn aan de overtreding. Een student meteen bij de eerste overtreding de maximale straf geven, lijkt niet verdedigbaar. Het behoort tot het scholingsproces van studenten wegwijs te worden gemaakt in de gebruiken van correct citeren en aanhalen. Pas wanneer de docent kwade trouw bewezen acht, liggen sancties voor de hand.

# Normeren en cijfergeven

*Henk van Berkel en Wynand Wijnen*

## Inleiding

Wat is het grote probleem bij normeren?

Welke twee conceptueel verschillende benaderingen zijn te onderscheiden?

Wat zijn compromisbenaderingen?

Welke compromismethoden zijn er?

Wat moeten docenten in de praktijk doen?

Wat zijn de eisen aan normeren?

**7.1 Twee methoden van normeren: relatief en absoluut – 74**

**7.2 Relatief normeren – 74**

**7.3 Absoluut normeren – 76**

**7.4 De relatieve en absolute methode vergeleken – 76**

**7.5 Compromismethoden – 77**

**7.6 Het proces van cijfergeven – 78**

Normeren, het relateren van een toetscore aan een criterium, en, daaruit afgeleid, cijfergeven, zijn in het onderwijs standaardactiviteiten voor een docent. Er worden cijfers gegeven voor examens en tentamens, voor proefwerken en mondelinge overhoringen, voor werkstukken en scripties, voor stages, praktische oefeningen, assessments, enzovoort. Vrijwel iedere activiteit van studenten binnen het onderwijs waarmee wordt bepaald of een student iets geleerd heeft (het summatieve toetsen), krijgt een waardering in de vorm van een cijfer. Een '5' of een '6' om het verschil tussen onvoldoende en voldoende aan te geven, een '8' of een '9' om onderscheid te maken tussen goed en nog beter. Meestal hebben de gegeven cijfers ook invloed op het verdere verloop van de studie. Behaalde onvoldoenden leiden er meestal toe dat bepaalde onderdelen opnieuw moeten worden gemaakt of uitgevoerd.

Normeren en cijfergeven vormen belangrijke gebeurtenissen in het onderwijs, in ieder geval voor studenten. Er is een veelheid aan methoden en procedures waaruit docenten iedere keer weer opnieuw een zo goed mogelijke keuze kunnen, en ook moeten, maken.

Het is niet zo verwonderlijk dat er geen eenduidige oplossing bestaat. Het allesoverheersende probleem bij normeren is het ontbreken van een criterium waartegen men de behaalde scores op toetsen kan afzetten om tot een cijfer te komen. De meeste opleidingen in het hoger onderwijs leiden op tot een bepaald beroep. Wanneer de afgestudeerde een slecht beroepsbeoefenaar wordt, heeft de opleiding niet goed geselecteerd. Kennelijk is het de opleiding dan niet gelukt een drempel op te werpen die de betreffende afgestudeerde had moeten stoppen. Met andere woorden, er zijn onderdelen in de opleidingen geweest waarvan de zak/slaaggrens niet correct is toegepast.

## 7.1 Twee methoden van normeren: relatief en absoluut

---

Bij het normeren gaat het erom een acceptabel criterium te vinden waartegen de toetscores kunnen worden afgezet om er een waardering aan te geven. Een cijfer drukt immers een waardering uit. Vooral de waarderingen onvoldoende en voldoende hebben grote gevolgen. Het grensgebied hiertussen is de zak/slaaggrens. Het uiteindelijke criterium is of de student waarover een zak/slaagbeslissing moet worden genomen, een competent beroepsbeoefenaar gaat worden. Docenten bezitten die informatie niet op het moment waarop het normeren aan de orde is en moeten zoeken naar een criterium dat wel voorhanden is. Zij kunnen dan kiezen uit twee principiële verschillende benaderingen om de toetscores van studenten te bepalen. De eerste is de toetscore van een student te vergelijken met die van andere studenten. Dit is *relatief normeren*, omdat cijfers worden bepaald door toetsresultaten vergelijkenderwijs te waarderen. De tweede benadering is de methode van *absoluut normeren*: de behaalde toetsresultaten worden afgezet tegen een vaste norm. In dat geval stelt de docent bijvoorbeeld dat studenten ten minste zestig procent van het maximale aantal punten moeten hebben behaald om te kunnen slagen.

## 7.2 Relatief normeren

---

Bij relatief normeren vergelijkt de docent de toetscore van een student met de toetscores van de overige toetsdeelnemers. Of een student een hoog cijfer verdient, is dan afhankelijk

van de toetsprestaties van anderen. Indien een student bijvoorbeeld 90 van de 100 vragen correct heeft beantwoord, en de medestudenten hebben allen een lagere score, dan zal de student een hoog cijfer krijgen. In feite plaatst de docent alle studenten op een denkbeeldige ladder. De student met de hoogste score op de toets, krijgt een plaats op de hoogste sport, de student met de laagste score krijgt een plaats op de onderste sport en de overige studenten krijgen een plaats ergens daartussenin, afhankelijk van hun toetsscore. Indien studenten eenzelfde score hebben behaald, staan zij even hoog op de ladder. De relatieve posities van studenten moeten daarna worden omgezet in een cijfer. De hoogte van hun cijfers is bij relatief normeren volledig afhankelijk van hun positie op de ladder ten opzichte van de medestudenten. Staan zij hoog op de ladder, dan verkrijgen zij een hoog cijfer, staan zij ergens onderaan, dan is hun cijfer laag. Wat voor cijfer zij precies krijgen, wordt bepaald door de docenten. Zij kunnen bijvoorbeeld besluiten dat de bovenste student een '10' krijgt, de onderste een '0' of '1', en de overigen een tussenliggend cijfer, overeenkomstig hun relatieve positie op de ladder. Omdat een dergelijke methode in praktijk erg bewerkelijk is vanwege het vele rekenwerk, zullen docenten die deze methode toepassen vaak met percentages werken die zij koppelen aan een bepaald cijfer. Zij kunnen bijvoorbeeld besluiten de vijf procent beste studenten een '10' te geven, de volgende tien procent een '9', enzovoort (Van Berkel, 1999).

De relatieve manier van normeren is vooral ingegeven door de gedachte dat studenten niet de dupe mogen worden van een toets die kwalitatief niet aan de maat is geweest. Daar zijn docenten verantwoordelijk voor geweest en niet de studenten. Studenten mogen daarvan geen negatieve gevolgen ondervinden. Een kwalitatief slechte toets is, net als slecht onderwijs, van invloed op de cijfers. Indien bijvoorbeeld de toets niet representatief was, zullen de toetsvragen niet aansluiten op wat studenten hebben bestudeerd. Bij toetsen waarvan de cijfers relatief zijn bepaald, hebben studenten hier geen last van. Hun score is immers gebaseerd op prestaties van de medestudenten op de toets en is niet een gevolg van een kwalitatief slechte toets. Overigens geldt wel dat bij een kwalitatief slechte toets, bijvoorbeeld een toets met een lage betrouwbaarheid (zie ► hoofdstuk 2), de relatieve positie van studenten op de ladder 'ruis' vertoont. Dat wil zeggen, je kunt er niet zeker van zijn of een bepaalde positie op de ladder, bijvoorbeeld plaats 20<sup>ste</sup>, inderdaad wordt bezet door de student die de 20<sup>ste</sup> plaats inneemt op de competentie die de toets beoogt te meten. Een onbetrouwbare toets plaatst de studenten verkeerd op de ladder en daardoor krijgen de studenten een verkeerd cijfer.

De relatieve methode is niet toepasbaar bij toetsen waaraan weinig studenten deelnemen. Relatief normeren is daarom problematisch wanneer de methode wordt toegepast op zogenoemde herkansingstoetsen, omdat aan deze toetsen doorgaans beduidend minder studenten deelnemen in vergelijking tot de oorspronkelijke toets. Het toeval speelt dan een te grote rol bij het bepalen van de relatieve posities van studenten en het daaraan gekoppelde cijfer. Pas bij meer dan vijftig toetsdeelnemers zal de relatieve methode een correcte indicatie geven over de relatieve positie van studenten. Bovendien geldt dat het bij relatief kleine groepen mogelijk is dat studenten onderling afspraken maken over de wijze van voorbereiding op de toets en op de wijze van invullen. Indien zij bijvoorbeeld afspreken een bepaald boek niet te gaan bestuderen, zal dit leiden tot een lager gemiddeld resultaat op de toets. Dit heeft echter geen invloed op cijfers die relatief zijn gegeven. De methode

is daardoor fraudegevoelig. Ook geldt, en dat is wellicht belangrijker, dat de groep deelnemers aan herkansingen een niet-representatieve steekproef vormt uit de gehele populatie. Het zijn meestal de minder competente studenten (anders waren ze oorspronkelijk niet gezakt). Het is daarom niet correct deze studenten weer op een denkbeeldige ladder te plaatsen en vanaf boven beginnen met het uitdelen van cijfers.

Een zuiver relatieve methode van cijfergeven is om bovenstaande redenen niet aan te bevelen om toe te passen.

### 7.3 Absoluut normeren

Bij absoluut normeren en cijfergeven houden docenten alleen rekening met de score van de student wiens toetsprestatie moet worden gewaardeerd. De docent vergelijkt de toetscore per student met een criterium dat vooraf is bepaald. In bijna alle gevallen zal het criterium bestaan uit een percentage van de maximaal te behalen score. Bijvoorbeeld, studenten die tachtig procent van de maximale score behalen, verkrijgen het cijfer '8', of: indien studenten veertig van de zestig driekeuzevragen correct beantwoorden, krijgen ze een '6'.

Veel docenten hanteren deze methode van normeren. Docenten, maar ook studenten, weten dan precies welke toetsscore tot welk cijfer leidt. Docenten kunnen hierdoor direct na de toets het modelantwoord plus een lijstje met voorlopige cijfers publiceren. Studenten verkrijgen dan een indicatie van de toetsuitslag. Dat schept duidelijkheid. Een groot nadeel van de absolute wijze van normeren is gelegen in onregelmatige slaagpercentages. Het ene jaar slaagt soms tachtig procent voor een toets, het andere jaar kan dat plotseling dertig procent zijn. Docenten streven er doorgaans naar hun toetsen jaar in jaar uit eenzelfde moeilijkheid te geven. Dit draagt bij aan de vergelijkbaarheid van toetsen over de jaren heen. In veel gevallen lukt dit niet. Het ene jaar is de gemiddelde moeilijkheid bijvoorbeeld zeventig procent ten opzichte van de maximale score, het andere jaar vijftig. Hiervoor is wel een reden te geven. Uit onderzoek (Van Berkel, 1984) blijkt namelijk dat zelfs ervaren docenten niet in staat zijn om voorafgaande de toetsafname een adequate schatting te geven van de moeilijkheid van de toetsvragen. Het lukt docenten niet, ondanks dat ze dat willen, een toets van een vooraf bepaalde moeilijkheid te maken. Dit verklaart de onverwacht hoge of lage slaagpercentages. Toetsen kunnen toevalligerwijs moeilijk of makkelijk uitpakken. Dit is het belangrijkste argument *tegen* een absolute wijze van normeren. In tegenstelling tot de relatieve methode van normeren houdt de absolute methode geen rekening met de groepsprestaties. Afspraken tussen studenten hebben dus geen invloed op de cijfers bij het absoluut cijfergeven.

### 7.4 De relatieve en absolute methode vergeleken

Onderstaand schema (■ tabel 7.1) is afkomstig uit Van Berkel (1999) en vergelijkt de relatieve en absolute methode van normeren op een aantal aspecten.



■ **Tabel 7.1** Een vergelijking tussen absoluut en relatief normeren (uit: Van Berkel, 1999).

	Absoluut	Relatief
tijdstip van vaststelling van de zak/slaaggrens	voorafgaande aan de toetsafname	na de toetsafname
wijze van vaststelling van de zak/slaaggrens	uitsluitend op basis van de toetsscore van de student	afhankelijk van de toetsresultaten van de andere toetsdeelnemers
invloed van niet-bedoelde moeilijkheid van de toets op het cijfer	beïnvloedt het cijfer	heeft geen invloed op het cijfer
percentage gezakten/geslaagden	100% gezakten of geslaagden is mogelijk	percentage ligt vast, ongeacht de feitelijke toetsresultaten
invloed van de kwaliteit van onderwijs en toets op het cijfer	ja	hebben bij betrouwbare toetsen geen invloed op het cijfer
aantal toetsdeelnemers	geschikt bij kleine en grote groepen	geschikt bij groepen > 50
geschiktheid bij herkansing	geschikt	niet geschikt
fraudegevoeligheid	ongevoelig	gevoelig: mogelijkheid tot afspraken tussen studenten
inzet studenten	inzet is van invloed op het cijfer	voldoende inzet leidt niet per se tot een voldoende

## 7.5 Compromismethoden

Het schema in ■ tabel 7.1 gaat uit van een zwart-witsituatie: of absoluut, of relatief. In de praktijk zal het normeren in veel gevallen wel iets van zowel het een als het ander in zich hebben. Veel docenten nemen de absolute methode als uitgangspunt. Daar is veel voor te zeggen. Relatief cijfergeven wordt niet vaak omarmd. Docenten vinden het vreemd om steeds per definitie een min of meer vast percentage van de toetsdeelnemers te moeten laten slagen. Om te kunnen slagen moeten studenten, naar hun idee, ieder voor zich kunnen aantonen dat ze de stof voldoende beheersen. Dat mag niet afhankelijk zijn van het kennisniveau van een medestudent. Het is daarom begrijpelijk dat er compromismethoden zijn voorgesteld die pogen de principiële tegenstelling tussen absoluut en relatief normeren te overbruggen.

Bij de compromismethoden gaan docenten over het algemeen uit van een absolute norm en stellen vast onder welke omstandigheden van deze absolute norm zal worden afgeweken, bijvoorbeeld bij te veel gezakten of bij een te laag bereikt kennisniveau. Maar ook de haalbaarheid van de maximale score speelt een rol. Er worden in praktijk vaak compromismethoden gebruikt. Daarom volgen hierna enkele voorbeelden.

- De *Hofstee-methode* (Hofstee, 1977) gaat uit van de situatie dat docenten in praktijk vaak een mix van absoluut en relatief normeren hanteren. De methode probeert dit te formaliseren en te rationaliseren. De methode gaat uit van twee absolute grenswaarden. De bovenste grens is het kennispercentage, bijvoorbeeld 65%, waarboven

de docent een student zal laten slagen, ook al zal dat betekenen dat *iedere* student slaagt. De onderste grens is het kennispercentage, bijvoorbeeld 45%, waaronder de docent een student zal laten zakken, ook al zal dat betekenen dat *iedere* student zakt. Tussen deze twee uitersten kan een docent laveren, afhankelijk van het aantal te verwachten onvoldoenden dat de docent acceptabel vindt.

- De *Absoluut-met-buffermethode* (Van Berkel, 1999) heeft als basis de waarschijnlijk meest gangbare praktijk van cijfergeven. Docenten stellen voorafgaande aan de toets vast welke ruwe score ten minste een voldoende oplevert. Deze methode van normeren gaat ervan uit dat docenten een juiste inschatting kunnen maken van de moeilijkheid van de toets. Dat is echter lang niet altijd het geval. Een toets kan onverwacht moeilijk uitvallen; dat leidt tot hoge zakpercentages. Er is daarom veel voor te zeggen dat de docent achteraf een zodanige correctie uitvoert dat het aantal gezakten binnen gebruikelijke proporties blijft. Bij toepassing van de Absoluut-met-buffermethode zal het zakpercentage niet al te veel afwijken van dat van voorgaande jaren.
- De *methode Cohen* (Cohen-Schotanus e.a., 1996) is een andere variant van absoluut normeren. De methode neemt de theoretisch hoogst bereikbare score als startpunt, maar stelt daarna bij op basis van de praktisch behaalde maximale score. Docenten stellen vervolgens de zak/slaaggrens vast, bijvoorbeeld bij 55% of 60% hiervan. Deze methode houdt echter geen rekening met de *haalbaarheid* van de hoogst bereikbare score. De maximale score wordt zelden behaald. Vooral bij toetsen bestaande uit meerkeuzevragen zal het weinig voorkomen dat er studenten zijn die alle vragen correct beantwoorden. Op de behaalde maximale score wordt vervolgens het vereiste kennispercentage toegepast. De achterliggende gedachte is dat er in een grote groep studenten altijd wel een paar studenten zijn die intellectueel begaafd zijn, zich optimaal hebben ingespannen tijdens het onderwijs en zich tevens goed op de toets hebben voorbereid. Deze studenten zouden dus eigenlijk een ‘10’ moeten krijgen. Als zij die niet krijgen, mankeert er iets aan de toets. Cohen-Schotanus neemt, om de invloed van een toevallige uitschieter te neutraliseren, als maximaal haalbare score de gemiddelde score die de vijf procent beste studenten heeft behaald. Maar variaties hierop zijn mogelijk.

## 7.6 Het proces van cijfergeven

---

Het voorgaande had vrijwel uitsluitend betrekking op het geven van cijfers uitgaande van een bepaalde methode van normeren. De relatieve methode plaatst studenten op basis van hun toetscore op een plaats op een ladder en afhankelijk van hun positie krijgen zij een cijfer. Bij de absolute methode vormt de score van de student ten opzichte van de maximaal mogelijke score het uitgangspunt. Vanwege allerlei problemen die ontstaan bij een zuivere toepassing van een van de twee methoden, hanteren docenten in praktijk een methode die het midden houdt tussen relatief en absoluut. Hebben docenten eenmaal een methode bepaald, dan gaan ze vrijwel automatisch over op het geven van cijfers aan de hand van een bepaalde berekeningswijze, veelal het toepassen van een formule. Dit is een mechanisch proces. Docenten denken dan dat ze op een rationele wijze de score hebben omgezet in een

cijfer. Dat is ook zo. Maar toch past een waarschuwing. Aan het begin van dit hoofdstuk is aangegeven dat er in feite geen correcte wijze van normeren is, omdat het criterium om te normeren (de kwaliteit van de toekomstige beroepsuitoefening) niet beschikbaar is op het tijdstip waarop de docent de toets afneemt. Het blijft dus een kwestie van 'het zo goed mogelijk doen'. Normeren is een proces, en niet het zomaar toepassen van een formule. Cijfergeven is een sluitstuk van een proces waarin docenten een keuze maken op grond van de informatie die voorhanden is. Docenten zullen bij het kiezen voor een bepaalde methodiek steeds afwegingen moeten maken.

# Formatief toetsen

*Desirée Joosten-ten Brinke en Dominique Sluijsmans*

## Inleiding

Wat is formatief toetsen?

Wat is de theoretische onderbouwing?

Hoe verhoudt formatief toetsen zich ten opzichte van summatief toetsen?

Welke methoden voor formatief toetsen zijn er?

Wat vereist formatief toetsen van docenten?

Hoe borg en evalueer je de kwaliteit van formatief toetsen?

## 8.1 Wat is formatief toetsen? – 82

## 8.2 Waarom is formatief toetsen belangrijk? – 83

## 8.3 Effectieve methoden van formatief toetsen – 84

8.3.1 Het geven van feedback – 84

8.3.2 Het voeren van een toetsdialoog – 85

8.3.3 Het stellen van effectieve vragen – 86

8.3.4 Zelfbeoordeling – 86

8.3.5 Peer-assessment – 87

8.3.6 Rubrics – 88

8.3.7 Formatief gebruik van summatieve toetsen – 88

## 8.4 De rol van docenten bij formatief toetsen – 88

## 8.5 Controlepunten voor docenten – 89

Toetsing wordt niet langer uitsluitend gebruikt ter afronding van een instructieblok of voor certificering, het zogenoemde summatief toetsen. Zowel maatschappelijk als wetenschappelijk is er steeds meer aandacht voor toetsen om het *leren* te bevorderen, ook wel formatief toetsen genoemd. Dit betekent dat naast gangbare kwaliteitscriteria als betrouwbaarheid en validiteit, ook wordt gelet op de aansluiting tussen leren, instructie en toetsing, en de mate waarin studenten worden betrokken bij toetsing (Baartman, Bastiaens, Kirschner & Van der Vleuten, 2006).

## 8.1 Wat is formatief toetsen?

Een gangbare definitie van formatief toetsen is:

» het proces van het zoeken, aggregeren en interpreteren van informatie die studenten en docenten gebruiken om te bepalen waar studenten staan in hun leerproces, waar zij naar toe moeten werken en op welke manier (ARG, 2002). «

Drie vragen staan hierbij centraal:

- Wat is de gewenste eindsituatie (eindtermen, leerdoelen)?
- Waar bevinden studenten zich in hun ontwikkeling ten opzichte van de gewenste eindsituatie?
- Hoe overbrug je de kloof tussen de huidige en gewenste eindsituatie?

De informatie moet zodanig betekenisvol zijn dat de student deze kan gebruiken om verder te leren zonder gedemotiveerd te raken. Bovendien is het van belang dat deze informatie de student helpt het leerproces mede zelf te sturen. Formatief toetsen kan op verschillende manieren plaatsvinden, bijvoorbeeld formeel door oefentoetsen of minder formeel door het stellen van de juiste vragen (Sluijsmans, Joosten-ten Brinke & Van der Vleuten, 2013). Cauley en McMillan (2009) onderscheiden een laag en hoog niveau van formatief toetsen op basis van elf aspecten (zie ■ tabel 8.1).

Het ontwerp en de inrichting van formatief toetsen worden bepaald door het doel dat een docent eraan geeft. Als studenten elkaar feedback moeten leren geven, is het belangrijk veel momenten te creëren waarop studenten dit met elkaar kunnen oefenen (aspect 2). Als een docent betrokkenheid van studenten in toetsing wil stimuleren, is het zinvol studenten integraal te betrekken bij bijvoorbeeld het proces om criteria op te stellen (aspect 9). Formatief toetsen kan *on-the-fly* plaatsvinden of ingebed en georganiseerd (Shavelson e.a., 2008). Met *on-the-fly* formatief toetsen worden die instructieactiviteiten bedoeld die onverwacht en ongepland plaatsvinden, bijvoorbeeld als een docent een groep studenten beluistert tijdens het werken aan een taak en ter plekke feedback geeft. Een vorm van georganiseerd formatief toetsen is het plannen van een bijeenkomst om met studenten beoordelingscriteria te formuleren.

■ **Tabel 8.1** Laag versus hoog niveau van formatief toetsen (Cauley & McMillan, 2009).

Aspecten		Laag niveau	→	Hoog niveau formatief toetsen
1	aard van het bewijs	objectief, gestandaardiseerd		gevarieerd, objectief en anekdotisch
2	structuur	meestal formeel, gepland		informeel, spontaan
3	betrokkenen bij beoordelen	docenten		docenten en studenten
4	feedback	vertraagd en algemeen		direct en specifiek voor laag presterende studenten, vertraagd voor hoog presterende studenten
5	wanneer uitgevoerd	na instructie		voor, tijdens en na instructie
6	aanpassing in de instructie	meestal voorgeschreven en gepland		flexibel, niet gepland
7	keuze van taken	door docent bepaald		door docent en student samen bepaald
8	aard van de interactie	interacties zijn gebaseerd op formele rollen		informeel, in vertrouwen, en eerlijke interacties
9	rol van student zelfevaluatie	weinig tot geen		integraal
10	type motivatie	extrinsiek		intrinsiek
11	toewijzen van succes	externe factoren		interne factoren (de moeite die is gedaan)

## 8.2 Waarom is formatief toetsen belangrijk?

Formatief toetsen is een manier om studenten zelf kennis en vaardigheden te laten ontwikkelen. Het wordt gezien als een proces waarin student en docent met elkaar communiceren over doelen, criteria en leerproblemen. Bennett (2011) relateert formatief toetsen vooral aan leren in sociale interactie, aan een dialoog over een leerproces in een onderwijskundige relatie tussen student en docent of tussen studenten onderling. Socio-culturele theorieën gaan ervan uit dat studenten effectiever leren in interactie met anderen, vooral met hen die meer domeinexpertise hebben. Formatief toetsen is gericht op het verhogen van de motivatie van studenten. Een van de bekendste motivatietheorieën is de *zelfdeterminatietheorie*, waarin competentie, autonomie en sociale verbondenheid de basisbehoeften zijn (Ryan & Deci, 2000). Het ontbreken van één van deze basisbehoeften kan de motivatie dusdanig verstoren dat leren niet plaatsvindt. Daarom staat actieve deelname van studenten aan het leerproces (aspect 3 uit ■ tabel 8.1) ook in socio-culturele theorieën centraal (Clark, 2012).

Feedback is een essentieel element van formatief toetsen (zie onder andere Hattie & Timperley, 2007; Shute, 2008) en levert een bijdrage aan de ontwikkeling van zelfregulerend leren bij studenten. In het proces van zelfregulatie beschrijft Sadler (1989) expliciet het belang van beoordelingsvaardigheden. Het systeem van instructie en leren moet

studenten ruimte bieden deze vaardigheden te ontwikkelen zodat zij niet alleen afhankelijk zijn van het oordeel van de docent.

Samenvattend: belangrijke elementen van formatief toetsen zijn de interactie tussen docenten en studenten, het behoud van motivatie om verder te willen leren en het gebruik van feedback om de zelfregulatie van studenten te bevorderen.

### 8.3 Effectieve methoden van formatief toetsen

---

Er zijn verschillende methoden van formatief toetsen die effectief zijn voor leren:

- het geven van feedback
- het voeren van een toetsdialoog
- het stellen van effectieve vragen
- zelfbeoordeling
- peer-assessment
- rubrics
- formatief gebruik van summatieve toetsen

Deze worden in de volgende paragrafen uitgelegd.

#### 8.3.1 Het geven van feedback

---

Effectieve feedback bevat volgens Hattie en Timperley (2007) antwoorden op drie vragen:

- Wat is het einddoel van de student (wat zijn de beoordelingscriteria en standaarden)? Dit wordt benoemd als *feed-up*.
- Hoe heeft de student de taak uitgevoerd (welke vooruitgang wordt geboekt ten aanzien van de beoordelingscriteria en standaarden)? (*feedback*).
- Welke aanpak is nodig om het doel te bereiken? (*feed-forward*).

Centraal staat het overbruggen van de kloof tussen wat de student wel beheerst en wat niet (Sadler, 1989). Om de feedback verder te verhelderen, kunnen de drie bovenstaande vragen vervolgens betrekking hebben op vier aspecten:

- criteria en standaarden van de taak ('De probleemstelling van je onderzoek is nog niet correct en nog niet van voldoende niveau')
- proces van probleem oplossen ('Bij het opstellen van de probleemstelling is het goed na te gaan hoe een aantal betrokkenen het probleem ervaren')
- niveau van het zelfsturingproces ('Bekijk je planning nog eens en geef aan waarom het niet gelukt is om de probleemstelling op tijd in te leveren')
- eigenschappen van de student ('Goed gewerkt!')

Deze laatste draagt er echter niet toe bij dat een student leert.

Duidelijke criteria en kwalitatief goede feedback zijn cruciaal voor het leerproces (Black & Wiliam, 2009). Het is bij het geven van feedback belangrijk om de juiste toon,

stijl en omvang te hanteren, zodat ook minder gemotiveerde studenten en studenten met weinig zelfkennis de feedback gebruiken (Wingate, 2010). Shute (2008) geeft de volgende richtlijnen voor effectieve feedback:

- Richt de feedback op de taak, niet op de student.
- Geef uitgewerkte feedback waarbij het wat, hoe en waarom van een gegeven probleem wordt beschreven.
- Geef de feedback in overzichtelijke eenheden om cognitieve overbelasting te voorkomen.
- Wees specifiek en helder in je feedbackboodschap.
- Houd de feedback zo eenvoudig mogelijk, maar niet te simpel.
- Geef zuivere, objectieve feedback, geschreven of via de computer.
- Geef feedback nadat studenten hebben geprobeerd een antwoord of oplossing te geven.
- Geef geen normatieve vergelijkingen.
- Wees terughoudend in het geven van algemene cijfers; de effectiviteit van feedback is gelegen in de inhoud van de opmerkingen.
- Geef geen feedback die het zelfvertrouwen van de student verlaagt en sluit aan op het niveau van de student.
- Probeer te voorkomen dat de feedback mondeling wordt gegeven in verband met objectiviteit.
- Onderbreek studenten niet met feedback als zij bezig zijn hun antwoord te formuleren of een probleem op te lossen.
- Zorg dat het gebruik van hints niet automatisch leidt tot het geven van het juiste antwoord.
- Beperk uitgebreide foutanalyses; deze kosten zeer veel tijd, zijn zelden volledig en leveren relatief weinig leerwaarde op.
- Geef feedback op het moment dat dat het effectiefst is: direct om fouten te corrigeren en vertraagd voor een betere transfer van leren.

Het geven van goede feedback is moeilijk en hangt sterk af van een specifieke onderwijscontext. Wanneer onvoldoende rekening wordt gehouden met deze context, kan feedback zelfs een negatief effect op leren hebben (Torrance & Pryor, 2001).

### 8.3.2 Het voeren van een toetsdialoog

---

De toetsdialoog is een specifieke gespreksvorm tussen docenten en studenten die plaatsvindt in de les tijdens instructie. Centraal in een toetsdialoog staan de te bereiken leerdoelen. In de vraagstelling van de docent wordt direct duidelijk welk leerdoel dat is. Een startvraag van een docent bij een leerdoel 'inzicht hebben in interpersoonlijk gedrag' kan bijvoorbeeld zijn: 'Waarom kijken we naar de manier waarop je met leerlingen praat om te bepalen of je interpersoonlijk competent bent?' Toetsdialogen helpen docenten om continu informatie te verzamelen over hoe het staat met het begrip van studenten en om hun instructieactiviteiten zodanig aan te passen dat de studenten de leerdoelen bereiken.



De voorbereiding van de docent op deze toetsdialogen betreft het zelf helder hebben van het doel van de les of lessenserie is. Toetsdialogen bestaan uit verschillende stappen die elkaar cyclisch opvolgen: initiëren, antwoorden en evalueren of follow-up. Hoewel de docent veelal degene is die een toetsdialoog start, kan de dialoog ook volgen op een opmerking van een student zonder dat er een vraag is gesteld. De docent blijft daarbij echter gespreksleider en ziet zo'n opmerking dan als 'trigger' om de dialoog aan te gaan. In een toetsdialoog kunnen studenten naar elkaar toe ook een effectieve rol spelen. Medestudenten kunnen reageren op de informatie die in een toetsdialoog wordt gegeven. De docent wordt dan meer een procesondersteuner om de dialoog tussen de studenten te leiden tot feedback of tot een gezamenlijke beslissing (Ruiz-Primo, 2011).

Een effectieve toetsdialoog 1) is gebaseerd op leerdoelen, 2) is interactief; de antwoorden van de studenten kunnen bestaan uit argumenten en input zijn voor discussie, 3) wordt ingezet om leren te 'scaffolden' met een goede balans tussen uitdagen en begeleiden van studenten richting de leerdoelen, 4) bevordert sociale participatie en de ontwikkeling van vaardigheden van hogere orde en 5) zorgt ervoor dat studenten worden ondergedompeld in de cultuur van het vak. Het goed afsluiten van een toetsdialoog is cruciaal voor het succesvol gebruik ervan (Ruiz-Primo, 2011).

### 8.3.3 Het stellen van effectieve vragen

---

Een derde methode van effectief formatief toetsen is het stellen van vragen aan studenten. Hun antwoorden geven docenten inzicht in hoe studenten denken en dit maakt het mogelijk om in te grijpen, misconcepties te weerleggen en dieper leren te stimuleren (Black & Wiliam, 2009). Voor het verzamelen van de juiste informatie is het stellen van de juiste vragen (*rich questioning*) een belangrijke vaardigheid van de docent. Vragen, die zowel door de docent als de student worden gesteld, kunnen verschillen in vorm en functie. De manier van vragen bepaalt of en hoeveel de studenten moeten nadenken over hun antwoord. Het beste resultaat wordt bereikt als de vragen: 1) open zijn omdat die de studenten stimuleren om hun gedachten te vormen, 2) zich richten op datgene wat een student denkt ('Welke formule denk je dat je moet gebruiken?' in plaats van 'Welke formule moet je gebruiken?'), 3) aansluiten bij verschillende soorten kennis, te weten beschrijvende kennis (*wat-vragen*), procedurele kennis (*hoe-vragen*), schematische kennis (*waarom-vragen*), en strategische kennis (*waarom wanneer- en hoe te gebruiken-vragen*), 4) de studenten ruimte tot nadenken geven (en niet onmiddellijk hoeven te worden beantwoord, maar bijvoorbeeld na tien seconden) en 5) niet bedreigend overkomen, maar als ondersteuning worden ervaren om verder te komen in het leerproces.

### 8.3.4 Zelfbeoordeling

---

Zelfbeoordeling (*self-assessment*) is een methode waarin studenten hun eigen werk lateren aan de leerdoelen. Het bestaat uit drie stappen: studenten beoordelen hun eigen werk, identificeren verschillen tussen hun werk en de gewenste uitkomsten, en voeren

nieuwe activiteiten uit totdat de einddoelen bereikt zijn. Zelfbeoordeling staat veelal niet op zichzelf, maar wordt gecombineerd met andere inhoudelijke opdrachten. Het kunnen uitvoeren van een goede zelfbeoordeling is effectief voor leren en noodzakelijk voor effectief leren. Studenten kunnen alleen reflecteren op hun eigen handelen als zij doelen en taken begrijpen, kunnen inschatten hoever ze zijn gevorderd en weten hoe zij hun doelen kunnen bereiken. Het uitvoeren van een zelfbeoordeling helpt de student een taak bij te stellen zodat de leeruitkomst beter aansluit bij de beoordelingscriteria. Zelfbeoordeling draagt bovendien bij aan het ontwikkelen van zelfregulatievaardigheden. De eigen verantwoordelijkheid van studenten groeit door het zelf beoordelen van eigen werk. Studenten krijgen een beter overzicht van de te bereiken doelen en hun positie ten opzichte van deze doelen. Studenten die zichzelf goed kunnen beoordelen, ontwikkelen een eigen autonomie, waardoor ze de taken beter begrijpen en weten welke stappen ze moeten zetten richting het gewenste einddoel. Het goed kunnen uitvoeren van een zelfbeoordeling reduceert bovendien de tijd die docenten nodig hebben om de prestaties van studenten te meten (Black & Wiliam, 2009).

### 8.3.5 Peer-assessment

---

Bij peer-assessments beoordelen medestudenten elkaars werk (zie ook ► hoofdstuk 24). Het proces van zelfbeoordelingen komt in grote mate overeen met het proces van peer-assessment, waarbij bij peer-assessments de nadruk ligt op de meerwaarde van samenwerking. Peer-assessment zorgt ervoor dat medestudenten worden betrokken bij het toetsproces door elkaars producten te beoordelen en elkaar te voorzien van feedback. Studenten geven aan dat peer-assessment bijdraagt aan het begrijpen van de criteria om te komen tot een goed resultaat (Black & Wiliam, 2009).

Als studenten gemeenschappelijke doelen hebben en het resultaat van hun prestatie afhankelijk is van het doen en laten van de andere groepsleden, gaan groepsdynamische processen een rol spelen. Er is sprake van positieve interactie als de groepsleden ervan overtuigd zijn dat ze deze groepsdoelen alleen kunnen bereiken als ze onderling samenwerken en elkaar voortdurend stimuleren en faciliteren (Johnson, Johnson & Smith, 2007). Vertrouwen in de groep is een belangrijke voorwaarde voor het succes van peer-assessment (Newby & Winterbottom, 2011). Daarnaast stelt groepswerk ook eisen aan de onderlinge communicatie en het kunnen geven en ontvangen van feedback.

Peer-assessment draagt bij aan het autonoom uitvoeren van zelfbeoordelingen en zelfregulerend leren (Allal & Lopez, 2005; Birenbaum, Kimron & Shilton, 2011). Self- en peer-assessments kunnen goed worden gekoppeld aan het werken met rubrics (zie volgende paragraaf en ► hoofdstuk 17) en aangezien het implementeren van peer-assessment veel tijd kost, is het verstandig om dit deel te laten uitmaken van al bestaande onderwijstaken (Newby & Winterbottom, 2011). Zoals ook al is aangegeven bij zelfbeoordeling, is training in peer-assessment van belang om het effectief te laten zijn voor leren (Birenbaum e.a., 2011; Black & Wiliam, 2009).

### 8.3.6 Rubrics

---

Een andere effectieve methode om formatief toetsen te ondersteunen is het werken met rubrics. Rubrics beschrijven aan de hand van een set criteria de verschillende niveaus van kwaliteit gerelateerd aan de leerdoelen. Rubrics kunnen zowel worden gebruikt voor formatieve doeleinden (om 'werk in ontwikkeling' te beschrijven) als voor summatieve doeleinden (als beoordelingscriteria om een leeruitkomst te beoordelen). In beide gevallen zijn ze geschikt om tijdige en informatieve feedback te geven. Dit zorgt voor een hogere motivatie en het bijsturen van het leren (zie ► hoofdstuk 17).

### 8.3.7 Formatief gebruik van summatieve toetsen

---

Een belangrijke vraag uit de onderwijspraktijk is hoe formatief en summatief toetsen samen goed in een toetsprogramma kunnen worden ingezet. Het primaire doel van formatief toetsen is het stimuleren van leren, het primaire doel van de summatieve toetsen is selectie of certificering. Secundair kunnen summatieve toetsen echter ook worden ingezet om ervan te leren en formatieve toetsen om te selecteren (Bennett, 2011). Zo kan een afsluitend examen door het bieden van feedback ook een formatieve functie hebben. Het doel en het ontwerp van de toets, de wijze van het verzamelen van informatie, het interpreteren van de informatie en de communicatie naar de student bepaalt of een toets formatief of summatief is. Het verschil ertussen hangt dus in hoge mate samen met de doelen die men beoogt en hoe de uitkomsten worden gebruikt. Het verschil heeft minder te maken met het tijdstip waarop de toetsen plaatsvinden (Sadler, 1989). Zo kunnen formatieve en summatieve toetsen op elk moment in een curriculum worden ingezet. Om dit te realiseren moet bij beide doelen sprake zijn van transparantie over criteria en wijze van beoordelen en is het belangrijk om ook bij toetsen met een summatief doel rijke feedback te geven. Daarnaast kunnen oude of vergelijkbare toetsen worden gebruikt als voorbereiding op de toets. Het vooraf maken van een (vergelijkbare) toets levert een positief effect op voor het uiteindelijk behalen van een toets (het zogenaemde *testing-effect*; Roediger & Karpicke, 2006).

## 8.4 De rol van docenten bij formatief toetsen

---

Scholing van docenten richt zich veelal op het *summatief* toetsen en beoordelen. Docenten zijn over het algemeen niet of nauwelijks geschoold in formatief toetsen (Popham, 2009). Formatief toetsen vereist een organisatie en een cultuur die openstaat voor vernieuwingen. De kwaliteit van de docent bepaalt in hoge mate het succes van formatief toetsen (Allal & Lopez, 2005; Birenbaum e.a., 2011). Docenten bepalen zelf, vanuit een bepaalde visie, op welke manier vorm wordt gegeven aan formatief toetsen. Docenten hebben hiervoor pedagogische kennis en vaardigheden nodig die hen in staat stellen de kloof tussen wat studenten kennen en moeten kennen vast te stellen en vervolgens om adequaat te handelen. De toetsbekwame docent is niet alleen bekwaam om een oordeel te geven, maar ook

bekwaam in het beslissen welke criteria van belang zijn in de onderbouwing van het oordeel. Vaak vinden docenten het moeilijk om vooraf alle criteria te formuleren of om criteria toe te passen op elke student. Hoewel het bij formatief toetsen vooral gaat om toetsen in functie van leren en minder over het statistisch analyseren van toetsen, is kennis over toetsanalyse en interpretatie van kwantitatieve toetsresultaten van belang. Door toetsen te analyseren krijgt de docent inzicht in de beheersing van de verschillende leerdoelen en kan daarmee adequate feedback geven (zie ► hoofdstuk 2).

Uit ervaringen met professionaliseringsactiviteiten op het gebied van formatief toetsen blijkt dat de wijze waarop een docent toetst ook in hoge mate wordt beïnvloed door de opvattingen en overtuigingen die docenten zelf hebben over leren en toetsing (Birenbaum e.a., 2011; Brookhart, Moss & Long, 2008). De opvattingen van docenten over formatief toetsen en al dan niet impliciete theorieën over leren, beïnvloeden sterk het effect van formatief toetsen (Sadler, 1989). Zo zal een docent met de opvatting dat toetsing is bedoeld voor certificering deze opvatting ook laten doorwerken in de implementatie van formatieve toetsen. Het veranderen van een toetspraktijk vereist ook een verandering in concepties. Een onderzoekende houding naar de eigen overtuigingen blijkt voorwaardelijk voor een structurele verandering in toetspraktijken, maar hangt ook vaak samen met persoonskenmerken. Een belangrijke conditie voor het vormgeven en implementeren van formatieve toetsen op een manier die draagvlak creëert bij alle docenten is het investeren in een leergemeenschap (*community of learning*). Wanneer de overtuigingen en opvattingen van docenten niet in overeenstemming zijn met het beleid van de opleiding en de opleidingscultuur en er te weinig wordt geïnvesteerd in de opleiding als leergemeenschap, bepalen deze overtuigingen en concepties de onderwijspraktijk. Het is van belang docenten van meet af aan te betrekken bij visievorming over toetsen in de opleiding en welke kernwaarden daarbij van belang zijn. Dit verhoogt niet alleen het draagvlak, maar maakt het ook mogelijk om hierover vanuit een gemeenschappelijk kader te communiceren naar studenten.

## 8.5 Controlepunten voor docenten

---

Aansluitend op de vorige paragraaf is het van belang docenten handvatten te bieden die hen helpt formatief toetsen in hun eigen onderwijspraktijk vorm te geven. In [tabel 8.2](#) zijn vijftien controlepunten opgenomen die docenten kunnen gebruiken om te bepalen wat ze doen aan formatief toetsen en wat ze ermee zouden willen bereiken.

Bij elk controlepunt kunnen docenten aangeven of ze dat doen of nog niet. Daarbij kunnen ze voor zichzelf aangeven hoe ze dat doen en welk effect het heeft op het leren van de studenten, op welke manier ze het wel zouden willen doen en welke hulp ze daarbij nodig zouden hebben.

De antwoorden kunnen vervolgens worden gebruikt om met collega's te delen en formatief toetsen in de opleiding of faculteit verder vorm te geven. De vragen hebben uitdrukkelijk geen normatief karakter, dat wil zeggen er kan niet goed of fout op worden geantwoord.

■ **Tabel 8.2** Controlepunten om te bepalen wat een docent al doet aan formatief toetsen.

Stelling		Doe ik al	Doe ik (nog) niet
1	Ik heb de eindtermen of leerdoelen voor mijn vak transparant beschreven.		
2	Ik communiceer naar mijn studenten welk bewijs (gedrag of antwoorden) studenten kunnen leveren om aan te tonen dat eindtermen of leerdoelen bereikt zijn.		
3	Ik voer met studenten gesprekken over het ontwerp van een toets.		
4	Ik geef feed-up aan mijn studenten.		
5	Ik geef feedback aan mijn studenten.		
6	Ik geef feed-forward aan mijn studenten.		
7	Ik stel mijn studenten vragen voor, tijdens en na de instructielessen.		
8	Ik pas mijn instructie aan op basis van toetsresultaten.		
9	Ik stel samen met de studenten beoordelingscriteria op.		
10	Ik maak gebruik van rubrics.		
11	Ik weet of mijn studenten de leerstof beheersen.		
12	Ik laat studenten zelfbeoordelingen uitvoeren.		
13	Ik laat studenten peer-assessment uitvoeren.		
14	Ik bespreek toetsresultaten met studenten.		
15	Ik weet hoe ik studenten kan motiveren voor een toets.		

# Digitaal toetsen

*Silvester Draaijer, Patris van Boxel en Marcel van Brunschot*

## Inleiding

Wat zijn de mogelijkheden en beperkingen van digitaal toetsen?

Wat zijn de mogelijkheden van een itembank?

Kunnen computers open vragen nakijken?

Wat zijn de ontwikkelingen in de toekomst?

## 9.1 Twee voorbeelden van toetsafname met de computer – 92

### 9.2 De computer als toetsmachine – 94

9.2.1 Afsluitende tentamens afnemen met de computer – 94

9.2.2 Formatieve toetsen afnemen met de computer – 95

### 9.3 Itembanken – 96

### 9.4 Nakijken van open vragen en werkstukken – 97

### 9.5 Echt andere toetsvormen: de toekomst? – 99

Digitaal toetsen gaat over het gebruik van de computer bij summatieve en formatieve toetsen. Het doel van het hoofdstuk is inzicht te bieden in de mogelijkheden en beperkingen van de computer bij toetsing. De nadruk ligt op relatief eenvoudig te gebruiken computerondersteunde toetsystemen en bevat praktische adviezen voor docenten die computerondersteund toetsmateriaal willen ontwikkelen, selecteren of gebruiken.

## 9.1 Twee voorbeelden van toetsafname met de computer

De computer is in de verschillende fasen van het toetsproces een belangrijk hulpmiddel. Het *ontwikkelen en afnemen* van toetsen op de computer staat in dit hoofdstuk centraal. Daarbij komt zowel de didactische inzet als de organisatie en techniek aan de orde. Hieronder twee voorbeelden waarin de computer een belangrijke rol speelt.

### Voorbeeld 1: De Nationale Kennistoetsbank voor de zorg

Stichting Leerstation Zorg (SLZ) beheert de Nationale Kennistoetsbank voor de Zorg, waarin leermateriaal is opgeslagen in een databank. Deze wordt ontsloten via een leermanagementsysteem. De kennistoetsbank bestaat op dit moment uit ruim 12.000 digitale toetsvragen. De toetsvragen worden kennisitems genoemd omdat er bij elke vraag feedback in de vorm van een kennisflits staat aangeboden. Een kennisflits is een (korte) toelichting bij het antwoord op de toetsvraag met een argumentatie van de constructeurs waarom de antwoordopties in deze context juist of onjuist zijn. Hiermee krijgt de student de uitleg waarom een bepaald antwoord goed of fout was meteen te lezen. Ook de bron staat vermeld, gebaseerd op de laatste actuele stand van de wetenschap of op basis van landelijk geldende richtlijnen en/of protocollen. Door bij de toetsvragen aan te geven uit welke bron de informatie is verkregen, kan op ieder moment worden nagegaan of een vraag nog voldoende actualiteitswaarde heeft. Een ander voordeel van bronvermelding is dat de student zelf actief op zoek kan gaan naar aanvullende informatie. De student kan met open tekst op elk kennisitem direct commentaar geven, waardoor er interactiviteit mogelijk is.

De vragen hebben verschillende vormen. Zo kan men kiezen uit meerkeuzevragen, stellingvragen, sleepvragen of aanwijsvragen. De ontwikkeling van de kennisitems gebeurt via een vast stramien. Er wordt gebruikgemaakt van de zogenaamde 'kritische incidentmethode'. Onder intensieve begeleiding van medewerkers van Leerstation Zorg stellen zorgverleners en docenten in onderling overleg kritische praktijksituaties vast en formuleren daar vragen over. Docenten kunnen zelf toetsen samenstellen uit de kennistoetsbank en gebruiken in het onderwijs. Het hanteren van toetsen via internet is niet aan tijd of plaats gebonden. Het leermateriaal kan formatief worden gebruikt: als leermiddel en voor (zelf)toetsing, om kennisontwikkeling te volgen, in combinatie met face-to-face onderwijs en andere digitale leermogelijkheden. Daarnaast kunnen de items summatief worden aangeboden: toetsen die meetellen voor beoordeling en examens.

Een goed voorbeeld van het inzetten van 'toetsen om te leren' zijn de stagetoetsen. Bij hbo-v-opleidingen worden ze gehanteerd als voorbereiding op de stage. Studenten krijgen dan een kennistoets over het specialisme van de organisatie waar zij stage gaan lopen. Door het maken van de toets krijgen zij inzicht in het vereiste kennisniveau en hun eigen kennisniveau.

#### Voorbeeld 2. Digitale Toetszaal VU Amsterdam

Bij de Vrije Universiteit worden sinds 2000 toetsen afgenomen via computers. Dit geldt specifiek voor vakken waarin studenten worden getoetst in het oplossen van problemen met behulp van software zoals SPSS, MatLab, en dergelijke. De toetswijze geldt ook voor vakken waarvoor het van belang is dat er vragen aan studenten kunnen worden aangeboden waarin kleurafbeeldingen, video of andere vraagtypen nodig zijn of waar een docent werkt met digitale vragenbanken. Ten aanzien van de uitvoering in de praktijk zijn er echter vele problemen. Denk aan het regelen en financieren van geschikte software, het technisch functioneren (lettertypes wisselen, scoring vergeten in te stellen, navigatieknoppen vergeten aan te zetten enzovoort), het roosteren en inrichten van meerdere zalen, het laten aanbrengen van de benodigde beveiliging tegen afkijken en fraude, het regelen en instrueren van surveillanten, het analyseren van toetsen en het bepalen en regelen van inzage: de docent moet voor al deze problemen een oplossing zoeken. Hierdoor haken docenten af of beginnen er überhaupt niet aan. Eind 2009 is de roep op de VU groter en groter om deze toetsen efficiënter en professioneler af te gaan nemen voor vooral grote cohorten studenten.

Op centraal niveau is besloten dat er een onderzoek moet komen naar de haalbaarheid van een omvangrijke digitale toetsvoorziening. Om de investering te kunnen rechtvaardigen, is het nodig dat alle stakeholders van de instelling (studenten, onderwijsdirecteuren, decanen, College van Bestuur, onderwijsbureaus, IT-diensten, facilitaire diensten, roosteraars) het nut en de noodzaak van een dergelijke voorziening inzien en actief bekrachtigen. Er is een voorlichtings- en besluitvoorbereidings-traject ingezet. Uiteindelijk is er een positief besluit genomen en in maart 2013 is deze voorziening gereedgekomen. Alle digitale toetsen kunnen vanaf dat moment in één zaal met ongeveer 385 volledig gecontroleerde en gemonitorde werkplekken worden afgenomen. Er is centrale ondersteuning met toetssoftware en gespecialiseerde toetsondersteuners. De docent concentreert zich op het ontwikkelen van goede toetsen en laat de techniek, beveiliging en logistiek over aan specialisten. De zaal staat borg voor een efficiënt en technisch betrouwbaar afnameproces voor zowel tentamens als tus-sentoetsen.



## 9.2 De computer als toetsmachine

Zoals uit de voorbeelden blijkt, kunnen toetsen worden afgenomen via de computer, maar kunnen er grote praktische problemen optreden. Een groot voordeel is dat de computer bij uitstek geschikt is om antwoorden op toetsvragen digitaal te verwerken.

### 9.2.1 Afsluitende tentamens afnemen met de computer

Bij digitaal toetsen is het mogelijk te toetsen met verschillende vraagtypen en verschillende vraagvormen. Digitale toetsen kunnen bijvoorbeeld bestaan uit toetsen waarbij gebruik moet worden gemaakt van beroepssoftware: dat is software die specifiek is voor een bepaalde beroepspraktijk (denk aan SPSS, databases, Excel en dergelijke). Alle antwoorden of responsen worden digitaal verzameld en verwerkt. Dat maakt het beoordelen van het werk efficiënter. Voor open vragen geldt dat studenten gebruik kunnen maken van de tekstverwerker, waardoor het ook mogelijk is om de lengte van de antwoorden beter te reguleren (zowel minimale als maximale antwoordlengte) en de leesbaarheid voor de beoordelaar wordt gegarandeerd. Alle gegevens kunnen in dezelfde database worden opgeslagen, wat het uitvoeren van berekeningen en gecombineerde analyses van open en gesloten vragen gemakkelijker maakt. Doordat beroepssoftware kan worden ingezet, sluiten de toetsen beter aan bij de uiteindelijke beroepspraktijk van de studenten.

Digitaal toetsen heeft ook de mogelijkheid multimediaal materiaal (foto's, grafieken, geluid, video, animaties, simulaties) in een toets op te nemen. Domeinen als Aard- en Levenswetenschappen, Verpleegkunde en Medicijnen maken daar veel gebruik van. De zogenaamde casusvragen (zie ► hoofdstuk 13) maken vaak gebruik van vragen waar studenten stapsgewijs vragen over een patiënt moeten beantwoorden. Bij gesloten vragen zijn docenten niet beperkt tot juist-onjuist- of meerkeuzevragen, maar kunnen zij ook gebruikmaken van 'meerdere antwoorden correct'-vragen (multiple response), matching-vragen, ranking-vragen, drag-and-drop-vragen, numerieke waarde-vragen, voice response-vragen en dergelijke. Voor voorbeeldvragen zie ► <http://bit.ly/QM-example-test-Use-IE-to-view2>.

Digitaal toetsen maakt adaptief toetsen mogelijk. Bij adaptieve toetsen worden tijdens de afname vragen geselecteerd waarvan de moeilijkheid aansluit bij de vaardigheid van de student. Dit betekent dat de vraag die studenten krijgen afhankelijk is van hun antwoorden op eerdere vragen. Deze toetsvorm vereist geavanceerde psychometrische modellen (bijvoorbeeld Item Response Theorie), waardoor het voor individuele docenten of opleidingen praktisch niet mogelijk is om dit soort toetsen te construeren. Ondersteuning door gespecialiseerde toetsdeskundigen is nodig.

Internet maakt het mogelijk dat studenten in beginsel vanaf iedere willekeurige plaats en op ieder tijdstip een tentamen maken. In de praktijk blijken zich echter problemen voor te doen, zoals de controle over de personen en bronnen: zijn studenten wel degenen voor wie ze zich uitgeven? Hoe is te voorkómen dat studenten ongeoorloofde bronnen raadplegen? Hoe kan de toets technisch voldoende betrouwbaar worden aangeboden? Inmiddels zijn er nieuwe technologieën in ontwikkeling die thuistoetsing beveiligingstechnisch binnen bereik brengen, zoals het gebruik van specifiek ingestelde camera's en identiteitscon-

trole aan de hand van de snelheid en de kracht van het typen. De kosten per toets zijn nu (nog) te hoog om ze in het reguliere hoger onderwijs toe te kunnen passen. Benaderingen waarbij studenten hun eigen laptop of mobiele apparaat meebrengen naar de instelling zijn ook erg gevoelig gebleken voor problemen die te maken hebben met zowel het raadplegen van internetbestanden als de connectiviteit. De conclusie is dat digitaal toetsen in het hoger onderwijs niet zonder een grootschalige digitale toetsvoorziening kan, zoals in het tweede voorbeeld van ► paragraaf 9.1 is gepresenteerd.

### ■ Toetsinzage

Een specifiek aandachtspunt bij digitale summatieve toetsen is het inzagerecht voor studenten. Opleidingen hebben veelal in de regelingen opgenomen dat studenten recht hebben toetsen na afname en beoordeling in te zien. Bij een aantal toetssystemen is ingebouwd dat de feedback pas zichtbaar is na een bepaalde datum. Dat maakt dit inzageproces eenvoudiger. Er zijn echter ook situaties dat opleidingen hun vragen willen ‘beschermen’ zodat deze vaker kunnen worden gebruikt. In dat geval moet inzage ook onder gecontroleerde condities plaatsvinden. In de praktijk komt het er dan op neer dat studenten op een bepaald moment na afname van de toets onder toezicht van een docent of surveillant in de gelegenheid worden gesteld de gemaakte toets in te zien. Als toetsen worden getrokken uit een digitaal systeem waarbij feedback per vraag beschikbaar is, kan deze feedback worden getoond tijdens de inzage.

## 9.2.2 Formatieve toetsen afnemen met de computer

Bij het interactief en/of thuis afnemen van toetsen gaat het meestal om *formatieve* toetsen. Het afnemen van formatieve toetsen stelt minder hoge eisen aan de beveiliging of betrouwbaarheid van afname of gelijkwaardigheid van toetsen voor alle studenten. Bij deze toetsen is de leerwaarde (oefenfunctie, de bijsturing van het leerproces en de zelfbeoordeling) belangrijker dan de beoordeling en selectie (zie ► hoofdstuk 8). Samenwerking om toetsvragen te beantwoorden, het zogenaamde *collaborative testing* (Slusser & Erickson, 2006), kan dan positief werken. Deze formatieve toetsen kunnen ook deels voor bijvoorbeeld bonuspunten worden ingezet, maar dan worden de eisen ten aanzien van beveiliging en controle weer hoger en neemt de kans op oneigenlijk beantwoorden van toetsvragen toe, bijvoorbeeld doordat studenten toetsvragen gaan verzamelen en delen via social media.

Op dit moment zijn er echter veel goede mogelijkheden om met de computer formatieve toetsen af te nemen. Deze mogelijkheden liggen binnen het bereik van steeds meer docenten. Afhankelijk van het karakter van de toets worden andere eisen gesteld aan de opbouw en detaillering van de toets. Zie hiervoor het kader.

### Aanwijzingen voor het maken van interactieve diagnostische toetsen en vraaggebaseerd leermateriaal

Als de toets een diagnostische functie heeft:

- moet de inhoud van de toets representatief zijn voor het niveau van de eindtoets en de leerdoelen goed dekken;

- moet de feedback zowel een totaalscore als informatie bevatten over de beheersing van de verschillende onderdelen van de leerstof om antwoord te kunnen geven op de vraag aan welke onderdelen de student extra aandacht moet besteden;
- wordt de feedback gegeven direct nadat de formatieve toets geheel is gemaakt;
- moet de registratie zowel gegevens per student als over de hele groep bevatten; eventueel moeten resultaten kunnen worden weggeschreven naar het studentvolgsysteem van de opleiding.

Als de toets bedoeld is om te leren door vragen te beantwoorden:

- ontwerp dan vragen die eventueel gemakkelijker zijn dan vragen van een eindtoets (bijvoorbeeld leren toepassen van sleutelconcepten uit de leerstof) en maak de studenten duidelijk dat het niveau niet representatief is voor de eindtoets;
- maak dan vooral vragen over lastige onderdelen van de leerstof;
- maak dan een keuzemenu, zodat studenten kunnen kiezen voor vragen over bepaalde onderwerpen of moeilijkheidsgraad;
- maak dan meer vragen per onderwerp;
- plaats dan de vragen in oplopende volgorde van moeilijkheid, zodat studenten niet direct stuklopen;
- maak dan gebruik van een variatie aan vraagtypen en van multimediaal materiaal om de student extra te motiveren om met het materiaal aan de slag te gaan;
- voorzie dan de eerste opgaven van aanwijzingen of suggesties in welke richting de oplossing kan worden gezocht of verdeel de opgaven in kleine stappen die voorafgaan aan de uiteindelijke hoofdplossing.

(Naar: Keller, 1983; Van Berkel, 1999; Van Vliet, 2001.)

### 9.3 Itembanken

Goede toetsprogramma's (software) bieden de mogelijkheid om een gestructureerd bestand van vragen aan te leggen (itembank) waaruit toetsen worden samengesteld. Een itembank bestaat uit items (vragen en de bijbehorende antwoordmogelijkheden of een antwoordmodel) en informatie over die items, zoals het onderwerp waarbij de vraag hoort, het niveau van de vraag, wat voor type vraag het is, wanneer de vraag voor het laatst in een toets is opgenomen. De structuur en opbouw van de itembank zijn bepalend voor welke informatie er in kan worden opgenomen. Er zijn vele indelingen mogelijk. In de praktijk blijkt het (inhoudelijk) ontwerpen van een itembank moeilijk te zijn door de heterogeniteit en dynamiek van curricula, cursussen, toetsvormen en vraagvormen die moet worden ondersteund. De meest werkbare indeling lijkt die op basis van toetsmatrices (zie ► hoofdstuk 1). Immers, als de inhoudelijke criteria (curriculum, fase, cursus) op basis waarvan toetsen worden samengesteld bekend zijn, is het relatief eenvoudig de indeling van de itembank te bepalen.

Wanneer eenmaal een vragenbestand beschikbaar is, is het samenstellen van een (formatieve of summatieve) toets mogelijk. De toetssamensteller, meestal de verantwoordelij-

ke docent, maakt dan een toets, conform de toetsmatrijs, bestaande uit een vooraf bepaald aantal vragen, met een bepaalde verdeling over onderwerpen en taxonomische niveaus. Daarnaast bepaalt de toetsmaker de exacte instellingen van de toets voor de student: wanneer is de toets beschikbaar, hoelang is deze beschikbaar, hoe vaak mogen de studenten de toets maken, geeft de toets automatisch inhoudelijke feedback of niet, hoe ziet de navigatie en lay-out eruit, en dergelijke. Daarnaast dient er op de instelling een vraagloket te zijn waar studenten terecht kunnen als er technische problemen zijn, zoals het per ongeluk afbreken van het toetsprogramma.

De computer maakt het mogelijk om bij het samenstellen van toetsen vragen en antwoordalternatieven in willekeurige volgorde aan te bieden, waardoor studenten verschillende versies van dezelfde toets krijgen. Hierdoor neemt de kans op fraude of samenwerken af en kunnen studenten op verschillende tijdstippen toch dezelfde toets maken. Het veranderen van de volgorde van de vragen en binnen antwoordalternatieven moet echter weloverwogen gebeuren. De volgorde in een toets is bij voorkeur oplopend in moeilijkheidsgraad of geordend op thema. Het is dan niet wenselijk dat de volgorde willekeurig wordt. Ook moet worden voorkomen dat alternatieven die gebaseerd zijn op een logische volgorde, bijvoorbeeld op basis van chronologie of processtappen, door elkaar aangeboden worden.

Meer en meer brengen uitgeverijen digitale lesmethoden op de markt waarbij digitale vragenbanken zijn inbegrepen. Die vragenbanken zijn vaak eenvoudig te downloaden en hergebruiken door docenten. Daarnaast wordt het ook steeds eenvoudiger om op basis van *interoperability*-standaarden (vooral de opkomende standaard Basic LTI) toetsen af te nemen via de *cloud*. Denk hierbij weer aan uitgeverijen die op hun eigen servers online toetsen beschikbaar stellen. Via een gestandaardiseerde koppeling kunnen docenten in hun leeromgeving (bijvoorbeeld Blackboard of Moodle) toetsen van uitgeverijen aan studenten aanbieden via één link, waarbij de resultaten automatisch worden opgeslagen in de eigen leeromgeving.

Overigens leidt een bestand met goede vragen niet automatisch tot goede toetsen. Het is van belang om de selectie van vragen voor een toets zorgvuldig uit te voeren (zie ► hoofdstuk 1) en om de onderlinge afhankelijkheid van de vragen, als ze in één toets staan, te controleren. Zeker als het bestand parallelvragen bevat (verschillende vragen over hetzelfde onderwerp) of casusvragen (een vraag bestaande uit een stam met daarbij deelvragen over verschillende onderwerpen), bestaat de kans dat er in een automatisch gegenereerde toets van elkaar afhankelijke vragen of antwoorden aanwezig zijn (een vraag geeft bijvoorbeeld het antwoord op een andere vraag). De moeilijkheid van de verschillende unieke toetsen die door de computer gegenereerd zijn samengesteld, kan verschillen. Dat maakt analyse en het bepalen van een goed onderbouwde cesuur lastiger.

## 9.4 Nakijken van open vragen en werkstukken

---

Er bestaan geen eenvoudig te gebruiken computerprogramma's die automatisch open vragen of werkstukken na kunnen kijken. De formulering van het correcte antwoord is vaak veelvormig en in sommige gevallen zijn er verschillende correcte antwoorden mogelijk.

Toch wordt de computer ook bij het nakijken van dit type vragen steeds belangrijker. Het grote voordeel van het digitaal nakijken van open vragen is de leesbaarheid van de antwoorden. Daarnaast wordt het organisatorisch eenvoudiger om bij het nakijken van open vragen dubbele correctie uit te voeren. Voor de dubbele correctie kunnen verschillende werkwijzen worden gehanteerd:

- De toetsen worden (willekeurig) verdeeld over verschillende correctoren, waarbij de correctoren allen het volledige tentamen nakijken.
- De vragen worden verdeeld over de correctoren. Elke corrector ziet een deel van het werk van elke student.
- De vragen worden dubbel nagekeken. Elke corrector kijkt de vragen van elke student na en daarna worden de scores met elkaar vergeleken.

De computer ondersteunt het proces bij deze werkwijzen. Daarnaast is het mogelijk om bij het digitaal nakijken ervoor te kiezen dat de namen van de studenten niet zichtbaar zijn. Hierdoor worden beoordelaareffecten, zoals het halo- of horn-effect (zie de bijlage achter in dit boek), beperkt. Een andere mogelijkheid om dit effect te beperken is een keuze tussen horizontaal (eerst van alle studenten het eerste item, dan van alle studenten het tweede item, enzovoort) of verticaal (eerst alle vragen van student 1, dan student 2, enzovoort) nakijken.

Voor werkstukken biedt het softwareprogramma *Turnitin* mogelijkheden om geschreven werkstukken van studenten digitaal te beheren. Vervolgens kan elk werkstuk digitaal van feedback worden voorzien door middel van tekstannotaties, audiofeedback en/of een beoordelingsrubric (rubric). Docenten dienen dan een bibliotheek aan te leggen met veelvoorkomende 'feedbackfragmenten' en deze te categoriseren in soorten feedback (zie ook ► hoofdstuk 17).

Deze digitalisering van het nakijken biedt, naast voordelen op het gebied van plagiaatcontrole, zowel op logistiek als inhoudelijk vlak mogelijkheden tot kwaliteitsverbetering:

- Het inleveren, downloaden, annoteren en verspreiden van werkstukken wordt voor de docent een stuk makkelijker en efficiënter. Alles gaat naar een centrale plek en wordt daar nagekeken, en de student haalt de feedback er zelf uit. De docent kan nagaan of studenten hun feedback al dan niet hebben geopend. Docenten bij de VU die *Turnitin* gebruiken, geven aan dat ze nauwelijks meer werkstukken uitprinten. Het programma archiveert de gegeven feedback en de student kan dit op elk gewenst tijdstip raadplegen. Ook voor de externe accreditatie vormen de digitale beschikbaarheid van feedback en de langere bewaartermijn een wenselijke ontwikkeling.
- Met het digitaliseren van nakijken is het mogelijk de feedback *sneller* aan studenten te doen toekomen. Dat is van belang omdat (te) laat gegeven feedback het beoogde effect doet verdwijnen. De inspanning van de docent zou dan voor niets zijn geweest (Gibbs, 2010).

De docent kan gaandeweg een inhoudelijk rijke pool aan *feedbackteksten* ontwikkelen die zich enerzijds richten op de inhoud van een werkstuk en anderzijds op specifiekere vaardigheden, bijvoorbeeld schrijf- of informatievaardigheden. Een van de mogelijkheden is dat docenten in een opleiding gezamenlijk deze fragmenten bundelen. Dit vormt dan de

aanzet voor een bredere aanpak van de inzet van feedback binnen een opleiding, bijvoorbeeld door het ontwikkelen van uniformere beoordelingscriteria.

Daarnaast hebben de meeste 'gewone' digitale leeromgevingen (als Blackboard en Moodle) de laatste jaren hun functionaliteit uitgebreid met de mogelijkheid om *rubrics* (zie ► hoofdstuk 17) te gebruiken. Ook kan de leeromgeving worden gekoppeld aan een externe rubric-applicatie als *Waypoint*. In *Turnitin* kunnen docenten een zelf ontwikkelde rubric koppelen aan een ingeleverd werkstuk en hiermee een beoordeling geven op verschillende opdrachtonderdelen van een werkstuk. De feedback kan kwalitatief en/of kwantitatief zijn. In het laatste geval wordt op basis van de rubric een cijfer berekend. Op deze wijze geeft de docent formatieve, inhoudelijke feedback en een summatieve beoordeling. Eenzelfde rubric kan in verschillende fases van de opleiding worden gebruikt om de voortgang van de student in kaart te brengen, bijvoorbeeld op het vlak van schrijfvaardigheden. Een meer longitudinalere inzet van een rubric vraagt in dat geval dan ook om bredere afstemming binnen een opleiding over het ontwerp en wijze van inzet van het instrument.

De verwachting is dat het digitaal nakijken van werkstukken de komende jaren in toenemende mate onderdeel zal gaan worden van de gewone werkpraktijk van de docent.

## 9.5 Echt andere toetsvormen: de toekomst?

---

Er worden brede horizonten geschetst voor digitaal toetsen. Eigenlijk is er geen enkel terrein meer denkbaar waar ICT geen rol speelt, ook niet bij toetsing. ICT kan altijd een rol spelen om het proces administratief te ondersteunen, maar ICT maakt ook nieuwe vormen van toetsen en taken mogelijk. Zo kan ook het digitaal schrijven en beoordelen van blogs of wiki-pagina's, het samen produceren van geschreven werkstukken, het spelen van digitale games, *inquiry based learning* door het werken met simulaties of het werken in virtuele werelden worden geschaard onder de noemer digitaal toetsen. Nog nieuwere ontwikkelingen gaan uit van online behaalde badges die aantonen dat een lerende zich in een digitale sociale context profijtelijk heeft gemaakt of met een bepaalde mate van expertise door zijn of haar peers is beoordeeld. Het voert te ver om dat in het kader van dit hoofdstuk verder uit te werken; verwezen wordt naar het werk van Crisp (2007) en diverse websites rondom badges, bijvoorbeeld ► [http://en.wikipedia.org/wiki/Digital\\_badges](http://en.wikipedia.org/wiki/Digital_badges). De waarde van deze nieuwe vormen van opdrachten en manieren van waardering dienen zich echter nog te bewijzen.

# Creatieve toetsvormen

*Titus Geerligs, Henk Schmidt, Irma Kokx, Erik de Graaf en Henk van Berkel*

## Inleiding

Waarom creatief toetsen?

Wat zijn creatieve toetsvormen?

Welke creatieve toetsvormen zijn creëeropdrachten?

Welke creatieve toetsvormen vallen onder presentatieopdrachten?

Welke creatieve toetsvormen zijn rollenspelen?

Welke creatieve toetsvormen zijn studieopdrachten?

Welke creatieve variaties zijn er op de schriftelijke toetsen?

## 10.1 Creatieve toetsvormen – 102

10.1.1 Creëeropdrachten – 102

10.1.2 Presentaties – 105

10.1.3 Rapporteren – 106

10.1.4 Rollenspelen – 107

10.1.5 Studieopdrachten – 109

10.1.6 Variaties in schriftelijke toetsen – 109

## 10.2 Ten slotte – 111

Toetsen in het onderwijs zijn doorgaans sterk gericht op het kunnen reproduceren van kennis. Deze benadering van toetsen heeft tot gevolg dat studenten het accent leggen op uit het hoofd leren, waardoor het toepassen van deze kennis in het gedrang komt. Zodra toepassing van die kennis wordt verlangd zullen studenten hun leerinspanning daar op gaan richten, want zoals bekend: de toets stuurt het studeergedrag. Het is gewenst dat studenten in het hoger onderwijs met een grote variëteit aan toetsvormen in aanraking komen waarin enerzijds een beroep wordt gedaan op kennis en vaardigheden sec en anderzijds op integratie ervan en het creatief kunnen omgaan met opgedane kennis. Docenten moeten dus creatiever toetsen.

## 10.1 Creatieve toetsvormen

---

In deze paragraaf komen verschillende toetsvormen aan de orde die een beroep doen op creatief en kritisch denken. De bruikbaarheid ervan is wisselend en is bijvoorbeeld mede afhankelijk van het aantal studenten dat eraan deelneemt, maar vooral van de doelen. De opbouw is in dit hoofdstuk steeds gelijk: eerst een schets, een korte uitleg en het doel waarvoor de toetsvariant kan worden gebruikt en vervolgens mogelijke beoordelingscriteria.

De genoemde toetsvormen in dit hoofdstuk zijn ongestructureerd. Het is niet mogelijk concrete richtlijnen te geven ten aanzien van de vorm in praktijk en de beoordelingscriteria. Daarvoor zijn ze te afhankelijk van de situatie in de opleiding en de doelen van het onderwijs. Met andere woorden, indien een docent een van de vormen wil toepassen in het eigen onderwijs, zal de docent zelf creatief moeten zijn.

### 10.1.1 Creëeropdrachten

---

#### ■ Concept mapping

Een conceptmap bevat, in de simpelste vorm, allerlei begrippen die in het onderwijs aan de orde zijn gekomen met lijnen ertussen. De lijnen symboliseren een relatie tussen die begrippen. Hoewel een conceptmap op de eerste plaats een middel is om studenten te leren dat de behandelde begrippen een bepaalde relatie hebben, is het in feite dus een onderwijsmiddel en kan het ook een toetsvorm zijn. Studenten krijgen dan een lijstje met begrippen en hen wordt gevraagd een tekening te maken met hun onderlinge relaties.

De beoordeling van het schema kan op de eerste plaats geschieden door het tegen het licht te houden van een ideaal schema. Dat kan echter problematisch zijn omdat er niet altijd een ideaal schema bestaat. Het is mogelijk dat begrippen, vanuit een bepaalde optiek, ook anders kunnen worden gerangschikt. Daarom is het raadzaam studenten bij het schema te vragen naar een nadere toelichting bij de relaties.

#### ■ Excursie/werkbezoek organiseren

Door het organiseren van een excursie of werkbezoek leren studenten niet alleen organiseren, maar ze leren ook de onderwerpen en vragen te kiezen die belangrijk zijn in de



opleiding die ze volgen. Veelal zal de organisatie van een excursie of werkbezoek worden uitgevoerd door een groepje studenten, onder supervisie van een docent. Dit laatste is belangrijk omdat een opleiding het zich niet kan veroorloven dat er ‘veldbederf’ plaatsvindt. Voor het welslagen van het werkbezoek is het van belang dat de deelnemers weet hebben van de bedoelingen zodat ze gericht kunnen deelnemen. Het ligt meer voor de hand eventueel schriftelijk informatiemateriaal enkele dagen van tevoren uit te delen dan op het werkbezoek zelf. Een excursie richt zich veelal op leerdoelen waarbij toepassing van opgedane kennis aan de orde is.

De beoordeling van de leeropbrengst van een excursie kan globaal gebeuren, bijvoorbeeld met een dichotome classificatie wel/niet voldaan aan de eisen, eventueel na een schriftelijke of mondelinge toelichting door de studenten die het werkbezoek hebben voorbereid.

#### ■ Interviewen

Een interview is geschikt als van studenten wordt verwacht dat ze in staat zijn op een systematische wijze mondeling informatie te verzamelen bij personen. Bij een interview stelt de interviewer mondeling vragen over bepaalde onderwerpen of problemen aan de geïnterviewde (dit kunnen er ook twee of drie zijn, niet meer). Meestal is de plaats van het interview de werkplek van de geïnterviewde, maar er zijn ook telefonische interviews of straatinterviews. Doel van het interview is het gericht verzamelen van informatie. De informatie kan betrekking hebben op kennis, opinies, gevoelens of ervaringen van de geïnterviewde. De interviewer kan gebruikmaken van een vooraf opgestelde vragenlijst. Dit is een vorm van een gestandaardiseerd interview. De resultaten van het interview kunnen schriftelijk worden genoteerd en/of op een geluidsband worden opgenomen. Achteraf notuleren op basis van het geheugen is af te raden.

Het vrije interview is een vorm zonder vragenlijst waarbij de ondervraagde wordt aangemoedigd zich uit te spreken over een door de interviewer aangedragen onderwerp. De interviewer spant zich in om het gesprek zodanig te laten verlopen dat de geïnterviewde de gewenste informatie verschaft. Een vrij interview is dus wat anders dan een gesprek waarbij het gaat om het uitwisselen van meningen, gevoelens of ervaringen tussen twee of meer personen.

Een zoekopdracht via Google (‘interviewtips’) levert talloze sites op met praktische tips voor het houden van een interview. Hieruit zijn beoordelingscriteria af te leiden.

#### ■ Peer review

Peer review is een toetsvorm waar studenten kritiek geven op een wetenschappelijk paper van een medestudent. Het doel van deze toetsvorm is na te gaan of studenten in staat zijn een paper kritisch te kunnen lezen en beoordelen. Student nemen de rol aan van reviewers van een tijdschrift. Zij moeten de richtlijnen van een nader te kiezen tijdschrift hanteren en een advies schrijven aan de editor. Een variant is dat reviewers direct feedback geven aan de schrijver.

Docenten beoordelen de review op basis van criteria die gebruikelijk zijn in de discipline.

### ■ Vragenlijst opstellen

De vragenlijst is een veelvuldig toegepast onderzoeksinstrument. Studenten leren door het construeren van een vragenlijst een onderzoeksvraag te operationaliseren. In het algemeen bevatten vragenlijsten drie soorten vragen: open vragen, gesloten vragen en schalen die bijvoorbeeld lopen van 1 tot 5 (de zogenaamde *Likert*-schaal). Het hoeft geen betoog dat open vragen meer verwerkingstijd vergen dan de andere twee. Het doel van de vragenlijst bepaalt de keuze uit de verschillende vraagsoorten.

De kwaliteit van een vragenlijst kan aan de hand van de volgende criteria voor een goede vragenlijst worden bepaald:

- de invulinstructie is duidelijk;
- de formuleringen zijn kort en begrijpelijk en bevatten geen vage begrippen;
- de formuleringen zijn aangepast aan de ‘denkwereld’ van de invullers;
- de formuleringen zijn eenduidig;
- de vragen zijn niet suggestief;
- een enkele vraag bevat niet meer dan één onderwerp;
- de vragenlijst bevat geen overbodige vragen;
- de vragenlijst bevat geen vragen waarop het antwoord zelf is te achterhalen;
- de vragen staan in een logische volgorde;
- de vragenlijst bevat geen vragen waarop de invuller sociaal wenselijke antwoorden kan geven;
- wanneer de vragenlijst uit verschillende vraagsoorten bestaat, dient steeds een nieuwe invulinstructie te worden ingevoegd;
- de vragenlijst heeft een overzichtelijke lay-out;
- iedere vraag is gerelateerd aan een vraagstelling van het onderzoek;
- de vragenlijst is uitgetoetst.

### ■ Toets construeren

Door studenten een toets te laten construeren leren zij hoofd- en bijzaken van elkaar te scheiden. De opdracht (vragen én modelantwoorden) kan individueel of in groepjes worden uitgevoerd. De onderwerpen waar de vragen over moeten gaan worden doorgaans gehaald uit het onderwijs dat studenten hebben gevolgd, of uit aangereikte artikelen of boeken. Het verdient geen aanbeveling studenten zelf een onderwerp te laten kiezen, omdat veel Amerikaanse studieboeken toetsvragen achter ieder hoofdstuk bevatten. Studenten komen dat snel te weten en zullen om die reden bepaalde onderwerpen kiezen.

Het is noodzakelijk een onder- en bovengrens aan het aantal te construeren vragen te stellen. Het laten construeren van gesloten vragen verdient de voorkeur, omdat de kwaliteit van deze vragen beter te beoordelen is dan die van open vragen en omdat het construeren van alternatieven een extra leeropbrengst geeft. Maar noodzakelijk is dat niet. Als studenten open vragen maken plus het antwoordmodel, krijgt een docent ook goed zicht op de leeropbrengst.

In ► hoofdstukken 11 en 12 staan checklists aan de hand waarvan de vragen kunnen worden beoordeeld.

## 10.1.2 Presentaties

---

### ■ Audiovisuele presentatie

De laatste jaren heeft het gebruik van audiovisuele producten een hoge vlucht genomen. Nog niet zo lang geleden waren dia- en videopresentaties en het hanteren van een overheadprojector (met transparanten!) het modernste van het modernste, tegenwoordig zijn daar talloze en aantrekkelijkere producten bijgekomen. Het is niet moeilijk meer de voor de presentatie relevante audiovisuele informatie van het internet te downloaden. Animaties, film- en geluidsfragmenten, delen van hoorcolleges, foto's, noem maar op. Dit alles is mogelijk via en in een PowerPoint presentatie of *Prezi*. Bij sommige presentaties denk je weleens 'hou het simpel'; ze lijken dan op een opgetuigde kerstboom en verbloemen vaak de boodschap, als die er is. Het devies aan de studenten is dat audiovisuele materialen functioneel moeten zijn voor de inhoud. Van het houden van een presentatie leren studenten onder andere hoofdzaken van bijzaken te scheiden en de juiste hulpmiddelen bij een betoog te verkrijgen en in te bouwen.

Bij de beoordeling kan worden gelet op de technische kwaliteit van beeld en geluid, de functionaliteit en het feit of beeld en geluid een samenhangend geheel vormen.

### ■ Poster

Op de meeste (wetenschappelijke) congressen is plaats ingeruimd voor postervoordrachten waarbij de maker van de poster een korte toelichting geeft. Net als bij het maken van een PowerPoint presentatie dienen studenten zich bij het maken van een poster op de essentie te richten. Op een poster moet alle informatie komen te staan waarin de lezer is in geïnteresseerd. Dat is lastig, maar kan goed in een onderwijssetting worden geoefend.

Een goede poster voldoet aan de volgende criteria:

- niet meer dan 500 woorden
- een duidelijke indeling
- ongeveer één vierkante meter
- letters niet kleiner dan één centimeter

Verder kan worden gelet op de vlakverdeling, lijnvoering, functioneel kleurgebruik, enzovoort. Uiteraard is ook de inhoud van belang. Vaak bevat de poster een zogenaamde conceptmap (zie de betreffende alinea in dit hoofdstuk) die als zodanig kan worden beoordeeld.

### ■ Voordracht

De leeropbrengst van het houden van een voordracht is hoog. Studenten moeten zich verdiepen in het onderwerp. Ze moeten hoofd- en bijzaken onderscheiden en de inhoud laten aansluiten op de voorkennis en interesse van de toehoorders.

Afhankelijk van het belang en het doel dat een beoordelaar toekent aan de presentaties, is onderstaande lijst bruikbaar.

- onderscheidbare delen zijn te herkennen
- aangeven van het doel
- uiteenzetten van de grote lijn

- aansluiting op de voorkennis van het gehoor
- belichten van de relevantie van het onderwerp
- onderscheid maken tussen hoofd- en bijzaken
- geven van voorbeelden en andere verduidelijkingen
- tussentijds recapitulieren
- helder formuleren
- spreeknelheid
- afwisselend stemgebruik
- contact met het gehoor
- hanteren van gebaren
- het geven van gelegenheid voor vragen
- beantwoording vragen
- enthousiasme

### 10.1.3 Rapporteren

---

#### ■ Specificatie literatuurstudie

Studenten kan worden gevraagd nauwkeurig aan te geven welke literatuur ze in het kader van het onderwijs hebben gelezen. Zij moeten daarbij bovendien aangegeven wat naar hun idee de kwaliteit is van de teksten die ze hebben bestudeerd en de relevantie voor het onderwijs. Studenten dienen dus door deze toetsvorm te tonen dat ze kritisch teksten lezen en de relevantie ervan kunnen beoordelen. Als hulpmiddel hierbij kan de docent de beoogde doelen van het onderwijs ter beschikking stellen en studenten vragen de link aan te geven tussen de gelezen literatuur en de doelen, waarbij studenten ook mogen aangeven welke doelen niet kunnen worden beantwoord met behulp van de opgegeven literatuur. De leeropbrengst van het maken van een dergelijk verslag is waarschijnlijk hoog, omdat studenten worden gedwongen na te denken over de literatuur en die te plaatsen tegen de achtergrond van de doelen. Zij kunnen niet volstaan met informatie die ze van het internet hebben gehaald.

De docent kan het verslag nakijken aan de hand van inhoudelijke criteria.

#### ■ Taakverslag

Het maken van een taakverslag leert studenten te reflecteren op het door hen gevolgde onderwijs. Bij het maken wordt studenten gevraagd individueel op te schrijven welke gedachtegang zij hebben gehad bij het behandelen van een onderwerp in het onderwijs. Ze moeten ook opschrijven welke verwachtingen ze hadden en of het onderwijs aan die verwachting heeft voldaan. Zouden zij andere literatuur hebben gebruikt? En welke? Een evaluatie van het geheel completeren de afsluiting van het maken van een taakverslag.

De beoordeling door de docent kan plaatsvinden door het verslag te beoordelen op consistentie, diepgang en volledigheid.

### ■ Tijdbestedingspecificatie

Door het bijhouden van hun studietijd leren studenten te reflecteren op het gevolgde onderwijs. In het kader van onderwijs voeren studenten allerhande activiteiten uit: lesvoorbereiding, bijwonen van onderwijs (colleges, werkgroepen, practica), literatuurstudie, toetsvoorbereiding enzovoort. Studenten kan worden gevraagd aan de hand van een formulier hun tijdsbesteding aan de verschillende activiteiten te noteren en hierover een verslag te maken dat vragen moet beantwoorden als: 'Was de tijdsbesteding in overeenstemming met wat je had gewild, in absolute zin en in relatieve zin? Was het genoeg? Was het doelmatig, oftewel was de verhouding inspanning-opbrengst goed? Ga je het in de toekomst weer zo doen? Waarom wel, waarom niet?'

De docenten beoordelen vervolgens op criteria die zijzelf belangrijk vinden.

## 10.1.4 Rollenspelen

---

### ■ Consensusconferentie

Een deelnemer aan een consensusconferentie leert de eigen mening te spiegelen aan die van anderen én te proberen die verschillende meningen samen te voegen tot één gezamenlijke mening. Kenmerkend voor een consensusconferentie is dat alle aanwezigen (studenten) hun bijdrage kunnen leveren bij het opstellen van de aanbevelingen en dat de conclusies door alle betrokkenen worden onderschreven. Het gaat erom de diverse meningen met elkaar in overeenstemming te brengen of eventueel duidelijk te maken op welke onderdelen de meningen uit elkaar blijven lopen. Er is een belangrijk verschil tussen consensusconferenties en wetenschappelijke congressen of nascholingscursussen. Staan bij de laatstgenoemde mededelingen van wetenschappelijk onderzoek en kennisoverdracht centraal, bij een consensusbijeenkomst is het de bedoeling om tot een gemeenschappelijk mening te komen. Gewoonlijk wordt de tekst van de consensus opgesteld door een panel van 'neutrale' deskundigen waarin, afhankelijk van de aard van het onderwerp, bijvoorbeeld ook juristen zitting kunnen hebben. Het panel moet antwoord geven op concrete vragen die door het organiserend comité zijn opgesteld. Het panel laat zich voorlichten door verschillende deskundigen in het bijzijn van het publiek. Vanuit de zaal ontstaat discussie met de inleiders, maar de panelleden kunnen om nadere toelichting vragen. Het panel stelt vervolgens een conceptconsensusverklaring op. De definitieve tekst wordt opgemaakt na uitgebreide discussie tussen de panelleden en aanwezigen. Is consensus niet haalbaar, dan bestaat de mogelijkheid tot het uitbrengen van een minderheidsstandpunt.

Er zijn verschillende methoden om ieders rol te beoordelen. Wat precies is afhankelijk van wat de docent belangrijk vindt. Op de eerste plaats dienen studenten hun eigen mening goed te kunnen verwoorden. Zij moeten kunnen luisteren naar andere argumenten en, wellicht, hun eigen mening daardoor veranderen.

### ■ Forumdiscussie

Met behulp van een forumdiscussie kunnen studenten laten zien of ze kunnen discussiëren. Een forumdiscussie als rollenspel kan er als volgt uitzien. 'Deskundigen' zitten voor de groep en worden voorgesteld door de voorzitter. De voorzitter leidt in en de forumleden

discussiëren vervolgens over een vooraf opgegeven onderwerp. Op aanwijzing van de voorzitter kunnen toehoorders vragen stellen.

De beoordeling van de bijdragen van de forumleden aan de discussie kan door de toehoorders plaatsvinden, bijvoorbeeld door middel van het geven van cijfers. Het is raadzaam dat de docent de aspecten bepaalt waarop de oordelen betrekking moeten hebben.

#### ■ Kort geding

In een kort geding leren studenten te pleiten voor of tegen een bepaalde stelling. Het kort geding leent zich voor toepassing bij onderwerpen waarover geen consensus bestaat. Bij een kort geding verschillen twee partijen van mening over een stelling. Elke partij bestaat uit een advocaat, of pleiter, en enkele getuigen-deskundigen. Daarnaast is er een onafhankelijk voorzitter die de regie heeft over het kort geding. Zij allen dienen hun bijdrage goed voor te bereiden. Partijen en voorzitter zitten vooraan in een ruimte. Bij de aanvang van het kort geding stemmen de toeschouwers (de 'jury') schriftelijk over de stelling. De voorzitter maakt de uitslag *niet* bekend. Het is belangrijk dat de stelling zodanig is geformuleerd dat er goede argumenten voor en tegen kunnen worden aangevoerd.

Beide advocaten ondervragen bij toerbeurt de getuigen. De voorzitter houdt de tijd in de gaten. Na het horen van de getuigen vinden de slotpleidooien plaats. Vervolgens stemmen de toeschouwers wederom en de voorzitter maakt de uitslag van beide stemmingen bekend.

De beoordeling van de deelnemers (advocaten, getuigen en voorzitter) kan door de toeschouwers gebeuren waarbij ze vooral letten op de kwaliteit van de argumentaties, bijvoorbeeld helderheid, formulering, bondigheid, (on)verantwoorde generalisaties, ter zake doende.

#### ■ Rollenspel toneelstuk

Het rollenspel is een toneelstukje zonder een expliciet uitgeschreven tekst waar alleen de personages en hun achtergronden zijn gegeven. Hoe het spel zich ontwikkelt, is sterk afhankelijk van de wijze waarop de spelers hun rol vervullen. Meestal gaat het in het rollenspel om een nabootsing van een situatie waarin personen problemen met elkaar hebben, bijvoorbeeld een groepsconflict.

De voorbereiding en nabespreking van het rollenspel moeten grondig zijn. Tijdens de voorbereiding vindt de 'warming up' plaats en worden de rollen van deelnemers en toeschouwers toegelicht. De studenten die als toeschouwers/beoordelaars het spel meemaken, moeten van de spelleider de nodige informatie verkrijgen over de te spelen situatie en de personages.

Tijdens de evaluatie (beoordeling) wordt het spelverloop geanalyseerd en de belevingen en leerervaringen van spelers en toeschouwers besproken. Er kan bijvoorbeeld worden nagegaan op welke punten toeschouwers anders gereageerd zouden hebben of waarin ze zich in de speler hebben herkend. De evaluatie verloopt gemakkelijker wanneer het spel op video is opgenomen. De beoordelingscriteria zijn afhankelijk van het doel van het rollenspel.

### ■ Simulatie

Bij simulatie gaat het om het nabootsen van een (vereenvoudigde) werkelijkheid, bijvoorbeeld het simuleren van een managementteam in een bedrijf. Simulatie is vooral geschikt voor het evalueren van situaties waarin beslissingsprocessen een cruciale rol spelen en voor het leren en evalueren van vaardigheden in het omgaan met mensen. Rollen en spelregels beperken de keuzevrijheid van de spelers, studenten. Van belang is dat het werkelijkheidsgehalte van simulaties zo groot mogelijk is.

De beoordeling van simulaties kan het beste geschieden aan de hand van beoordelingslijsten, die gezien de verscheidenheid aan onderwerpen voor iedere simulatie apart moeten worden ontwikkeld.

## 10.1.5 Studieopdrachten

---

### ■ Hink-stap-sprong (*triple-jump test*)

De hink-stap-sprong is oorspronkelijk bedoeld om te toetsen of studenten zich de methodiek hebben eigen gemaakt die van hen wordt verwacht in een probleemgestuurd onderwijscurriculum (PGO). Het is een mondelinge toets bestaande uit drie stappen. In stap 1 legt de docent een probleem, vraagstelling, gevals- of situatiebeschrijving voor die betrekking heeft op het voorafgaande onderwijs en die in zekere mate representatief is voor de latere beroepsuitoefening. Studenten gaan in de toetssituatie hardop denken hoe ze een dergelijk probleem zouden aanpakken, stellen hypothesen op en passen hun dan aanwezige kennis toe om tot een oplossing te komen. De taak van de docent is om studenten via doorvragen te laten formuleren wat ze nog niet kunnen verklaren en/of welke informatie ontbreekt. Deze discussie eindigt met het formuleren van vragen voor verdere studie. In de tweede stap gaan studenten op zoek naar aanvullende informatie om de vragen te kunnen beantwoorden. Zij bezoeken bijvoorbeeld de bibliotheek, raadplegen internet en/of lezen kranten. Zij zullen doelmatig moeten zoeken, omdat de docent de opzoektijd heeft gelimiteerd. Nadat die tijd is verstreken, komt de student in stap 3 terug bij de docent met een synthese van het probleem en met antwoorden. Dan volgt de beoordeling waarbij de volgende aspecten van belang zijn:

- het toepassen van de aanwezige kennis op de eerste probleemanalyse;
- het denkproces (hypotheseformulering, het verwerken van de nieuwe informatie, het trekken van conclusies);
- het proces van zelfstudie;
- de eigen beoordeling.

## 10.1.6 Variaties in schriftelijke toetsen

---

### ■ Probleemvragen

Een probleemvraag beoogt een situatie in de toekomstige beroepsuitoefening te weer spiegelen. In de beschrijving van een probleem, de casus, staat een herkenbaar probleem dat studenten moeten oplossen. Door middel van enkele korte vragen is het mogelijk het

denkproces van studenten te sturen. Bij de beoordeling van probleemvragen is het van belang tevens een antwoordmodel op te stellen dat de volgende elementen bevat:

- een weergave van goede, gedeeltelijk goede en, eventueel, pertinent foute antwoorden;
- een puntentoeckenning per onderdeel.

De probleemvraagtoets lijkt op de korte casustoets (► hoofdstuk 13).

#### ■ Open boektoets

Bij een open boektoets mogen studenten boeken, artikelen en aantekeningen, of een specifieke selectie daaruit, meenemen en raadplegen. Vragen die een beroep doen op parate kennis zijn voor deze toetsvorm ongeschikt omdat ze simpelweg kunnen worden opgezocht. Overigens, dit soort toetsvragen kunnen eventueel wel worden gehanteerd wanneer er veel van in de toets zijn opgenomen en studenten binnen een korte periode de vragen moeten beantwoorden. De toets doet dan, gezien het *speed* karakter, een beroep op de opzoekvaardigheden. Probleemvragen en gestructureerde open vragen zijn in het bijzonder geschikt voor de open boektoets, omdat ze een beroepssituatie weerspiegelen en in de beroepssituatie is het raadplegen van bronnen in voorkomende gevallen juist aan te raden.

#### ■ Schriftelijke opdrachten

Een schriftelijke opdracht verschilt van een schriftelijke toets. Bij een toets denkt men al gauw aan vragen en antwoorden, terwijl een opdracht meer te vergelijken is met een taak die een student moet uitvoeren en die na het gereedkomen door een docent wordt beoordeeld. Er zijn vele opdrachten te bedenken, van zwaar tot licht:

- herschrijven van een ‘moeilijk’ artikel in gemakkelijk leesbaar Nederlands;
- verbeteren van onjuiste definities;
- een artikel beoordelen op de methodologische kwaliteit;
- tekortkomingen opsporen in een probleemstelling en verbeteren;
- redeneerfouten in een betoog benoemen;
- in een tekst meningen, interpretaties en resultaten uit elkaar halen;
- formuleren van geldige gevolgtrekkingen uit een aangereikte tekst;
- formuleren van aanbevelingen bij een nota;
- formuleren van verschillende scenario’s om hetzelfde doel te bereiken;
- uitschrijven van een trainingsprogramma;
- opstellen van brieven aan diverse instanties of personen;
- schrijven van een ingezonden brief aan krant of tijdschrift;
- formuleren van procedurevoorstellen voor de afhandeling van rapporten of beleidsvoorstellen;
- interpreteren van gegeven feiten in het licht van verschillende theorieën;
- beschrijven van de mate van overeenkomst tussen een gegeven model en werkelijkheid;
- beredeneren van de keuze voor een bepaalde wijze van steekproeftrekking bij een gegeven onderzoeksvraagstelling;



- kritisch vergelijken van verschillende onderzoekontwerpen in relatie tot een geformuleerde onderzoeksvraagstelling en een eigen voorkeur formuleren en beargumen-teren;
- de interne validiteit van een onderzoek becommentariëren;
- de generalisatiereikwijdte van onderzoeksresultaten aangeven;
- de theoretische relevantie van gegeven onderzoeksresultaten bepalen;
- toepassingsmogelijkheden van een bepaalde theorie verwoorden;
- beredeneren van de maatschappelijke relevantie van verschillende onderzoeksresultaten;
- kritisch bespreken van een onderzoeksverslag ten aanzien van vorm en inhoud;
- het schrijven van een boekbespreking aan de hand van een leidraad;
- schrijven van een onderzoeksontwerp;
- een gegeven stelling beargumenteert verdedigen of aanvallen;
- schrijven van een folder op basis van onderzoeksresultaten;
- het maken van slides voor een PowerPoint presentatie op basis van een wetenschappelijk artikel.

De beoordelingscriteria voor bovenstaande opdrachten kunnen in eerste instantie worden gehaald uit de beoordelingscriteria voor werkstukken (► hoofdstuk 23). De nabespreking met de student over de producten kan of individueel plaatsvinden, of in groepjes. Dat laatste heeft als voordeel dat het waarschijnlijk doelmatiger is en dat er ook een leereffect zal uitgaan naar andere studenten.

## 10.2 Ten slotte

---

In dit hoofdstuk staat een groot aantal alternatieven voor de gebruikelijke ‘potlood-en-papier’ toets. De alternatieven zullen lang niet in alle onderwijssituaties hanteerbaar zijn. Van docenten wordt verwacht zelf een keuze te maken. Waar het om gaat is dat de kennis en vaardigheden van studenten in de loop van hun opleiding met meerdere toetsvormen worden beoordeeld. Niet alle doelen van het onderwijs zullen met een enkele toetsvorm kunnen worden getoetst. Met andere woorden, de diversiteit aan toetsvormen dient hoog te zijn.

Ook zijn de alternatieven in dit hoofdstuk niet tot in detail uitgewerkt. Dat zou ook niet mogelijk zijn omdat opleidingen hun eigen accenten leggen. Wat de ene opleiding belangrijk vindt is voor de andere een detail. Waar het om gaat is dat de inhoud van dit hoofdstuk docenten aan het denken zet om variëteit in toetsvormen aan te brengen binnen de kaders van de doelen van de opleiding. Het blijkt in praktijk dat studenten andere dan gebruikelijke toetsvormen niet alleen hoog waarderen, maar dat ze hen ook tot ander studiegedrag aanzetten.

# Toetsen met gesloten vragen

*Henk van Berkel en Anneke Bax*

## Inleiding

Wat zijn de meest gebruikte vormen van gesloten toetsvragen?

Welke varianten zijn er?

Wanneer zijn gesloten toetsvragen te gebruiken?

Hoe is naast reproductiekennis, ook inzicht en toepassing te bevragen met gesloten vragen?

Wat zijn de constructieregels?

Uit hoeveel toetsvragen moet een toets met gesloten vragen bestaan?

Hoelang doen studenten over het beantwoorden van gesloten toetsvragen?

**11.1 Meerkeuzevragen – 114**

**11.2 Juist/onjuist- en ja/nee-vragen – 116**

**11.3 Andere gesloten vraagvormen – 117**

**11.4 Samenstellen van de toets en controleren van vragen – 117**

11.4.1 Constructieregels – 117

**11.5 Toetsen van hogere cognitieve vaardigheden met gesloten vragen – 119**

**11.6 Aanwijzingen en tips voor het construeren van gesloten vragen – 122**

**11.7 Checklist – 127**

De gesloten toetsvraag is een beproefde vorm om kennis van studenten in kaart te brengen, ondanks dat deze toetsvorm in de loop der jaren met de nodige scepsis is omgeven. Over het algemeen worden gesloten toetsvragen gebruikt bij grote groepen studenten (meer dan vijftig) omdat het nakijken ervan snel kan gebeuren met behulp van de computer. De valkuil is dat docenten met gesloten toetsvragen in de verleiding komen om uitsluitend reproductiekennis te bevragen. Dergelijke vragen zijn namelijk vrij eenvoudig te construeren. Zelden echter zullen doelstellingen voorschrijven dat van studenten uitsluitend reproductiekennis wordt verwacht. Docenten verlangen van studenten dat ze inzicht hebben in het geleerde en dit kunnen toepassen in verschillende situaties. Met gesloten vragen zijn – naast reproductiekennis – heel goed de zogenoemde hogere cognitieve vaardigheden als inzicht en toepassing te toetsen, mits de vragen zorgvuldig worden geconstrueerd (zie ook ■ tabel 11.2).

Gesloten toetsvragen komen in verschillende vormen voor. De meest bekende is de meerkeuzevraag, waarbij studenten een vraag beantwoorden door middel van het aankruisen van een van de gepresenteerde alternatieven. Daarnaast is een veel gehanteerde variant de juist/onjuist-vraag (die vergelijkbaar is met de ja/nee-vraag). Hierbij krijgen studenten een stelling, uitspraak of propositie voorgelegd waarvan zij de (on)juistheid moeten aangeven.

## 11.1 Meerkeuzevragen

---

Een meerkeuzevraag bestaat uit twee delen: de *stam* en de alternatieven. De *stam* is het gedeelte van de vraag dat het probleem bevat. De *alternatieven* zijn de antwoordmogelijkheden waaruit studenten kiezen. *Afleider* is de benaming voor een niet-correct alternatief; het correcte alternatief is de *sleutel*.

### ■ Varianten van meerkeuzevragen

Meerkeuzevragen komen in verschillende varianten voor:

- vragen met één correct alternatief;
- vragen met een alternatief dat te midden van andere het beste is;
- vragen met meer dan één correct alternatief;
- vragen waarvan de stam moet worden aangevuld met een van de alternatieven;
- vragen waarbij van de alternatieven alleen de eerste letter wordt vermeld;
- vragen waarbij een combinatie van alternatieven correct is.

De eerste twee varianten, ‘*één correct*’ en ‘*het beste alternatief*’, komen in de praktijk verreweg het meest voor. Daar is wel wat voor te zeggen omdat de andere varianten vaak nogal gekunsteld zijn en taalkundig moeilijk te lezen. Zij leiden de aandacht van studenten af van waar het om gaat: het meten van de gevraagde kennis of kunde. De laatste variant (vragen waarbij een combinatie van alternatieven correct is) is niet wenselijk om te gebruiken. Een voorbeeld van een dergelijke vraag is:

Beoordeel de juistheid van de volgende twee stellingen.

1. Piaget is vooral bekend om zijn theorie over schildpadden.
2. Het studiehuis is conceptueel verwant aan probleemgestuurd onderwijs.
  - a) 1 is juist en 2 is onjuist
  - b) 1 is onjuist en 2 is juist
  - c) 1 en 2 zijn beide juist
  - d) 1 en 2 zijn beide onjuist

Studenten hebben bij dergelijke vragen het gevoel dat ze twee vragen moeten beantwoorden. Dat is ook zo. In feite zijn het twee vragen van het type juist/onjuist, gewrongen in één vierkeuzevraag. Soms weten studenten het antwoord op een van de stellingen en gokken op de andere. Als zij de andere stelling fout gokken, verkrijgen zij doorgaans voor de hele vraag 0 punten. In zo'n situatie kan een conflict ontstaan waarover het College van Beroep voor de Examens zich moet buigen. Er zijn beroepszaken bekend waar studenten in dergelijke gevallen in het gelijk zijn gesteld. Zij behoren dus een half punt te krijgen op vragen waarvan ze de helft correct beantwoorden.

Het verdient aanbeveling deze variant op de vierkeuzevraag niet te hanteren. Een dergelijke vraagvorm kost veel lees/denktijd, frustreert studenten en vormt regelmatig een bron van conflicten. Door er bijvoorbeeld twee juist/onjuist-vragen van te maken kunnen er meer vragen in een toets worden opgenomen, wat de betrouwbaarheid van de toets doorgaans verhoogt (zie ook ► hoofdstuk 2).

#### ■ Alternatieven: aantal en volgorde

Het aantal alternatieven in meerkeuzevragen kan variëren. In Nederland is vier het gebruikelijke aantal terwijl bijvoorbeeld in de Verenigde Staten vijf alternatieven gebruikelijk zijn. Naarmate het aantal alternatieven groter is, wordt de raadkans kleiner, waarmee de vraag over het algemeen de kennis beter in kaart brengt. Het is echter niet mogelijk een oneindig aantal plausible alternatieven te bedenken. Al bij een vierkeuzevraag is het vaak zo dat de derde afleider er met de haren lijkt te zijn bijgesleept.

Een ander nadeel van veel alternatieven is dat de leestijd evenredig toeneemt. Gegeven een beperkte toetstijd is de keuze waarvoor de docent komt te staan: minder vragen met veel alternatieven of meer vragen met minder alternatieven. Er is veel onderzoek verricht naar het optimale aantal alternatieven. Hierbij is gezocht naar een optimum wat betreft de doelmatigheid, betrouwbaarheid en snelheid van constructie. De conclusie is dat het aanbeveling verdient meerkeuzevragen uit drie alternatieven te laten bestaan. Dat wil zeggen, één correct alternatief en twee afleiders.

Een ander probleem heeft betrekking op de rangschikking van de alternatieven. Een eenvoudige oplossing is de alternatieven in oplopende volgorde te plaatsen wanneer ze uit cijfers bestaan, en ze alfabetisch te rangschikken als ze uit zinnen of woorden bestaan. Studenten kunnen dan gemakkelijk zien welk systeem is gevolgd en zij zullen de volgorde van de alternatieven niet als een aanwijzing zien voor een bewuste manipulatie door de docent.

## 11.2 Juist/onjuist- en ja/nee-vragen

De populariteit van de juist/onjuist-vraag is voornamelijk te danken aan de Amerikaanse toetsdeskundige Ebel. Hij stelt dat alle kennis kan worden uitgedrukt in proposities, waarbij een propositie een stelling of uitspraak is die juist of onjuist is. Omdat de mate waarin studenten kennis hebben van leerstof wordt bepaald door hun vaardigheid juiste en onjuiste proposities uit die leerstof te herkennen, zijn juist/onjuist-vragen zeer geschikt, zo luidt zijn redenering (Ebel & Frisbie, 1991).

### ■ Varianten van juist/onjuist-vragen

Juist/onjuist-vragen komen in verschillende varianten voor. De meest bekende variant bestaat uit een stelling waarvan studenten moeten aangeven of deze juist of onjuist is. Bijvoorbeeld:

Volgens Rorschach heeft de inktvlekkentest een zeer hoge betrouwbaarheid ( $> 0,90$ ).	juist/onjuist
--	---------------

Soms is het nodig bij een stelling een wat meer uitgebreide context aan te geven. De stam bevat dan ook de context. De informatie in een stam is een gegeven en hoeft de student niet op juistheid te beoordelen. Het deel dat de student wel op juistheid moet beoordelen – de eigenlijke vraag – moet duidelijk in de lay-out onderscheiden zijn. Een stam kan worden gevolgd door een of meer vragen die alle betrekking hebben op dezelfde stam. Een voorbeeld:

Gegeven: In de DSM-IV-classificatie wordt de 'depressieve episode' gedefinieerd. Hierbij kunnen zogenaamde 'vitale symptomen' aanwezig zijn.

Stelling: Een vitaal symptoom is/vitale symptomen zijn:

- |                         |               |
|-------------------------|---------------|
| – gewichtsverlies       | juist/onjuist |
| – diarree               | juist/onjuist |
| – moeite met inslapen   | juist/onjuist |
| – concentratieproblemen | juist/onjuist |

Hierbij gaat het dus eigenlijk om een vraag die bestaat uit vier vragen over dezelfde context. Zo'n vraag wordt ook wel een matrixvraag genoemd.

### ■ Nadelen

Er kleven nadelen aan juist/onjuist-vragen. De stellingen moeten óf honderd procent juist óf honderd procent onjuist zijn. Vooral bij leerstof waar verschillende auteurs verschillende opvattingen over eenzelfde onderwerp aanhangen, is het risico aanwezig dat een uitspraak niet geheel (on)juist blijkt te zijn. Voorzichtigheid is dus geboden.

Een tweede nadeel van deze vraagvorm is dat studenten al vijftig procent van de vragen correct kunnen beantwoorden door te gokken. Nu is het bij de beoordeling weliswaar mogelijk daarvoor te corrigeren, maar het blijft jammer dat de helft van de vragen geen enkele informatie geeft over de kennis of het niveau van de student en om die reden in feite ‘vulling’ is voor de toets. Het gevolg is dat er veel meer vragen in een toets moeten worden opgenomen om een betrouwbaar oordeel te geven.

### 11.3 Andere gesloten vraagvormen

---

Naast de meerkeuzevragen en de juist/onjuist- of ja/nee-vragen zijn er nog meer ‘innovatieve’ gesloten vragen te bedenken. Dit zijn veelal vraagvormen die geschikt zijn voor digitale afnamen, zoals de sleepvraag of de hotspotvraag. Ook numerieke vragen waarbij de computer beoordeelt of het antwoord juist of onjuist is, valt onder de gesloten vragen. Bij deze vraagvormen is de raadkans zeer miniem.

### 11.4 Samenstellen van de toets en controleren van vragen

---

Een gesloten toets kan binnen een gegeven toetstijd relatief veel vragen bevatten. Daarmee is het mogelijk dat de toets een betere representatieve steekproef uit de leerstof vormt en – mede door het grote aantal vragen – een hoge betrouwbaarheid bereikt. Toch doen zich regelmatig problemen voor. Uit toetsanalyses blijkt dat de formuleringen van de toetsvraag en de alternatieven, maar ook de betrouwbaarheid van gesloten toetsen, nogal eens te wensen overlaten. Het construeren van gesloten vragen en het samenstellen van de toets vergt meer aandacht dan er doorgaans aan wordt besteed. Het hanteren van constructieregels voor de toets als geheel en voor de afzonderlijke toetsvragen werpt hierbij vruchten af.

#### 11.4.1 Constructieregels

---

##### ► Vooraf een specificatietabel maken

Voor het construeren van een toets bestaande uit gesloten vragen is – net zoals bij andere toetsvormen – het opzetten van een specificatietabel van belang (zie ook ► hoofdstuk 1, ► tabel 1.3). De tabel is meestal samen te stellen uit de doelstellingen. Men kan echter ook uitgaan van wat in de praktijk is onderwezen. Hiertoe maakt de docent een lijst met te toetsen onderwerpen, waarbij tevens wordt aangegeven op welk cognitief niveau vragen moeten staan om een representatieve toets te verkrijgen (zie ► hoofdstuk 1). Naarmate een onderwerp belangrijker is, dienen er meer vragen over te worden gesteld.

##### ► Meer vragen bedenken dan nodig voor de toets

In eerste instantie moet een groter aantal vragen worden geconstrueerd dan in een toets zal worden opgenomen. Hoe zorgvuldig de constructie van een vraag ook heeft plaatsgevon-

den, het is altijd noodzakelijk de conceptvragen voor te leggen aan collega's of vakgenoten. Ervaring leert dat een vraag vaak kan worden verbeterd, vormtechnisch en/of inhoudelijk.

Een goed begin van een bespreking met collega's is om hen te vragen de vraag te beantwoorden. Wanneer geen eensgezindheid bestaat over het juiste alternatief, is er alle reden om over de vraag van gedachten te wisselen. In de bespreking moeten per vraag steeds de volgende punten aan de orde komen:

- Is de vraag relevant ten opzichte van de doelstellingen?
- Is de vraag gericht op het gewenste niveau?
- Is het belangrijk dat de student het gevraagde weet?
- Is het goede antwoord (de sleutel) inderdaad correct?
- Is de literatuur eenduidig over de te toetsen kennis?

Slechts wanneer alle vragen met 'ja' worden beantwoord, is de vraag in aanzet een goede vraag. In de praktijk van het hanteren van de geschetste procedure blijkt dat vaak slechts 25% van de conceptvragen de bespreking met vakgenoten doorstaat. Het is niet zo dat de betreffende vragen naar de prullenbak moeten worden verwezen; meestal zijn ze te verbeteren.

De procedure impliceert dat men tijdig met de vraagconstructie moet beginnen.

### ➤ Minimaal aantal vragen

De omvang van het studieonderdeel, bijvoorbeeld uitgedrukt in het aantal te behalen studiepunten of het aantal bladzijden van de leerstof, én de beschikbare toets-tijd bepalen het aantal vragen in de toets. Een toets is op te vatten als een steekproef uit alle mogelijke vragen uit de leerstof. Belangrijke onderwerpen nemen doorgaans in de leerstof meer ruimte in. Dat impliceert dat over belangrijke onderwerpen ook meer vragen in de toets moeten komen. Dit verhoogt de representativiteit van de toets (zie ► hoofdstuk 2).

Naast de representativiteit is de betrouwbaarheid van de toets een belangrijke indicator voor de kwaliteit van de toets. Bij steekproeven geldt: hoe groter de steekproef, des te meer geven de uitkomsten een correct beeld van de populatie. Dat geldt ook voor toetsen: hoe meer vragen in de toets, des te beter zullen de toetsscores de kennis van studenten weerspiegelen. In de praktijk blijkt dat met een gesloten toets bestaande uit veertig vierkeuzevragen een acceptabele betrouwbaarheid ( $> 0,80$ ) is te behalen. Indien de vragen drie alternatieven bevatten, zijn minimaal zestig vragen vereist; een toets met tweekeuzevragen dient minimaal tachtig vragen te bevatten. Dit zijn echter zeer algemene richtlijnen die niet in alle situaties gelden. Indien bijvoorbeeld een toets gaat over 1500 bladzijden leerstof, zullen veertig vierkeuzevragen doorgaans geen goede afspiegeling zijn van de totale leerstof.

In de praktijk blijkt dat het aantal op te nemen vragen in de toets eerder wordt bepaald door de beschikbare toets-tijd, dan door de omvang van de leerstof, ofschoon dit niet de 'koninklijke' weg is om het aantal vragen te bepalen. Met behulp van ■ tabel 11.1 is het mogelijk het aantal op te nemen vragen te berekenen, gegeven de toets-tijd.

■ **Tabel 11.1** Antwoordtijd per vraagtype.

Vraagtype	Beantwoordingstijd
juist/onjuist-vraag	50 seconden
meerkeuzevraag met 2 antwoordmogelijkheden	50 seconden
meerkeuzevraag met 3 antwoordmogelijkheden	60 seconden
meerkeuzevraag met 4 of 5 antwoordmogelijkheden	75 seconden

Uit: Van Berkel, 1999.

## 11.5 Toetsen van hogere cognitieve vaardigheden met gesloten vragen

Gesloten toetsvragen hebben in veel kringen een slechte naam. Voor velen is het ondenkbaar dat kennis van studenten in het hoger onderwijs in kaart kan worden gebracht door middel van het zetten van een streepje. Toegegeven, wie de meerkeuzevragen die gangbaar zijn in het hoger onderwijs aan een kritische blik onderwerpt, is gauw geneigd die mening te onderschrijven. De aversie tegen gesloten vragen is daartoe te herleiden; gesloten vragen neigen te sterk naar het bevragen van reproductiekennis. Deze paragraaf beoogt aan te tonen dat het ook anders kan.

Hoe kan een docent de vraag zo stellen dat het gewenste cognitieve niveau daadwerkelijk wordt bevestigd? In ► hoofdstuk 1 zijn de zes cognitieve niveaus die Bloom (1972) hanteert onderscheiden. Het is een hiërarchie die loopt van het niveau van reproduceren tot evalueren. Over het algemeen geldt dat hoe hoger in de hiërarchie, des te lastiger het is om er een vraag over te construeren. In de onderwijspraktijk zijn deze zes niveaus terug te brengen tot drie: reproductiekennis, inzicht en toepassen, waarbij de volgende definities gangbaar zijn:

- *Reproductiekennis*: kennisvragen zijn op te lossen door herinnering of herkenning. De vraag bevat een bekende probleemsituatie, zij het vaak in een geparafraseerde vorm.
- *Inzicht*: oplossingen voor inzichtvragen vereisen een hoge mate van reorganisatie van het probleem en zelfstandig denken. Deze vragen doen een beroep op intellectuele vaardigheden als interpretatie, analyse, synthese en evaluatie.
- *Toepassen*: bij toepassingsvragen worden studenten voor een voor hen nieuw probleem gesteld. Het probleem wordt opgelost met een formule, wet of oplossingsprincipe.

Er bestaat een handig hulpmiddel om ervoor te zorgen dat een vraag betrekking heeft op een van deze niveaus. Met behulp van een specifiek werkwoord in de vraag wordt een bepaald cognitief niveau getoetst. In ■ tabel 11.2 staat een overzicht van werkwoorden en beoogd niveau (zie ook ► <http://www.celt.iastate.edu/pdfs-docs/teaching/RevisedBloomsHandout.pdf>). Dit hulpmiddel is uiteraard ook bruikbaar bij andere vraag- of toetsvormen.



■ **Tabel 11.2** Werkwoordvormen als hulpmiddel bij vraagconstructie per cognitief niveau.

Reproductiekennis		Inzicht			Toepassen
<i>feitenkennis</i>	<i>samenvatten</i>	<i>analyseren</i>	<i>synthetiseren</i>	<i>evalueren</i>	<i>toepassen</i>
benoemen beschrijven citeren definiëren identificeren labelen opsommen	concluderen extrapoleren interpreteren onderscheid maken parafaseren samenvatten schatten selecteren uitleggen vertalen vertellen in eigen woorden	deduceren illustreren onderscheiden ontrafelen relateren schetsen toelichten	categoriseren componeren formuleren herschrijven maken ontwerpen samenvatten	concluderen bekritisieren contrasteren interpreteren ondersteunen rechtvaardigen valideren verdedigen vergelijken waarderen	berekenen demonstreren gebruiken maken ontwikkelen oplossen organiseren produceren relateren transfereren veranderen verklaren voorbereiden wijzigen

Uit: Van Berkel, 1999.

Op het eerste gezicht zijn niet alle werkwoorden direct hanteerbaar in gesloten vragen. Hoe kunnen studenten een begrip uitleggen in hun eigen woorden? Hier is binnen de context van de gesloten vraag een oplossing voor. Bijvoorbeeld:

Gegeven: Arthur loopt samen met zijn blinde broertje te wandelen in een bos. Arthur ziet een paddenstoel, een boleet, en vertelt dat aan zijn broertje. Zijn broertje vraagt hem hoe de boleet eruitziet.

Vraag: Welk antwoord zou jij hebben gegeven?

- Een boleet is ongeveer 10 cm hoog en is wit.
- Een boleet is ongeveer 15 cm hoog, de stengel is wit en de top is rood.
- Een boleet is ongeveer 20 cm hoog en groeit aan een boom.

Voorbeelden van vragen op *feitenkennisniveau* zijn:

Bij hoeveel graden Celsius gaat water bij een druk van 1 atmosfeer koken?

- 100°
- 180°
- 360°

De huidige hoofdstad van België is Hasselt.

juist/onjuist

Bij vragen op *inzichtniveau* moeten studenten onderscheid maken tussen verschillende feiten, opinies of hypothesen, maar ook de merites kunnen aangeven van methoden of producten ten opzichte van een bepaald doel en eventueel iets origineels kunnen maken. Enkele voorbeelden zijn:

Gegeven: Uit een studie van de onderwijstheorie van Dalton naar het aspect 'zelfwerkzaamheid', blijkt dat de theorie van Dalton hier geen uitspraken over doet. Toch stelt Bieren (1999) dat het studiehuisconcept volgt uit de onderwijstheorie van Dalton.

Vraag: Is dit een correct standpunt van Bieren? ja/nee

Gegeven: Twee gezonde ouders zijn beiden drager van een autosomaal recessief overervend gen. Bij een deel van hun kinderen zullen zich manifeste ziekteverschijnselen vertonen, bij de rest niet. De kinderen zonder ziekteverschijnselen hebben een bepaalde kans om drager te zijn van dit gen.

Stelling: Deze kans bedraagt  $2/3$ . juist/onjuist

Enkele voorbeelden van vragen op *toepassingsniveau*, die een beroep doen op het hanteren of toepassen van eerder verworven informatie in een andere situatie:

Gegeven: Een studietoets bestaat uit 70 vragen. De vragen hebben een gemiddelde intercorrelatie van 0,14.

Vraag: Welke waarde benadert het best coëfficiënt alfa?

- a) 0,50
- b) 0,75
- c) 1,00

Voor het maken van mayonaise is meer centiliter olie nodig dan water. juist/onjuist

Gegeven: Zoutzuur bestaat uit de elementen waterstof, zwavelzuur en zuurstof. Wanneer aan zoutzuur een loog wordt toegevoegd, ontstaat een bruine damp.

Stelling: Die damp is broom. juist/onjuist

Deze laatste twee voorbeelden sluiten uiteraard alleen aan bij de hogere cognitieve niveaus als de studenten de juistheid ervan alleen kunnen bepalen door hun analysevaardigheid. Als de tekst letterlijk in het boek staat vallen de vragen weer in de categorie 'reproductie'.

## 11.6 Aanwijzingen en tips voor het construeren van gesloten vragen

De onderstaande aanwijzingen en tips voor het construeren van gesloten vragen gelden voor zowel meerkeuze als voor juist/onjuist en ja/nee-vragen.

### ➤ Vermijd vage of te absolute aanduidingen

Vage aanduidingen in de stelling, de stam of in een van de alternatieven als veel, soms, goed, effectief, belangrijk, groot en klein, geven te grove en daardoor discutabele grenzen aan.

Regenwolken hebben een licht gewicht. juist/onjuist

De bedoeling van deze vraag is te achterhalen of studenten weten dat vochtige lucht lichter is dan droge lucht. Deze vergelijking ontbreekt echter in de stelling. Bovendien is de zinsnede 'licht gewicht' relatief. Regenwolken wegen vaak tonnen, wat nauwelijks licht is te noemen. Om het probleem van relatieve termen in de stellingen op te lossen, kan in de stelling een vergelijking worden opgenomen:

Vochtige lucht heeft een lager soortelijk gewicht dan droge lucht. juist/onjuist

Toetsvragen worden duidelijker wanneer begrenzingsen staan aangegeven. Zo kunnen vage aanduidingen als 'klein', 'soms' en 'vaak' makkelijk worden omgezet:

'Een klein percentage' wijzigen in: 'Tussen tien en twintig procent';  
 'Soms' wijzigen in: 'In de minderheid van de gevallen';  
 'Vaak' wijzigen in: 'In de meerderheid van de gevallen'.

### ➤ Bevraag één kennisaspect per vraag

Een van de meest voorkomende fouten bij het construeren van gesloten vragen – in het bijzonder bij juist/onjuist-vragen – is de zogenoemde 'dubbele vraag':

In Nederland is sinds het oprichten van het CITO in 1967 de trend ontstaan leerlingen op basisscholen regelmatig met geobjectiveerde schooltoetsen lastig te vallen. juist/onjuist

Wat wordt hier precies gevraagd? Of het CITO in 1967 is opgericht? Of dat studenten weten dat basisscholen in 1967 nog niet bestonden? Of schooltoetsen zijn geobjectiveerd? Of het maken van toetsen 'lastigvallen' is te noemen?

Dubbele vragen brengen studenten in verwarring. Wanneer moet met 'juist' worden beantwoord: als alle deelvragen juist zijn of als één stelling juist is? Als de naar het oordeel van de lezer belangrijkste stelling juist is? Vragen dienen zich te richten op één kennisaspect, idee, feit, onderwerp enzovoort. Het is niet moeilijk om dergelijke vragen op te splitsen in verschillende vragen c.q. stellingen, bijvoorbeeld:

- Het CITO is in 1967 opgericht.
- In 1967 bestonden er basisscholen.
- Volgens De Groot en Van Naerssen (1971) is de meerderheid van de schooltoetsen geobjectieerd.
- Volgens Popper ervaren kinderen het maken van een toets als een traumatische ervaring.

### ➤ Gebruik een positieve in plaats van negatieve formulering

Vragen die negatief zijn geformuleerd, toetsen dezelfde kennis als positief gestelde vragen, maar bij de negatieve formulering wordt van de student tevens een taaltruc gevraagd, namelijk een omkering van het gestelde. Dit verhoogt de moeilijkheid op oneigenlijke gronden, namelijk niet op het beoogde kennisgebied, maar op taalvaardigheid.

- Welke van de onderstaande Nederlandse steden is *geen* provinciehoofdstad?
- a) Amsterdam
  - b) Maastricht
  - c) Utrecht

Als de negatieve term toch wordt gebruikt, moet deze in elk geval opvallen, bijvoorbeeld door het negatief gestelde woord te onderstrepen of te cursiveren.

### ➤ Vermijd dubbele ontkenningen

Bij vragen die een dubbele ontkenning bevatten, geldt ook dat ze tevens taalvaardigheid meten. Dat zal meestal niet de bedoeling zijn. In de meeste gevallen kan een dubbele ontkenning worden omgezet in een positieve zin, zij het dat soms de antwoordsleutel moet worden gewijzigd.

### ➤ Voorkom subjectiviteit in de vraag

Vragen die opinies bevatten, zijn subjectief. Zij komen vooral voor in zinnen met werkwoorden als ‘moeten’, ‘dienen’ en ‘behoren’. Vakgenoten zullen het lang niet altijd eens zijn over wat het correcte alternatief is.

- |  |               |
|--|---------------|
| De buitenspelregel bij voetbal verlaagt de aantrekkelijkheid van het spel. | juist/onjuist |
|--|---------------|

In dit voorbeeld is het beter een neutrale formulering te gebruiken of een contextverruiming in te voeren:

‘Volgens Crujff verlaagt ...’

### ➤ Geef bij het gebruik van citaten de context

Citaten worden dikwijls gebruikt zonder dat er een stroming of kader staat aangegeven.

Het geheel is meer dan de som van de delen.

juist/onjuist

Met een dergelijke vraag kunnen studenten vanuit diverse invalshoeken redeneren. Bij dit soort vragen is het voor de duidelijkheid van belang om een kader te vermelden wanneer of door wie dit wordt beweerd:

In de gestaltpsychologie ...

In plaats van het vermelden van de stroming of de context, is het ook toegestaan een naam te noemen, mits die persoon synoniem is aan een bepaalde stroming:

Freud bespreekt het geval van ...

### ➤ Vermijd afleidende informatie

Overbodige informatie in de vraag leidt studenten af en kan ze op het verkeerde been zetten.

Gegeven: Mulisch stelt in zijn roman *Het stenen bruidsbed* de zoektocht naar wat werkelijk gebeurd is aan de orde.

Vraag: De uitspraak 'de leugen regeert' is afkomstig van

- a) de hoofdfiguur in de roman.
- b) Harry Mulisch.
- c) koningin Beatrix.

De vraag is onzinnig gesteld. Er zijn dingen door elkaar gehaald die weliswaar een goed antwoord niet in de weg staan, maar die uitsluitend afleidend zijn. Het gegeven in de vraag is overbodig. Het kost studenten veel extra tijd en energie om erachter te komen of de gegeven informatie daadwerkelijk van belang is.

### ➤ Geef de context bij vragen

De Nederlandse taal bevat veel woorden die zonder context betekenisloos zijn, bijvoorbeeld: aantal, bepaalde, aspecten, punten, onderdelen.

Het draaien van hun kop is een betekenisvol aspect bij zonnebloemen.

juist/onjuist

Door de context aan te geven of woorden te hanteren die wel een eenduidige betekenis hebben, wordt de vraag veel duidelijker.

Gegeven: Sommige zonnebloemen draaien hun bloem in de richting van de zon.

Stelling: Zonnebloemen die dat doen, groeien gemiddeld hoger dan zonnebloemen die dat niet doen.

juist/onjuist

### ➤ Vermijd vragen naar details en strikvragen

Strikvragen en listigheden geformuleerd in details, zijn uit den boze. ‘Het kan toch niet de bedoeling van het onderwijs zijn, de student te laten struikelen?’, luidt de mening van Crombag (1981). Als docenten toetsvragen richten op de knapste jongetjes van de klas, dan dienen docenten zich te realiseren dat zij daarmee de negentig procent anderen behoorlijk frustreren. Een (fout) voorbeeld uit een mavo-C-examen is:

Wanneer een vliegtuig neerstort in de Noordzee tussen Engeland en België, moet volgens de Zeewet (1918) de helft van de overlevenden in Engeland worden begraven en de andere helft in België.	juist/onjuist
--	---------------

Overlevenden worden niet begraven. Deze vraag is een voorbeeld van een strikvraag.

In zijn boek <i>A brief history of time</i> stelt Hawking dat zich in zwarte gaten het verschijnsel singulariteit voordoet.	juist/onjuist
---	---------------

De inhoud van deze vraag stijgt ver uit boven datgene wat van mavo-C-leerlingen kan worden verwacht. Docenten dienen zich te allen tijde te realiseren in hoeverre een gestelde vraag echt van belang is.

### ➤ Zorg voor gelijke lengte van alternatieven

Het is niet gemakkelijk een correct alternatief kort en bondig te formuleren. De lengte van de alternatieven is daarom een aanwijzing voor de correctheid. Een correct antwoord vereist vaak meer woorden dan een afleider. Uit ervaring weten studenten dat een dergelijk alternatief vaak de sleutel is.

Epistemologie is een filosofische stroming die betrekking heeft op

- de aard en de oorsprong van kennis, in het bijzonder de wijze waarop mensen externe stimuli kunnen omzetten in kennis.
- het wezen van het bestaan.
- schoonheid.
- sterfte.

In dit voorbeeld zou het eerste alternatief kunnen worden geherformuleerd tot: ‘de aard van kennis’. Ook kan worden geprobeerd de alternatieven ongeveer even lang te maken als de sleutel, waardoor het voor studenten die de stof niet beheersen, moeilijker is het juiste antwoord te gokken.

### ➤ Zorg voor plausibele alternatieven

Bij een goede meerkeuzevraag dienen studenten alle alternatieven serieus te overwegen voordat ze een alternatief aankruisen. Wanneer een van de alternatieven bij voorbaat al door studenten is te elimineren, verhoogt dit de raadkans. Bijvoorbeeld:

Welke Nederlandse sportman won een gouden medaille op een van de Olympische Spelen?

- a) Eef Kamerbeek
- b) Reinier Paping
- c) Stephan van den Berg
- d) Steve Jobs

Wanneer studenten het correcte antwoord niet weten, kunnen zij in ieder geval alternatief 'd' als niet plausibel overslaan om vervolgens te gokken met een kans van één op drie om de vraag correct te beantwoorden.

### ➤ Zorg ervoor dat de alternatieven elkaar niet overlappen

Alternatieven mogen elkaar niet overlappen. Dit kan immers impliceren dat er meer antwoorden correct zijn, terwijl niet is aan te geven welke het beste is. Het werkt het best als slechts één alternatief het correcte is.

Kegeltjes worden aangetroffen in:

- a) de blinde vlek
- b) de fovea
- c) de periferie van de retina
- d) in het oog

In dit voorbeeld is dat niet het geval doordat alternatief 'd' alle alternatieven omvat. Indien het vierde alternatief wordt vervangen door 'de cornea', is de vraag correct.

### ➤ Vermijd het gebruik van het alternatief 'geen van de vermelde alternatieven'

In de toetspraktijk is een vraag met het alternatief 'geen van de vermelde alternatieven' berucht.

Gegeven: De Vries (1999) trekt uit de bewering van De Swaan ten aanzien van het geven van colleges, de conclusie dat docenten meer entertainer moeten zijn.

Vraag: Welke filosoof zal deze conclusie ondersteunen?

- a) Hegel
- b) Kierkegaard
- c) Popper
- d) Geen van deze drie

Op zichzelf zijn dit geen slechte vragen; ze blijken alleen moeilijk te zijn. Dit is verklaarbaar. De student moet immers van alle alternatieven kunnen aangeven of het correct of niet correct is. Het is efficiënter meer vragen te construeren uit die ene vraag.

## 11.7 Checklist

### Checklist voor het construeren van gesloten vragen

#### *Vormtechnische aspecten*

- Bevat de vraag geen vage aanduidingen?
- Bevat de vraag slechts één duidelijk probleem?
- Is de vraag niet onnodig negatief geformuleerd?
- Zijn ontkenningen als niet onderstreept of gecursiveerd?
- Bevat de vraag geen dubbele ontkenningen?
- Bevat de vraag geen subjectieve uitspraken?
- Bevat de vraag geen woorden als altijd, nooit, meestal, zeker?
- Staat bij een citaat in de vraag de context?
- Bevat de vraag geen overbodige informatie?
- Is de vraag opsplitsbaar in een gegeven en een vraaggedeelte?
- Heeft de vraag geen betrekking op een detail?
- Is de vraag geen strikvraag?
- Is de stelling 100% juist of 100% onjuist?

#### *Bij meerkeuzevragen geldt nog:*

- Is de vorm 'geen van de vermelde alternatieven' beslist nodig?
- Komen er geen overlappingsen in de alternatieven voor?
- Zijn alle alternatieven ongeveer even lang?
- Zijn de alternatieven in oplopende/alfabetische volgorde geplaatst?
- Zijn alle alternatieven plausibel?



# Toetsen met open vragen

*Tom Erkens*

## Inleiding

Welke vormen van open vragen zijn er?

Wat zijn de voor- en nadelen?

Welke regels moeten bij de constructie in acht worden genomen?

Wat zijn veelvoorkomende beoordelaarsverschillen en hoe kunnen die worden beperkt?

Aan welke eisen moet een correctievoorschrift voldoen?

**12.1 Vormen van open vragen – 130**

**12.2 Voor- en nadelen van open vragen – 131**

12.2.1 Voordelen – 131

12.2.2 Nadelen – 133

**12.3 Correctievoorschrift – 133**

**12.4 Constructieregels – 134**

**12.5 Oorzaken voor verschillen in beoordeling – 139**

**12.6 Oplossingen voor beoordelaarsverschillen – 140**

**12.7 Checklist – 140**

Open vragen zijn vragen waarbij studenten zelf het antwoord moeten formuleren. Dat antwoord kan in lengte variëren van één woord of enkele woorden tot een volledig betoog. De open vraag heeft voordelen, maar geeft ook problemen die doorgaans pas worden ontdekt tijdens of na de afname van de toets. Voor een niet onbelangrijk deel zijn de nadelen terug te voeren tot kwesties die met het formuleren te maken hebben. Met andere woorden: het gaat veelal om taalproblemen. Een niet juist of niet volledig juist geformuleerde vraag levert vaak een niet verwacht of ongewenst antwoord op en veroorzaakt daardoor problemen bij de beoordeling op.

Open vragen zijn bij uitstek geschikt om hogere cognitieve vaardigheden te toetsen. Dat betreffen vaardigheden als toepassen, analyseren, synthetiseren, evalueren. In de onderwijspraktijk worden open vragen ook gebruikt voor het toetsen van feitenkennis. Men kan zich afvragen of open vragen daarvoor gebruikt moeten worden. Het is bewezen (Ebel & Frisbie, 1991) dat feitenkennis op een efficiëntere en meer betrouwbare wijze met behulp van gesloten vragen kan worden bevraagd dan met open vragen. Bij de keuze voor een vraagvorm kunnen echter andere overwegingen een rol spelen, bijvoorbeeld inhoudsvaardigheid, betrouwbaarheid, objectiviteit en doelmatigheid. Soms kan daardoor de situatie ontstaan dat de toetsvorm een (iets) andere is dan door de doelstellingen geïndiceerd.

## 12.1 Vormen van open vragen

Open vragen komen in diverse vormen voor. Er is daarbij een onderscheid te maken naar de *lengte* van het antwoord, en naar de *benodigde tijd* voor de beantwoording.

### ■ In- en aanvulvragen

Bij in- en aanvulvragen verwachten docenten dat studenten een onvolledige zin, berekening of tekening completeren. Lezen en beantwoorden van een dergelijke vraag kost weinig tijd, hooguit tien seconden. Voorbeelden zijn:

De langste rivier van Europa is de ...  
De moord op ... vormde de directe aanleiding tot de Eerste Wereldoorlog.

### ■ Korte-antwoordvragen

Op korte-antwoordvragen moet met een citaat, enkele woorden, een enkelvoudige zin, een getal, een (eenvoudige) tekening of formule worden geantwoord. De tijd voor het lezen en de beantwoording zal meestal niet meer dan een minuut vergen. Een voorbeeld van een dergelijke vraag gesteld aan de hand van een hier niet vermelde korte context, is:

Van welke vorm van werkloosheid is hier sprake?

### ■ Lange-antwoordvragen

Studenten dienen lange-antwoordvragen te beantwoorden met samenhangende zinnen. Ook een toelichting, een gecompliceerde berekening, tekening of bewijs bestaande uit

verschillende stappen, zijn voorbeelden van antwoordmogelijkheden. Voor het lezen en beantwoorden van een dergelijke vraag is al gauw drie tot vijf minuten nodig. Een voorbeeldvraag is:

Leg uit waarom het voor sommige bedrijven economisch gezien voordelig kan zijn afval opnieuw te gebruiken.

#### ■ Opstelvragen/betoogvragen

Typische antwoorden op opstelvragen/betoogvragen zijn een samenhangende tekst, gestructureerd bijvoorbeeld naar inleiding, midden en slot; een zeer gedetailleerde tekening of berekening die beide als een afgerond geheel worden beschouwd. Het lezen en beantwoorden van een dergelijke vraag is een tijdrovende aangelegenheid. Vijftien tot dertig minuten voor een dergelijke opdracht is niet ongebruikelijk. Een voorbeeld is:

De heer Anders reageerde met een ingezonden brief op een artikel in Het Hollands Dagblad over bijbaantjes voor studenten. Zijn brief begon als volgt:

» In uw krant verscheen onlangs een artikel waarin de minister van Onderwijs, Cultuur en Wetenschappen zijn bezwaren tegen bijbaantjes van studenten uiteenzet. Hieronder wil ik de voordelen van bijbaantjes voor studenten voor de Nederlandse economie uiteenzetten ... «

Opdracht: Schrijf het vervolg van de brief. In de brief moeten de volgende aspecten aan de orde komen:

- a) de bijdrage van bijbaantjes van studenten aan het terugdringen van de frictie-werkloosheid in de toekomst
- b) het multipliereffect van de stijgende bestedingen van studenten.

Veelal blijken studenten meer tijd voor de beantwoording nodig te hebben dan de docent voor afname had geschat. Er is onderzoek verricht naar de antwoordtijd op vragen. Deze richttijden staan weergegeven in de [tabel 12.1](#). De vermelde beantwoordingstijd betekent dat de overgrote meerderheid van studenten het betreffende vraagtype in de vermelde tijd kan beantwoorden.

## 12.2 Voor- en nadelen van open vragen

### 12.2.1 Voordelen

Sommige doelstellingen zijn beter te toetsen met open vragen dan met gesloten vragen. Dit geldt voor doelen waarbij originaliteit of creativiteit een rol spelen (bijvoorbeeld bij het ontwerpen van een onderzoeksopzet), bij de toetsing van complexere vaardigheden of wanneer het te toetsen kennisdomein omvangrijk is. Het gaat hierbij om productieve

■ **Tabel 12.1** Beantwoordingstijd per open vraagtype.

Vraagtype	Beantwoordingstijd
open vraag (antwoord vergt 1 woord of zin)	1 minuut
open vraag (antwoord vergt 1/4 pagina-A4)	5 minuten
open vraag (antwoord vergt 1/2 pagina-A4)	10 minuten
open vraag (antwoord vergt 1 pagina-A4)	25 minuten
open vraag (antwoord vergt 2 pagina's-A4)	60 minuten

Uit: Van Berkel, 1999.

doelen in tegenstelling tot reproductieve doelen. Dat wil zeggen: bij de beantwoording van een open vraag (vooral de lange-antwoordvraag) hoeven de studenten niet in hun vrijheid te worden beperkt. Bij de beantwoording kunnen de studenten zodoende creativiteit aan de dag leggen. Een voorbeeld van een vraag die moet worden beantwoord met behulp van twee bronnen is:

Gebruik bron 16 en 17.

In de prent geeft de tekenaar zijn mening over de toespraak van president Truman.

Wat is de mening van de tekenaar? Ondersteun je antwoord door twee elementen uit de prent te verbinden met twee elementen uit de toespraak.

Daar waar taalvaardigheidsaspecten bij de toetsing een rol spelen, verdient de open vraag de voorkeur boven de gesloten vraag. De studenten kunnen immers bij de beantwoording van open vragen productief gebruikmaken van de taal. Overigens is deze regel niet voor iedere student gunstig omdat sommige studenten wel baat hebben bij gesloten vragen waar ze kunnen gokken om tot het correcte antwoord te komen. Een voorbeeld:

Maak een goedlopende samenvatting in correct Nederlands van maximaal 200 woorden van de tekst 'De liefde tot zijn land is ieder aangeboren ...' Zorg ervoor, dat je samenvatting begrijpelijk is voor iemand die de oorspronkelijke tekst niet kent.

Uit je samenvatting moet duidelijk worden:

- welke overeenkomsten er zijn tussen de identiteit van het gezin en de natie
- welk gevaar zich voordoet op het niveau van de identiteit van de natie
- welke factoren dit gevaar beperken en waardoor deze factoren remmend werken
- met welke redenering kosmopolitisme geen alternatief wordt gevonden voor de behoefte aan een nationale identiteit
- wat wordt geconcludeerd ten aanzien van het trots zijn op de eigen nationale identiteit.

### 12.2.2 Nadelen

---

Voor taalzwakke studenten leveren open vragen extra problemen op. Het is niet altijd duidelijk of een 'foute' beantwoording van de vraag uitsluitend te wijten is aan onvoldoende kennis dan wel (mede) het gevolg is van een onvoldoende taalbeheersing. Met andere woorden is de taal of de vakinhoud de bottleneck? Het eenduidig formuleren van open vragen is lastig. Niet eenduidig geformuleerde vragen zijn voor studenten op meer manieren te interpreteren. Zodra studenten moeten interpreteren, zal het door de docent gewenste of verwachte antwoord waarschijnlijk niet worden gegeven, althans in mindere mate. Niet eenduidig geformuleerde open vragen leveren beoordelingsproblemen op; de studenten vertonen een ander antwoordgedrag dan was beoogd. Ook indien de vraag eenduidig is geformuleerd, kunnen verschillen ontstaan tussen beoordelaars omdat ze bijvoorbeeld van mening verschillen over het juiste antwoord, of een beoordeling van één docent wordt niet consistent op iedere student toegepast.

De beantwoording van een toets bestaande uit open vragen kost meer tijd dan de beantwoording van een toets met gesloten vragen. Voor het beantwoorden van een gesloten vraag is een minuut meestal voldoende, terwijl het beantwoorden van open vragen veel meer tijd vergt, soms meer dan een kwartier, afhankelijk van de vraagstelling. Dit heeft nadelige gevolgen voor het aantal vragen dat, gegeven de beschikbare toetstijd, gesteld kan worden. Dat gaat ten koste van de inhoudsvaliditeit en betrouwbaarheid van de toets (zie ► hoofdstuk 2).

## 12.3 Correctievoorschrift

---

Het beoordelen van open toetsvragen is zeer arbeidsintensief. Een hulpmiddel bij het beoordelen, is het hanteren van een antwoordmodel, ook wel aangeduid als correctievoorschrift. Dat dient bij of direct na het construeren van de toetsvragen te worden gemaakt.

Wanneer beoordelaars een correctievoorschrift hanteren, verkleint dat de verschillen tussen beoordelaars. Een correctievoorschrift leidt ertoe dat (extreem) strenge en (extreem) soepele beoordelaars zich matigen omdat ze niet hun eigen intuïtieve oordelen mogen geven. Het hanteren van een dergelijk voorschrift voorkomt ook dat docenten verschillende criteria hanteren bij de beoordeling. Een correctievoorschrift is meer dan een eenvoudige checklist aan de hand waarvan de beoordelaar de antwoorden naloopt. Een correctievoorschrift bevat algemene en vakspecifieke regels voor de beoordeling en voorbeelden van juiste, gedeeltelijk juiste en (voor de hand liggende) onjuiste antwoorden. Ook bevat het modelantwoord de maximaal te behalen score per vraag en (waar zinvol en mogelijk) de verdeling van deze maximumscores in deelscores. De algemene regels gaan in op de wijze van beoordelen en geven aan hoe de docent kan omgaan met gedeeltelijk juiste antwoorden, antwoorden die niet in het modelantwoord voorkomen, zogenoemde doorwerkende fouten (een eerder gemaakte fout die de rest van het antwoord ook fout maakt), identieke fouten en taalfouten. Daarnaast kunnen de regels enkele tips bevatten die de hiervoor vermelde beoordelaarseffecten tegengaan.

## 12.4 Constructieregels

Een aantal problemen met open vragen kan worden voorkomen indien de constructeur zich houdt aan een aantal regels.

### ➤ **Formuleer bij open vragen eerst het modelantwoord en vervolgens de vraag.**

Een vraag kan uiteenlopende antwoorden uitlokken die stuk voor stuk te verdedigen zijn omdat de antwoorden zijn terug te voeren op uiteenlopende interpretaties van de vraag. Het probleem wordt voorkomen door andersom te werk te gaan. Formuleer eerst het wenselijke antwoord, het zogenoemde *modelantwoord*, en dan pas de bijbehorende vraag. Een voorbeeld:

Bedenk een reden waarom politieagenten die hun werk doen, toch voor een rechter moeten kunnen verklaren waarom zij geweld hebben gebruikt.

### ➤ **Voeg zo nodig antwoordrestricties aan de vraag toe.**

Een ander probleem dat zich kan voordoen, is dat studenten in hun antwoord uitvoeriger dan wel beknopter zijn dan de docent verwacht. Dit gebeurt vooral bij vragen waarbij de docent geen of onvoldoende antwoordrestricties heeft geformuleerd. Hoeveel voorbeelden moeten er worden gegeven? Uit hoeveel woorden mag het antwoord maximaal bestaan?, enzovoort. Een voorbeeld van een dergelijke vraag is:

Noem twee redenen om alcohol of drugs te gaan gebruiken.

### ➤ **Geef bij iedere toets(deel)vraag aan hoeveel punten kunnen worden behaald.**

Studenten zijn niet altijd in staat een inschatting te maken van het relatieve belang van de vraag respectievelijk de vraagonderdelen. Hierdoor komt het voor dat zij aan bepaalde toetsvragen of vraagonderdelen die slechts een beperkt aantal scorepunten opleveren, relatief veel tijd besteden.

### ➤ **Ontwikkel open vragen samen met collega's.**

Bij het formuleren van een open vraag realiseren docenten zich niet altijd dat studenten een bepaalde manier van lezen hebben. Zodoende zijn docenten zich niet bewust van de vele interpretaties die mogelijk zijn. Het volgende, aan de examenpraktijk ontleende voorbeeld, is in dit verband illustratief. De vraag luidde:

Waarom bepaalt het KNMI elk uur de temperatuur op een groot aantal meetpunten?

Bij de correctie van deze vraag signaleerden docenten diverse verschillende antwoorden. Deze antwoorden bleken afhankelijk te zijn van de wijze waarop, in dit geval, leerlingen de vraag hadden gelezen. Er was nadruk op:

- *waarom*: dat levert nuttige informatie op voor boeren en tuinders;
- *KNMI*: dat behoort tot de taken van dat instituut;
- *elk uur*: dat levert een grote betrouwbaarheid van meting op vergeleken bij een kleiner aantal meetmomenten;
- *temperatuur*: voor het doen van een goede weersvoorspelling zijn temperatuurgegevens onmisbaar;
- *een groot aantal meetpunten*: temperaturen kunnen van plaats tot plaats nogal verschillen.

Een *pretest* van deze vraag – bij een pretest legt de toetsconstructeur voorafgaande aan de ‘echte’ afname de vraag of hele toets voor aan een vergelijkbare groep – en vervolgens een bestudering van de antwoorden, had dit probleem kunnen voorkomen. Het *pretesten* van toetsvragen is echter een kostbare aangelegenheid en gebeurt daarom nauwelijks. Een (goedkoper) alternatief is een screening en beantwoording van de vragen door collega’s, ook collega’s van buiten het vakgebied. Het is steeds weer verrassend te ervaren hoe (anders) collega’s de vraag interpreteren. Laat collega’s de vragen beantwoorden, zonder dat zij beschikken over een modelantwoord. Verzamel hun antwoorden en eventuele op- en aanmerkingen. Gebruik deze informatie om de vraag (en het antwoord) bij te stellen.

### ➤ **Splits de vraag op in een informatiegedeelte en een vraaggedeelte.**

Het geschetste probleem had ook kunnen worden voorkomen door de vraag te splitsen in een informatiegedeelte en een vraaggedeelte. Zo zou de KNMI-vraag als volgt kunnen worden opgesplitst:

Het KNMI bepaalt elk uur de temperatuur van een groot aantal meetpunten.

- a) Geef één reden waarom dat elk uur gebeurt.
- b) Geef één reden waarom dat op een groot aantal meetpunten gebeurt.

### ➤ **Hanteer concrete opdrachten en vragen en vermijd formuleringen die het gewenste antwoord niet oproepen.**

Sommige vraagformuleringen leiden tot ongewenste dan wel ontwijkende antwoorden omdat ze onduidelijke antwoordinstructies bevatten. In al deze formuleringen ontbreekt een concrete opdracht. Voorbeelden van dergelijke formuleringen zijn:

*Probeer eens uit te leggen waarom...*

De vraag suggereert een poging maar het gaat om de uitleg.

*Kun je aangeven op welke manier...*

De vraag suggereert dat de studenten iets kunnen aangeven maar het gaat erom dat zij de manier aangeven.

*Zal de maatregel leiden tot een stijging of een daling...*

Studenten volstaan met het antwoord 'ja'; beide ontwikkelingen zijn immers mogelijk.

**► Kom tot afspraken voor wat betreft vraagformuleringen: gebruik een beperkt aantal vraagformuleringen en geef aan vraagformuleringen steeds een zelfde betekenis.**

In de praktijk worden voor hetzelfde type open vragen verschillende formuleringen gebruikt. Deze verschillende formuleringen kunnen een, niet beoogd, verschillend antwoordgedrag oproepen. En er kunnen zich problemen voordoen als bepaalde formuleringen niet steeds consequent worden gebruikt. Voor studenten is het prettig als zij bij een bepaalde opdracht- of vraagvorm bij alle vakken eenzelfde soort antwoord moeten geven. Het is daarom wenselijk in de opleiding een lijst van vraagformuleringen op te stellen. Door het hanteren van een dergelijke lijst ontstaat duidelijkheid over het per vraag vereiste antwoord.

Als voorbeeld van een dergelijke afspraak staat hieronder het overzicht van formuleringen dat bij de centrale examens vmbo, havo en vwo wordt gebruikt. De formuleringen zijn ook toepasbaar in het hoger onderwijs.

### Standaardformuleringen voor korte-antwoordvragen

Formulering	Eventuele aanwijzing met voorbeeld
Noem/Schrijf op	Te gebruiken bij een vraag naar iets concreets, zoals termijn, plaats, gebeurtenis, kenmerk, aspect bijvoorbeeld: → Noem twee gasvormige afvalproducten van elektriciteitscentrales die, zonder tegenmaatregelen, de voornaamste bijdrage leveren aan de verzuring van het milieu.
Geef (aan)	Dit kan een korte- of lange-antwoordvraag zijn, moet dus nader gespecificeerd worden. Als korte-antwoordvraag te gebruiken voor concrete zaken. Voorbeelden: → Geef de formule van de prijselasticiteit van de vraag. → Geef één voorbeeld van een winkelketen die aan ... voldoet. → Geef aan of de volgende beweringen juist zijn of onjuist.
Hoe	Te gebruiken bij kwantitatieve vragen en vragen naar concrete zaken of verschijnselen, bijvoorbeeld: → Hoeveel % van de pakken zal minder dan 1 kg bevatten? → Hoe groot is de afstand van ... naar ... ? → Hoe vaak ...? Hoelang duurt ...? → In een winkel worden schappen op drie verschillende hoogten onderscheiden. Bukhoogte en grijphoogte zijn er twee van. Hoe heet de derde hoogte?
Wie	Te gebruiken bij reproductie- of identificatievragen, bijvoorbeeld: → Wie heeft het buskruit uitgevonden?
Wat	→ Wat is de betekenis van de afkorting AIVD? (Of: Waarvan is AIVD een afkorting?) → Wat wordt verstaan onder arbeidsproductiviteit?



### vervolg Standaardformuleringen voor korte-antwoordvragen

Formulering	Eventuele aanwijzing met voorbeeld
Waar	Te gebruiken bij een concrete plaatsaanduiding, bijvoorbeeld: → Waar heeft ... plaatsgevonden?
Welke	Te gebruiken bij een vraag naar een concreet object of concrete categorie, bijvoorbeeld: → Welke drie muziekinstrumenten hoor je achtereenvolgens? → In welke vier delen van Nederland is volgens de kaart de vervuiling door ammoniakuitstoot het grootst? → Voor welke waarde van ... geldt ...?
Wanneer	→ Wanneer vond de Tweede Wereldoorlog plaats?
Arceer	→ Arceer op het uitwerkblad het gebied dat onder water loopt als de dijk doorbreekt.
Journaliseer	Te gebruiken voor het debiteren en crediteren van bedragen in grootboekrekeningen, bijvoorbeeld: → X en Y brengen hun zaken bijeen in een VOF. Journaliseer de inbreng van X.
Citeer	→ Citeer het eerste en laatste woord van de zin waarin dat staat.

### Standaardformuleringen voor lange-antwoordvragen

Formulering	Eventuele aanwijzing met voorbeeld
Geef (aan)	Kan korte- of lange-antwoordvraag zijn, moet dus nader gespecificeerd worden. Als lange-antwoordvraag te gebruiken bij een vraag naar iets abstracts zoals reden, oorzaak. (Dus niet: waarom ...) Voorbeeld: → Geef de oorzaak voor het ontstaan van conjuncturele werkloosheid. → Geef een verklaring voor het ontstaan van ...
Geef de berekening van	Bij deze vraag moet ook de berekening zelf gegeven en beoordeeld worden. Voorbeeld: → Geef de berekening van de totale constante kosten per jaar.
Laat met een berekening zien	Te gebruiken wanneer de uitkomst van de berekening al in de vraag staat aangegeven, bijvoorbeeld: → Laat met een berekening zien dat de prijs € 10,25 is.
Stel op/samen	Te gebruiken bij het opstellen van vergelijkingen, bijvoorbeeld: → Stel de balans samen.
Leid af	Te gebruiken om op basis van verstrekte gegevens het gevraagde af te leiden, bijvoorbeeld: → Leid uit het model de hoogte van het nationale inkomen waarbij sprake is van inkomensevenwicht af.
Onderzoek	Te gebruiken voor het onderzoeken van een serie eigenschappen van een functie of voor beweringen, o.a. in de statistiek, Voorbeeld: → Onderzoek of de steekproef groot genoeg is om de vereiste nauwkeurigheid te bereiken.

vervolg Standaardformuleringen voor lange-antwoordvragen	
Formulering	Eventuele aanwijzing met voorbeeld
Bewijs	Te gebruiken wanneer op deductieve wijze bewezen moet worden dat een bewering waar is door deze te herleiden tot definities, stellingen of axioma's. Voorbeeld: → Bewijs dat de steekproef groot genoeg is om de vereiste nauwkeurigheid te bereiken.
Toon aan	Te gebruiken wanneer een bewering aannemelijk moet worden gemaakt zonder deductieve redenering, bijvoorbeeld: → Gegeven is .... Toon aan dat bij het inkomensevenwicht sprake is van een tekort op de overheidsbegroting.
Leg uit (waarom)	Te gebruiken wanneer een redenering of argumentatie gegeven moet worden die uit enkele afzonderlijke denkstappen bestaat, bijvoorbeeld: → Leg uit waarom het voor sommige bedrijven economisch gezien voordelig kan zijn om afval opnieuw te gebruiken.
Beschrijf	Bij 'beschrijf' moet een antwoordrestrictie worden opgenomen. In het volgende voorbeeld is de antwoordrestrictie: 'in een gesloten economie zonder overheid'. → Beschrijf de werking van de multiplier in een gesloten economie zonder overheid.
Beredeneer Beargumenteer Beoordeel/Geef een beoordeling Vat samen Geef een verklaring voor	Deze standaardformuleringen komen voor in de examens Nederlands, en worden vaak gevolgd door een aanwijzing over de lengte van het antwoord. → Beredeneer dat voorlichting over de gevaren van roken een averechts effect kan hebben. Gebruik maximaal 40 woorden. → De schrijver van fragment 2 maakt een verwijt naar de schrijver van fragment 1 ... Beoordeel of dit verwijt terecht of niet terecht is. Geef bij je standpunt één op de tekst gebaseerd argument. → Maak een samenvatting in correct Nederlands van maximaal 180 woorden van de tekst ... Zorg ervoor dat deze samenvatting begrijpelijk is voor iemand die de oorspronkelijke tekst niet kent.

➤ **Formuleer de vraag zodanig dat studenten de vraag taalkundig begrijpen. Laat de conceptvragen zo nodig beoordelen door ter zake deskundige collega's.**

Toetsconstructeurs moeten voorkomen dat studenten fouten maken uitsluitend als gevolg van te moeilijk of te vaag taalgebruik. Cultuurgebonden verwijzingen, (te) lange zinnen en teksten, figuurlijk taalgebruik en het gebruik van verwijswoorden leveren bij minder taalvaardige studenten problemen op. Taalgebruik moet echter ook weer niet zo worden versimpeld dat het inhoudelijke niveau van een examen omlaag gaat. In de onderwijspraktijk blijken bepaalde opgaven soms moeilijker te zijn voor allochtone dan voor autochtone studenten. Dit als gevolg van hun andere culturele en linguïstische achtergrond. Ook blijken opgaven soms moeilijker te zijn voor meisjes dan voor jongens als gevolg van hun andere interesses en achtergrondkennis. Het verschijnsel dat een verschil in moeilijkheidsgraad van een opgave voor verschillende groepen wordt veroorzaakt door aspecten van de opgave die niet relevant zijn voor wat de opgave beoogt te meten, wordt in de testliteratuur 'itembias' genoemd (Bügel & Sanders, 1998).

Allochtone studenten of studenten die niet erg taalvaardig zijn, mogen niet worden benadeeld door onnodig complex of al te zeer cultuurgebonden taalgebruik in een toets. Men

moet zich er altijd van bewust zijn dat er onbedoelde en ongewenste culturele aspecten kunnen worden getoetst. Dat gevaar dreigt vooral als er meerdere culturen in het spel zijn.

- **Gebruik contexten die aansluiten op het algemene niveau dat van de studenten mag worden verwacht. Let daarbij in het bijzonder op de aansluiting van het materiaal bij de voorkennis en het taalniveau van de doelgroep.**
- **Hanteer geen contexten (en kies geen voorbeelden) die overwegend inspelen op de belangstelling en ervaring van bepaalde groepen studenten.**

Bij veel open vragen worden vragen gesteld naar aanleiding van een context (situatie, tekst, afbeelding, cartoon, tekening, advertentie enzovoort). Die context moet functioneel zijn met betrekking tot de te toetsen doelen. De context moet inhoudelijk en qua vormgeving overeenkomen met teksten en afbeeldingen zoals men deze in werkelijkheid tegenkomt (binnen grafische mogelijkheden uiteraard). Het mag niet zo zijn dat de studenten de vraag net zo gemakkelijk kunnen beantwoorden zonder zich in de context te hebben verdiept. Tevens dient te worden voorkomen dat de context of het voorbeeld overwegend aansluit bij de voorkennis van een bepaalde groep studenten, bijvoorbeeld uitsluitend bij vrouwen of mannen. Ook wanneer toetsvragen overwegend aansluiten bij de voorkennis, ervaring en belangstelling van steeds dezelfde groep studenten, wordt deze groep bevoordeeld ten opzichte van andere.

## 12.5 Oorzaken voor verschillen in beoordeling

---

Bij open vragen wordt een zo objectief mogelijke beoordeling nagestreefd. Het is belangrijk dat beoordelaars de antwoorden nauwkeurig en steeds op gelijke wijze beoordelen. Dat verhoogt de betrouwbaarheid van de toets. Uit diverse onderzoeken blijkt dat er in praktijk grote verschillen bestaan tussen de wijze waarop de antwoorden op toetsvragen worden beoordeeld. De belangrijkste oorzaken hiervoor worden hieronder kort besproken:

### ■ Geringe intra- en interbeoordelaarsovereenstemming

De beoordelaar die hetzelfde antwoord op een later tijdstip opnieuw beoordeelt (bijvoorbeeld omdat studenten hebben geprotesteerd tegen de uitslag), komt nogal eens tot een ander oordeel. In een dergelijke situatie is er sprake van een geringe *intra*beoordelaarsovereenstemming. Van een geringe *inter*beoordelaarsovereenstemming is sprake indien twee of meer beoordelaars onafhankelijk van elkaar tot aanzienlijk verschillende beoordelingen komen.

Deze lage beoordelaarsovereenstemming kan ontstaan door allerlei beoordelaarseffecten, zoals normverschuiving, 'restriction of range', sequentie-effect, contaminatie-effect, halo- of horn-effect of een signifisch effect. De uitleg van deze beoordelaarseffecten staat in de bijlage van dit boek.

## 12.6 Oplossingen voor beoordelaarsverschillen

---

Het is een utopie te geloven dat er een situatie kan ontstaan waarin beoordelaars volledig met elkaar overeenstemmen. Tegen ongewenste beoordelaarseffecten zijn oplossingen voorhanden die het effect enigszins teniet kunnen doen:

- Om ervoor te zorgen dat de studenten van wie de vragen steeds als eerste worden nagekeken niet de dupe worden van de normverschuiving bij beoordelaars, is het raadzaam de volgorde van antwoordbladen regelmatig wijzigen.
- Het sequentie-effect wordt beperkt door het beoordelen van de antwoorden per vraag (in plaats van per student). Ook het wijzigen van de volgorde van antwoordbladen tijdens de beoordeling is hiervoor een oplossing.
- Het contaminatie-effect is moeilijk oplosbaar zolang er geen scheiding is aangebracht tussen doceren en beoordelen. Intercollegiaal uitwisselen van beoordelingsactiviteiten (docent A beoordeelt de toets van studenten van docent B, en andersom) biedt een oplossing.
- Het halo- en horn-effect wordt beperkt door de toets anoniem na te kijken. Studenten wordt gevraagd niet hun naam op de toets te zetten, maar hun ID-nummer. Het halo-effect kan eveneens worden gematigd door het inschakelen van een tweede beoordelaar, onafhankelijk van de eigen docent. Een dergelijk tweede beoordelaar beperkt ook het contaminatie-effect. Indien een buitenstaander als beoordelaar optreedt (in plaats van de eigen docent), doen beide genoemde effecten zich niet voor.

Soms zal het inzetten van een extra (tweede) beoordelaar een negatief effect verminderen. Het nadeel hiervan is, naast het kostenaspect, dat de voor de beoordeling benodigde tijd toeneemt. Of een tweede dan wel onafhankelijke beoordeling echt nodig is, hangt van de functie van de toets af. De beoordelaarsbetrouwbaarheid van toetsen op grond waarvan wordt bepaald of een student slaagt of zakt (dan wel of een student wel of niet wordt toegelaten tot een vervolgstudie) moet uiteraard hoger zijn dan die van een formatieve toets. Dus vooral bij toetsen waarvan de uitslag vergaande gevolgen heeft voor de student, is een betrouwbare uitspraak over een voldoende of onvoldoende van groot belang. Het is daarom aan te bevelen de toetsen met een toetsscore rond de zak/slaaggrens door een tweede beoordelaar te laten nakijken.

## 12.7 Checklist

---

Ter afronding van dit hoofdstuk een checklist die hanteerbaar is bij de controle van geformuleerde vragen, controle door de samensteller van de toets of door de collega die de toets screent. De genoemde punten zijn in dit hoofdstuk bij het onderdeel over de constructie-regels aan de orde gekomen of zijn een verdere uitwerking daarvan.

## Checklist voor het samenstellen, beoordelen en redigeren van open vragen

### Taalgebruik

- Is de vraag grammaticaal juist geformuleerd?
- Bevat de vraag ingewikkelde zinsconstructies?
- Bevat de vraag een dubbele ontkenning?
- Bevat de vraag onnodige tussenvoegsels?
- Is de vraag onnodig negatief gesteld?
- Kan de formulering van de vraag aanleiding geven tot misverstanden?
- Kan de vraag door een verschuiving van de klemtoon een andere betekenis krijgen?

### Informatie

- Bevat de vraag voldoende informatie voor beantwoording?
- Geeft de vraag voldoende informatie over de gewenste lengte en vorm van het antwoord?
- Is duidelijk dat een antwoord moet worden verklaard/toegelicht?
- Zijn informatie en probleemstelling duidelijk gescheiden?

### Relevantie

- Blijkt uit de vraag welke inhoud wordt getoetst?
- Kan de vraag met gebruik van een andere dan de beoogde kennis worden beantwoord?
- Suggereert de vraag een probleem dat niet aan de orde is? (strikvraag)
- Bevat de vraag onbedoelde hints voor de beantwoording? (testwise)
- Is de moeilijkheidsgraad van de vraag respectievelijk van de gehele toets acceptabel?
- Wordt de moeilijkheidsgraad van de vraag onnodig verhoogd door irrelevante gegevens?
- Is de toets representatief voor de doelstellingen?

### Context

- Is het gebruik van context (tekeningen, grafieken, teksten, afbeeldingen) functioneel?
- Is de context duidelijk en correct weergegeven?
- Bevat de context irrelevante informatie?

### Presentatie

- Zijn de vragen en de vraagonderdelen duidelijk van elkaar te onderscheiden?
- Is de nummering van de vragen logisch en overzichtelijk?
- Zijn de conventies wat betreft spelling, symboolgebruik, interpunctie en dergelijke, in acht genomen?
- Zijn de verwijzingen in de vraag naar teksten, tekeningen en dergelijke juist?

# Toetsen met korte casussen

*Lambert Schuwirth*

## Inleiding

Wat is het doel van toetsen met korte casussen?

Waarom geen lange casussen?

Wat maakt een korte casus geschikt?

Welke vragen passen goed bij casusgerichte toetsing?

Wat zijn de constructieregels?

**13.1 Korte casussen geschikter dan lange – 144**

**13.2 Construeren en beoordelen van een korte-casustoets – 145**

**13.3 Schrijven van casussen – 147**

**13.4 Tips bij het schrijven van casussen – 148**

13.4.1 Casusbeschrijving – 148

13.4.2 De vragen – 151

**13.5 Ten slotte – 154**

**13.6 Tips – 155**

Als mensen spreken over casustoetsen, bedoelen ze vaak lange, vertakte simulaties van de werkelijkheid. Een typisch voorbeeld van zo'n toetsvorm is de examinering van verkeerspilooten met behulp van simulatoren. De simulator is in staat om vele verschillende scenario's te genereren, waarin de piloot allerlei beslissingen moet nemen of handelingen moet verrichten. Iedere beslissing wordt gescoord en heeft gevolgen voor hoe het scenario verder gaat. Een dergelijke aanpak is misschien wel geschikt voor die opleiding, maar dat wil niet zeggen dat het voor iedere discipline de geschikteste toetsvorm is. Vooral voor die disciplines waarin het oplossen van probleemsituaties niet sterk geprotocolleerd of gestandaardiseerd kan worden, zoals geneeskunde, blijken lange simulaties minder zinvol te zijn dan korte casussen.

Dit hoofdstuk gaat over het construeren van toetsen die bestaan uit korte casussen met een zeer beperkt aantal vragen, en die gericht zijn op het nemen van essentiële beslissingen. Enkele nuanceringsen zijn hierbij echter op hun plaats. In dit hoofdstuk wordt bedoeld op formele en voornamelijk summatieve toetsvormen met als doel het probleemoplossend vermogen van studenten te toetsen. Dat lange casuïstiek in veel onderwijssituaties een plaats heeft als leermiddel en als feedbackmiddel, staat niet ter discussie. Voor *formele* examensituaties blijkt een toets met korte casussen echter beter. Ook moet worden gemeld dat de voorbeelden zijn gebaseerd op het medische domein; daar bestaat een rijkdom aan ervaring met en onderzoek naar de korte casus als toetsmethode. De geldigheid van wat in dit hoofdstuk aan bod komt, reikt echter veel verder en geldt ook voor natuurkunde, rechten, economie, enzovoort. De voorbeelden die worden gebruikt, zijn ook voor niet-medici duidelijk.

### 13.1 Korte casussen geschikter dan lange

Lange simulaties lijken de werkelijkheid beter te benaderen dan korte, en daarmee als toetsvorm authentiekere te zijn. Bij het huidige streven naar meer *authentic assessment* lijkt het gebruik van korte casussen 'tegenintuïtief' en een stap terug in de tijd. Er zijn echter goede redenen waarom korte casussen geschikter zijn dan lange, vertakte casussen, en de belangrijkste is dat ze leiden tot toetsen met een hogere betrouwbaarheid en validiteit.

#### ■ Toetsen met korte casussen hebben een hogere betrouwbaarheid

De betrouwbaarheid van een toetsuitslag geeft aan in welke mate de scores van studenten overeenkomen met de scores die zij op basis van hun competentie zouden moeten hebben. Met andere woorden, de betrouwbaarheid geeft aan in hoeverre toeval, geluk of pech een rol spelen bij het behalen van het toetsresultaat (zie ook ► hoofdstuk 2). Een belangrijke factor die een toets onbetrouwbaar maakt, is dat iedere toets slechts een (kleine) steekproef is uit een groot domein aan mogelijke vragen of items. Dit zou geen probleem opleveren als men er zonder meer van zou mogen uitgaan dat iemand die bijvoorbeeld goed presteert op een bepaalde casus, ook wel goed zal presteren op een willekeurige andere casus. Dat blijkt echter niet zo te zijn; kennis en probleemoplossend vermogen zijn erg domeinspecifiek, dat wil zeggen: hoe studenten presteren op één casus zegt weinig over hoe zij zullen presteren op een willekeurige andere casus. Om de rol van toeval in de steekproef zo veel

mogelijk te beperken, moet een toets zo veel mogelijk verschillende casussen bevatten. Per uur toetstijd kunnen meer korte casussen aan de orde komen dan lange, waardoor toetsen met korte casussen meer betrouwbare scores opleveren.

#### ■ Toetsen met korte casussen zijn meer valide

De validiteit van een toets geeft aan in hoeverre de toets meet wat bedoeld is te meten (zie ook ► hoofdstuk 2). Wanneer probleemoplossend vermogen (expertise) een belangrijk doel is, is het van belang vast te stellen of het *probleemoplossend vermogen* van een student ook daadwerkelijk wordt gemeten, en niet pure feitenkennis of toetshandigheid. Toetsen met korte casussen blijken expertise beter te voorspellen dan lange casussen. Dit hangt samen met de ontwikkeling van expertise en probleemoplossend vermogen. Uit vergelijkingen tussen experts en niet-experts komen enkele belangrijke aspecten naar voren:

- Experts zijn efficiënter dan niet-experts en slaan vaak stappen in het oplossingsproces over.
- Experts en niet-experts genereren ongeveer even veel en even snel hypotheses over de oplossing voor het probleem, maar die van experts zijn beduidend accurater.
- Experts verschillen met elkaar van mening over hoe een probleem het beste opgelost kan worden (het proces), maar zijn het veelal meer eens over wat de uiteindelijke oplossing (de uitkomst) zou moeten zijn. Dit wordt de idiosyncrasie van het oplossingsproces genoemd.
- Over het algemeen lossen experts problemen niet analytisch op; ze herkennen het probleem en daarmee ook de oplossing.

Toetsen met korte casussen sluiten beter aan op de essentie van succesvol probleemoplossen doordat deze – in tegenstelling tot toetsen met lange casussen – kort blijven en op de *uitkomst* of essentiële beslissingen focussen in plaats van op het proces; efficiëntie wordt daarmee beloofd (en niet bestraft).

Bovengenoemde aspecten zijn uitgebreid beschreven in de literatuur (Swanson, Norcini & Grosso, 1987; Elstein e.a., 1978). Verder zijn er veel cognitief-psychologische experimenten op het gebied van probleemoplossingsvermogen bekend die aantonen waarom een korte-casusaanpak beter werkt dan lange, vertakte simulaties (zie bijvoorbeeld Chi, Glaser & Rees, 1982).

Tegenwoordig wordt aan verschillende toetsinstituten steeds meer gewerkt met casussen volgens een *key-feature approach*. Hierbij wordt de casusbeschrijving kort, maar wel zo authentiek mogelijk gehouden. Er wordt slechts een beperkt aantal vragen gesteld die alle betrekking hebben op *essentiële beslissingen*. Het volgende deel van dit hoofdstuk gaat in op het schrijven en beoordelen van dergelijke toetsen.

## 13.2 Construeren en beoordelen van een korte-casustoets

#### ■ Blauwdruk of toetsmatrijs

Een belangrijk hulpmiddel bij het construeren van een korte-casustoets is het opzetten van een blauwdruk of matrijs waarin de toets inhoudelijk wordt ingedeeld. De blauwdruk



geeft per vakgebied of per onderdeel aan hoeveel casussen hierover zouden moeten handelen om een representatieve toets te krijgen. In een blauwdruk staat bijvoorbeeld welke onderwerpen in het onderwijs aan bod zijn gekomen, maar ook hoe frequent de verschillende problemen in de dagelijkse praktijk voorkomen. Vaak bestaat de neiging om een blauwdruk heel gedetailleerd op te zetten. Op die manier verdwijnt het bos echter achter de bomen. Een blauwdruk moet juist niet te gedetailleerd zijn, en alleen een grove indeling van de toets aangeven, om te voorkomen dat de nadruk te veel op één aspect komt te liggen. Een voorbeeld van een blauwdruk voor een casustoets over huisartsgeneeskunde ziet er als volgt uit:

categorieën:	in toets opnemen
algemeen	6%
spijsverteringsstelsel	2%
oog	7%
oor	5%
hart en vaatstelsel	11%
bewegingsapparaat	12%
enzovoort	57%
<i>totaal</i>	<i>100%</i>

#### ■ Verzamelen van casusmateriaal

Een volgende stap bestaat uit het verzamelen van materiaal voor casussen. Hierbij kan gebruik worden gemaakt van de ervaringen van mensen die in de praktijk werken. Door de onderwerpen, problemen of praktijksituaties die zij tegenkomen als uitgangspunt te nemen, komt de casus dicht bij de praktijk te liggen. De informatie voor casussen die op deze manier wordt verkregen, zal vaak niet direct bruikbaar zijn. Doorgaans moet de formulering of de precieze manier van vragen nog worden bijgeschaafd. Verderop in dit hoofdstuk staan hiervoor diverse tips beschreven.

#### ■ Scoren en beoordelen

Vervolgens moeten de antwoorden worden gescoord en beoordeeld. Het heeft de voorkeur dit eenvoudig te houden. Uit de literatuur blijkt keer op keer dat ingewikkelde scoringsmechanismen niet beter werken dan simpele '1-0' of 'goed-min-fout' scoringsystemen. Het is beter om veel energie te steken in de productie van goede casussen en goede vragen (op grond waarvan essentiële beslissingen kunnen worden genomen) dan in ingewikkelde scoringsalgoritmes.

Een goed geschreven casus vraagt van studenten dat zij de gegevens uit de casus met elkaar combineren en relateren aan de eigen kennis, waarschijnlijkheden tegen elkaar afwegen en zo tot een correcte beslissing komen. Het is daarom meestal onmogelijk om met één enkele literatuurreferentie als onderbouwing van de antwoordsleutel te volstaan.

Geadviseerd wordt om de casussen, vóór de toetsafname, voor te leggen aan collega's of aan panels van experts om te komen tot een validering van de antwoordsleutel. Pas als er voldoende consensus is over het juiste antwoord of de juiste beslissing, mag de vraag worden voorgelegd aan studenten. Als er geen consensus is, kan er een verdedigbare minderheidsmening bestaan. In dat geval is het het eenvoudigst om deelscores te gebruiken voor degenen die dit antwoord hebben gekozen. Wanneer er helemaal geen overeenstemming over de juiste beslissing bestaat, is de casus niet bruikbaar en moeten de experts met elkaar overleggen om de casus en/of vraag zo te veranderen dat wel voldoende overeenstemming kan worden bereikt.

### 13.3 Schrijven van casussen

#### ■ Een algemeen raamwerk voor casussen

Een casusgerichte vraag bestaat uit twee onderdelen, de casus en de vraag. De casus beschrijft de situatie of geeft studenten de informatie die moet worden gebruikt om het probleem op te lossen. Soms staat het probleem al in de casustekst:

Je bent huisarts. Je maakt een visite bij mevrouw Van Doorn (65 jaar oud). Ze heeft plotseling hevige pijn in haar buik gekregen. Ze vertelt je verder dat ... Je besluit lichamenlijk onderzoek te doen en je vindt ... Je vraagt je af of je haar moet insturen naar het ziekenhuis of dat je het voorlopig beter thuis kunt aanzien. Wat is de beste actie op dit moment?

- a. Je stuurt haar naar de SEH.
- b. Je verwijst haar naar de polikliniek van de internist.
- c. Je verwijst haar naar de polikliniek van de chirurg.
- d. Je verwijst haar naar de polikliniek van de gynaecoloog.
- e. Je wacht even af en ziet het beloop van de klachten aan.

Hoewel vaak wordt gedacht dat meerkeuzevragen niet geschikt zijn om hogere cognitieve vaardigheden (zoals probleemoplossend vermogen) te toetsen, is het toch niet zo zwart-wit (zie ook ► hoofdstuk 11). Uit onderzoek blijkt dat de *vraagvorm* er minder toe doet dan de *vraaginhoud*. Het is veel belangrijker *wat of welke beslissing* er wordt gevraagd dan of dit in een open of gesloten vraagvorm gebeurt.

In sommige casussen staat het probleem pas in de vraag:

Je bent huisarts. Je maakt een visite bij mevrouw Van Doorn (65 jaar oud). Ze heeft plotseling hevige pijn in haar buik gekregen. Ze vertelt je verder dat ... Je besluit lichamenlijk onderzoek te doen en je vindt ...  
Wat is de waarschijnlijkste diagnose?

De informatie in de casus is van vitaal belang. De casus moet in ieder geval voldoende informatie bevatten om het de student mogelijk te maken de vraag te beantwoorden. Te

veel, en vooral te veel irrelevante informatie is echter ook niet goed, omdat dat kan afleiden van waar het werkelijk om gaat. Verschillende soorten informatie kunnen in de casus worden beschreven. Bij een medische casus bijvoorbeeld niet alleen inhoudelijke informatie (klachten van de patiënt, bevindingen bij lichamelijk onderzoek, enzovoort) maar ook contextuele informatie (waar wordt de patiënt gezien: in de huisartsenpraktijk of in het ziekenhuis, is de patiënt getrouwd, wat voor werk doet de patiënt). Ook mag 'negatieve informatie' niet worden weggelaten. Dit zijn gegevens die wel zijn onderzocht, maar die niet afwijkend of opvallend waren.

Authenticiteit van de casusbeschrijving is essentieel. De informatie wordt bij voorkeur zo ruw en waarde vrij mogelijk aangeboden (er staat niet in dat de patiënt flinke koorts heeft, maar dat de temperatuur 39,3 °C bedraagt). Daar waar het nuttig is, wordt informatie visueel of akoestisch aangeboden (bijvoorbeeld door multimediegebruik bij computers).

Vragen zijn gericht op essentiële beslissingen. Algemeen gezegd zijn dat die beslissingen die bepalen of een casus succesvol afgehandeld kan worden of niet. Veelal is het zo dat als de essentiële beslissingen goed zijn genomen, de rest er vanzelf uit volgt. Bij een patientencasus hoeft de diagnose niet altijd de essentiële beslissing te zijn. In de eerste casus over mevrouw Van Doorn wordt bijvoorbeeld niet naar de diagnose, maar naar de acuutheid van het probleem gevraagd. In de huisartsensetting kan dat bij sommige problemen belangrijker zijn dan de diagnose.

## 13.4 Tips bij het schrijven van casussen

---

In het onderstaande staat een overzicht van tips en trucs bij het schrijven van een casus, bij het soort op te lossen problemen en bij de formulering van de vragen. De voorbeeldcasussen zijn enigszins karikaturaal van aard.

### 13.4.1 Casusbeschrijving

---

#### ➤ Gebruik de representatie van werkelijke situaties.

Er zijn diverse argumenten te geven om de representatie van werkelijke situaties te gebruiken bij het schrijven van casussen. Op de eerste plaats biedt de praktijk een rijke bron voor mogelijke casussen. Dat voorkomt dat docenten veel tijd kwijt zijn met het bedenken van onderwerpen of problemen voor casuïstiek.

Ten tweede biedt dit de mogelijkheid om onderwerpen min of meer aselekt te kiezen (in de geneeskunde bijvoorbeeld iedere vijfde patiënt uit de praktijk voor een casus selecteren of in een lerarenopleiding een willekeurig incident of interventie met een leerling). Dit voorkomt dat er alleen maar zeldzame dingen worden bevraagd, ook al zijn deze vaak wel indrukwekkend.

Ten slotte verdient het gebruik van echte simulaties de voorkeur, omdat tekstboekcasussen vaak kunstmatig aandoen en authenticiteit missen (zie ook ► hoofdstuk 9).

### ➤ Zorg ervoor dat de casusbeschrijving zo helder en concreet mogelijk is.

Als studenten de gegeven informatie moeten verwerken, is het nodig dat deze eenduidig en helder is. In het taalgebruik tijdens conversaties komen regelmatig (te) vage termen voor. Dit levert echter meestal geen problemen op, omdat de mimiek en intonatie van de gesprekpartners veel extra informatie geven en er bij onduidelijkheden nadere vragen kunnen worden gesteld. Bij geschreven casussen ontbreekt deze informatie, waardoor een casusbeschrijving concreter moet zijn dan gesproken woorden.

Mevrouw Van Dam komt bij je in de praktijk omdat ze last heeft van een vage pijn in de buik. De pijn is nu al enige tijd aanwezig en is krampend van aard. Ze vertelt je verder dat de pijn niet lijkt samen te hangen met de menstruatie, maar dat de menstruatie wel een week later was dan normaal. Als je haar onderzoekt, geeft ze een lichte drukpijn in de buik aan, en je voelt ook een mogelijke zwelling diep in de buik. Wat is de waarschijnlijke diagnose?

In deze casus is de gegeven informatie te onduidelijk: wat is ‘vage pijn’, wat is ‘enige tijd’ en wat is ‘een mogelijke zwelling’? In een gewone conversatie zouden dit soort opmerkingen al niet concreet genoeg zijn, laat staan in een schriftelijke casus. Verder ontbreekt er belangrijke informatie. Er staat bijvoorbeeld niet hoe oud de patiënte is, of ze een partner heeft, en of ze misschien zwanger zou kunnen zijn.

Het komt regelmatig voor dat vraagauteurs proberen om teksten van casussen zo kort mogelijk te houden en bang zijn om te veel woorden te gebruiken. Dit kan ten koste van de duidelijkheid gaan. Het kost studenten echter veel meer (denk- en overlees)tijd om een onduidelijke formulering te doorgronden dan om een langere, maar heldere tekst te lezen.

### ➤ Zorg voor voldoende inhoudelijke en contextuele informatie.

Niet alleen moet alle *benodigde* informatie in de casus zijn vermeld, deze moet ook zo veel mogelijk expliciet zijn. Vaak staan samenvattende termen in de casus, zoals ‘routineonderzoek’ voor een hele set aan testen, of er staan vakspecifieke afkortingen en jargon in. Het advies is de informatie in de casus zo concreet mogelijk te maken.

Bij het zoeken naar een oplossing voor een probleem, is het van belang te schetsen in welke context het probleem zich voordoet. Bij een medische casus betekent dit bijvoorbeeld dat duidelijk moet zijn in welke setting de arts de patiënt ziet, wat de rol van de arts is, wat de leeftijd van de patiënt is, wat de thuissituatie van de patiënt is. In een lerarenopleiding zullen bijvoorbeeld de setting, het soort school, de onderwijsbenadering van de instelling enzovoort, bekend moeten zijn.

Mevrouw Webers heeft sinds twee dagen last van een hernia. Tijdens het consult vertelt ze je dat ze geen last heeft van zwakte in de voet, en dat er geen problemen zijn met de ontlasting of urineren. Je doet een routine herniaonderzoek en besluit haar één week strikte bedrust voor te schrijven.

Dit is een correct besluit.

juist/onjuist

Deze casus bevat te weinig inhoudelijke informatie. Wat voor pijn heeft de patiënte? Straalt de pijn uit? Heeft ze dit ooit eerder gehad? Leeftijd? Dit zijn allemaal zaken die van belang zijn bij de afweging welke therapie het beste is. Een 'routine herniaonderzoek' is te vaag, omdat de student nu niet weet welke handelingen wel of niet zijn gedaan. Verder is onduidelijk wat de bevindingen ervan waren, en moet de student maar aannemen dat er afwijkingen waren die het voorschrijven van bedrust rechtvaardigen. Ten slotte weet de student niets van de thuissituatie (als contextueel gegeven), wat wel van belang is om een inschatting te kunnen maken van de 'therapietrouw' van de patiënte. Wanneer mevrouw Webers bijvoorbeeld alleenstaand is en drie jonge kinderen heeft, is de kans erg klein dat ze het advies van bedrust strikt opvolgt.

### ➤ **Zorg voor voldoende 'negatieve' informatie.**

Onder alle benodigde informatie die een casus moet bevatten, valt ook de 'negatieve' informatie, ofwel de informatie die geen bijzonderheden oplevert maar wel van belang is. Wanneer in het kader van een bepaald probleem naar informatie is gezocht en geen afwijkingen zijn gevonden, moet dit in de casusbeschrijving zijn vermeld. Wanneer een arts bijvoorbeeld onderzoekt of er drukpijn aanwezig is, moet vervolgens expliciet worden aangegeven of er wel of geen drukpijn is.

### ➤ **Presenteer de informatie ongeïnterpreteerd ('ruw').**

Om efficiëntieredenen staat informatie vaak in een geïnterpreteerde vorm. Ook bij het schrijven van een casus bestaat de neiging dit te doen. Het interpreteren van gegevens of het samenvoegen van afzonderlijke gegevens tot grotere informatiebrokken is echter voor studenten een belangrijke stap in het oplossen van problemen. Het is daarom beter om te beschrijven dat een patiënt een ademfrequentie heeft van 35/minuut dan dat deze een versnelde ademhaling heeft (een zogenaamde tachypneu).

### ➤ **Wees voorzichtig met het geven van afleidende informatie.**

Het gebruik van computers biedt allerlei mogelijkheden om visuele of akoestische informatie toe te voegen aan de casus. Het is echter raadzaam multimedia alleen te gebruiken daar waar het echt nodig is, dus alleen wanneer informatie ook in de werkelijkheid visueel of akoestisch moet worden geïnterpreteerd. Overdaad schaadt, en het gebruik van multimedia als larding bij een casus heeft het risico af te leiden van waar het werkelijk om gaat.

Dit geldt ook voor het presenteren van afleidende informatie in een casus. In de werkelijkheid moet bij het oplossen van een probleem ook altijd onderscheid worden gemaakt tussen relevante en niet-relevante informatie. Maar meestal zijn er dan wel bepaalde aanwijzingen die hierbij helpen. Bij een arts-patiëntcontact is bijvoorbeeld de non-verbale informatie van een patiënt van belang, iets wat in geschreven tekst moeilijk is te simuleren. Daarom wordt aangeraden voorzichtig om te gaan met het geven van afleidende informatie.

### ➤ **Vermijd problemen die niet in de dagelijkse praktijk bestaan.**

Vaak worden bij casustoetsen beperkingen ingebouwd die in de praktijk niet bestaan. Zo komt het voor dat studenten binnen een korte tijd een antwoord moeten geven op iets waarover zij in de praktijk gerust een dag mogen nadenken. Ook is het studenten bij

een examen vaak niet toegestaan boeken te gebruiken om snel iets op te zoeken, terwijl dit in de praktijk gebruikelijk is. Artsen kunnen bijvoorbeeld tijdens een consult even de dosering van een medicament opzoeken als zij dit niet meer weten. Het is daarom aan te raden studenten deze mogelijkheid bij een casusgericht examen niet te onthouden. Sterker nog, als het antwoord op een vraag daadwerkelijk snel is op te zoeken, dan is de casus doorgaans niet goed geschreven of de vraag niet gericht op het nemen van een essentiële beslissing.

### 13.4.2 De vragen

---

#### ➤ **Relateer de vraag direct aan de casus.**

De casus mag geen losse illustratie zijn bij een vraag, die dan vervolgens algemene kennisaspecten bevraagt; de zogenaamde kapstokcasus. Het is essentieel dat casus en vraag één geheel vormen, waarbij de vraag niet kan worden beantwoord zonder dat de casus is gelezen.

Mevrouw Smit heeft aanvallen van hevige krampende pijn in de bovenbuik. Als een dergelijke aanval optreedt, moet ze gaan rondlopen en kan ze niet stil blijven zitten. Ze heeft 5 tot 6 aanvallen per dag, nu al sinds 3 dagen. Een aanval duurt 4-5 minuten. Het treedt voornamelijk op nadat ze heeft gegeten, vooral bij vet eten. Haar temperatuur heeft ze opgenomen, die was 37,3 °C.

Bij lichamenlijk onderzoek vind je een lichte drupijnlijkheid in de rechter bovenbuik, maar verder geen afwijkingen.

Hoe groot is de kans dat als iemand galstenen heeft, deze bij een echo van de buik te zien zijn?

- a. 20%
- a. 40%
- b. 60%
- c. 80%

Dit is een duidelijk voorbeeld van een vraag die eenvoudig is te beantwoorden (voor wie de kennis heeft) zonder de casus te hebben gelezen. Ook een vraagstelling als: 'De kans dat *bij deze persoon* bij een echo van de buik galstenen te zien zijn' is niet ideaal. Weliswaar wordt hierbij al meer gerefereerd aan de casus, maar het blijft de vraag of een essentiële beslissing wordt gevraagd. De inschatting van die kans is namelijk in de praktijk een van de factoren die een rol spelen bij de beslissing om al dan niet een echo aan te vragen. Andere, ook belangrijke, overwegingen zijn: de tijd die ermee is gemoeid, de kosten, de belasting voor de patiënt, enzovoort. Idealiter sluiten casus en vraag zo op elkaar aan dat verschillende informatiedelen uit de casus met elkaar *moeten worden gecombineerd* om tot het juiste antwoord en de essentiële beslissing te komen.

#### ➤ **Richt de vraag zo veel mogelijk op essentiële beslissingen.**

Het bepalen van wat essentiële beslissingen zijn en wat niet, is wellicht het moeilijkste deel van casusgerichte toetsing. Hoewel dit in hoge mate een inhoudelijke discussie is die voor ieder vakgebied anders ligt, zijn drie algemene eisen aan te geven.

Ten eerste moet de vraag alleen kunnen worden beantwoord door combinatie van verschillende gegevens uit de casus. Een tweede eis is dat een casus antwoordgedrag ontlokt dat op de essentie ingaat. Soms is het voor een huisarts bijvoorbeeld niet echt noodzakelijk de precieze diagnose te weten, maar is het veel belangrijker om goed in te schatten of iemand onmiddellijk met een ambulance naar het ziekenhuis moet of dat de situatie thuis nog even kan worden aangezien. In een dergelijk geval is niet de vraag naar de diagnose het essentieelst, maar de vraag naar de spoed waarmee de patiënt naar het ziekenhuis moet worden gebracht. De derde eis betreft de mate van overeenstemming over het antwoord of de te nemen beslissing. Het is verstandig om te controleren of collega's het eens zijn met de keuze die is gemaakt ten aanzien van de beslissing die wordt gevraagd. Het advies is daarom simpel: vraag collega's of zij op grond van de casusbeschrijving dezelfde beslissing zouden bevragen.

Meneer Jansen (67 jaar) bezoekt je in je huisartsenpraktijk, omdat hij plotseling een scherpe pijn in zijn linkerschouder heeft gekregen. Hij heeft deze pijn nu 3 weken, maar heeft de hele tijd gedacht dat het wel over zou gaan. Nu wil hij toch graag dat je er eens naar kijkt. Bewegingen in de schouder blijken geen invloed te hebben op de pijn, ook zijn er geen duidelijke punten die drukpijnlijk zijn. De pijn straalt uit naar de rug, vooral naar het schouderblad. Verder vertelt hij je dat hij nog nooit echt ziek is geweest in zijn leven. Hij rookt sinds ongeveer 40 jaar ruim een pakje per dag en drinkt 1-2 glazen bier per dag. Bij lichamelijk onderzoek valt je niets afwijkends op. Wel zie je dat zijn linker-ooglid afhangt en dat zijn linkerpupil is vernauwd. Tot welk syndroom behoren de laatstgenoemde bevindingen?

In dit voorbeeld is niet essentieel om de naam van het syndroom te weten – iets waarvoor overigens alleen de laatste zin uit de casus voldoende is – maar om te herkennen dat deze patiënt mogelijk zeer ernstig ziek is. Deze vraag leidt dus niet tot het nemen van een essentiële beslissing en moet anders worden geformuleerd.

#### ➤ **Beperk het aantal beslissingen dat per casus wordt gevraagd.**

De wet van de verminderde meeropbrengst geldt sterk bij casustoetsing. De eerste vraag bij een casus geeft relatief veel informatie over de competentie van de kandidaat, de tweede wat minder, de derde nog minder, enzovoort. Het al eerder genoemde probleem van de domeinspecificiteit maakt duidelijk dat het zinvoller is één of slechts enkele vragen per casus te stellen en daarna door te gaan naar een andere casus, dan om één casus helemaal uit te putten.

#### ➤ **Formuleer de vraag zo helder mogelijk.**

De regels die gelden voor allerlei vraagvormen (zie ► hoofdstukken 11 en 12) zijn natuurlijk ook van toepassing op vragen bij casusgerichte toetsing. Elke vraag moet daarom worden gecontroleerd op de meest voorkomende constructiefouten. Hierbij moet zo veel mogelijk in de gaten worden gehouden of een vraag correct wordt beantwoord door een student die

de stof beheerst, en niet, of fout, wordt beantwoord door iemand die het niet, of verkeerd, weet. Alle keren dat iemand zonder kennis de vraag toch goed heeft of dat iemand met voldoende kennis de vraag fout heeft, gelden als een verstoring van de geldigheid (validiteit) van de toets (zie ► hoofdstukken 11 en 12 voor constructieregels).

### ► Richt de vraag precies op de te bevragen beslissing.

Dit lijkt misschien iets wat al eerder hier aan de orde is geweest, maar er is een subtiel verschil. De vraag kan zo worden gesteld dat het niet duidelijk wordt welke beslissing nu precies is bedoeld. Een voorbeeld:

Theo (4 jaar oud) heeft sinds gisteren koorts. Zijn ouders hadden de temperatuur opgenomen, en deze bleek 39,3 °C te zijn ('s avonds gemeten). Zijn ouders zijn erg ongerust: 'Zo'n hoge koorts voor zo'n klein kind.' Ook hoest hij er behoorlijk bij; een vieze, droge hoest waarvan hij soms een beetje buiten adem raakt. De huisarts onderzoekt het kind en komt tot de conclusie dat er niets anders aan de hand is dan een gewone verkoudheid. Hij geeft de ouders het advies Theo een paar dagen wat meer binnen te houden. Zo nodig mogen ze hem een tabletje paracetamol geven om de koorts wat te drukken. Is dit een correct beleid?

- a. ja
- b. nee

Los van het punt of deze vraag nu de essentie betreft van deze casus, is de vraag slordig gesteld. Beide antwoorden zijn te verdedigen. Wat de huisarts hier doet is niet fout, en meer doen is daarom ook niet nodig. Wel staat nog aangegeven dat de ouders erg ongerust zijn. De huisarts had er dus verstandig aan gedaan om de ouders gerust te stellen, bijvoorbeeld omdat zij anders geen goede nachtrust zouden hebben. Kortom, de beslissing van de arts was niet fout, maar er had meer bij gekund. Als de essentie van de casus is eventuele fouten in de beslissing op het spoor te komen, dan moet de vraag daarop zijn gericht. Een mogelijke oplossing is om de casus uit te breiden; er kan vervolgens een verkeerde stap in het beleid worden beschreven met als vraag welk aspect van het handelen van de arts verkeerd was.

### ► Zorg ervoor dat het bedoelde antwoord verdedigbaar correct is, en dat 'foute' antwoorden ook echt fout zijn.

Het lijkt een open deur dat een 'fout' antwoord daadwerkelijk fout moet zijn, maar dat is het niet altijd. Vaak worden vragen zo geformuleerd dat de grenzen tussen foute en goede antwoorden erg vaag zijn.

Mevrouw Witteman (45 jaar) bezoekt je huisartsenspreekuur. Ze vertelt je dat ze forse pijn in de buik heeft. De pijn heeft ze nu al een week, en het lijkt wel of het erger wordt.

Welke vragen stel je?



In deze vraagstelling staan twee fouten. Ten eerste wordt gevraagd wat de student zou vragen aan de patiënt (met de impliciete bedoeling dat de student aangeeft wat een correcte beslissing zou zijn), in plaats van dat direct wordt gevraagd welke actie op dit moment correct of aangewezen zou zijn. Een tweede en groter probleem is dat er geen echt fout antwoord kan worden gegeven. In principe kan iedere vraag worden gesteld die niet beledigend of choquerend is. Vragen die geen relevante informatie opleveren, hoeven geen grote schade op te leveren. Het wordt een heel ander verhaal wanneer zou worden gevraagd welke vragen in dit geval de grootste kans hebben relevante informatie op te leveren. Er kan ook worden gekozen voor een wat kunstmatige oplossing, namelijk om een beperking in te voeren: ‘Stel dat je maar vijf vragen zou mogen stellen, welke vijf zijn dan het meest aangewezen?’

➤ **Laat de inhoud van de vraag de vraagvorm definiëren in plaats van andersom.**

Het is verleidelijk om een bepaalde vraagvorm te gebruiken, en aan te nemen dat deze in alle gevallen het beste is. Feit is echter dat de meeste studies aantonen dat de vraagvorm er weinig toe doet: de inhoud is van meer belang. Elke vraagvorm heeft zijn voor- en nadelen. Sommige inhouds laten zich beter in een gesloten-vraagvorm gieten, andere weer beter in een open-vraagvorm. Als de mogelijkheid in logistieke zin bestaat, verdient het zonder meer aanbeveling om de vraagvorm te laten afhangen van het aantal reële alternatieven dat er in de praktijk bestaat. Bij de beslissing of iemand naar de SEH moet of dat nog even thuis kan worden afgewacht, zijn er in feite maar twee reële opties. Het hanteren van een meerkeuzevraag ligt dan voor de hand. Indien het er bij een ziektegeschiedenis echter expliciet om gaat dat de kandidaat *spontaan* aan een bepaalde aandoening denkt, is een gesloten vraag vrij zinloos. De keuze om één bepaalde vorm (meestal meerkeuzevragen) te gebruiken, kan zinvol zijn uit logistieke overwegingen, maar het argument dat een bepaalde vorm in alle gevallen superieur is aan andere, is inhoudelijk niet houdbaar.

### 13.5 Ten slotte

Het gebruik van toetsen met korte casussen is een betrouwbare en valide methode om het probleemoplossingsvermogen van studenten te toetsen. Het vereist wel een zorgvuldige kwaliteitscontrole van het toetsmateriaal. Er moet vooral voldoende aandacht worden besteed aan de selectie van de casuïstiek in een toets, aan zorgvuldige formulering van casustekst en vragen, en aan een zorgvuldige validering van de antwoordsleutel. Het schrijven van casustoetsen is geen eenvoudig proces en vereist meestal de inzet van meerdere personen die kritisch naar de teksten kijken. Indien er constructiefouten in een casus worden gesignaleerd, is dat geen teken dat de vakinhoudelijke expertise van een docent niet goed zou zijn; het schrijven van casussen is een kunde die moet worden geleerd.

Het construeren van toetsmateriaal in het algemeen en van korte casussen voor casusgerichte toetsing is een dure en tijdrovende zaak. Een gemiddelde tijdsduur van twee tot drie uur voor het maken, beoordelen en reviseren van één goede casus is geen uitzondering. Maar de ervaring leert ook dat ervaren vragenmakers gemakkelijk vier tot vijf casussen per uur kunnen produceren. Daarom luidt het advies te zoeken naar mogelijkheden

voor samenwerking met andere opleidingen of instituten. Dit levert bijna altijd een win-winsituatie op, waarbij ieder van de partners slechts een deel van het werk hoeft te doen om de beschikking te krijgen over een volledige set casussen.

## 13.6 Tips

### Tips voor het toetsen met casussen

#### *Tips voor korte-casustoetsen in het algemeen*

- Maak bij het construeren van een korte-casustoets gebruik van een blauwdruk.
- Gebruik de ervaringen van mensen uit de praktijk voor het verzamelen van materiaal voor casussen.
- Leg de casus vóór afname onder studenten ter controle voor aan collega's of aan een panel van experts.
- Verander de casus en/of vraag wanneer er onvoldoende overeenstemming over de correcte beslissing tussen collega's bestaat.
- Houd het scoringsmechanisme simpel.
- Werk voor het maken van casustoetsen samen met andere opleidingen of instituten.

#### *Tips voor het schrijven van de casus*

- Gebruik de representatie van werkelijke situaties.
- Zorg ervoor dat de casusbeschrijving zo helder en concreet mogelijk is.
- Zorg voor voldoende inhoudelijke en contextuele informatie.
- Zorg voor voldoende 'negatieve' informatie.
- Presenteer de informatie ongeïnterpreteerd (ruw).
- Wees voorzichtig met het geven van afleidende informatie.
- Vermijd problemen die ook niet in de dagelijkse praktijk bestaan.

#### *Tips voor het schrijven van de vragen*

- Relateer de vraag direct aan de casus.
- Richt de vraag zo veel mogelijk op essentiële beslissingen.
- Beperk het aantal beslissingen dat per casus wordt bevestigd.
- Formuleer de vraag zo helder mogelijk.
- Richt de vraag precies op de te bevragen beslissing.
- Laat de inhoud van de vraag de vorm definiëren in plaats van andersom.
- Zorg ervoor dat het bedoelde antwoord verdedigbaar correct is, en dat 'foute' antwoorden ook echt fout zijn.

# Toetsen met een mondelinge toets

*Henk van Berkel en Anneke Bax*

## Inleiding

Wanneer valt de keuze op een mondelinge toets?

Waarom moet een mondelinge toets minimaal voldoen?

Welke functies vervult een mondelinge toets?

Wat kan er wel en wat kan er niet mee worden getoetst?

Hoe kan een mondelinge toets betrouwbaar en valide zijn?

Wat zijn de voor- en nadelen van een mondelinge toets?

Hoeveel docenten moeten aanwezig zijn?

## 14.1 Overwegingen voor mondelinge toetsen – 158

## 14.2 Voordelen – 159

## 14.3 Kwaliteitseisen voor de beoordeling en haalbaarheid – 160

14.3.1 Inhoudsvaliditeit – 160

14.3.2 Betrouwbaarheid en objectiviteit – 161

14.3.3 Transparantie – 164

14.3.4 Bruikbaarheid – 165

## 14.4 Beroepsmogelijkheden – 165

14.4.1 Oplossingen – 166

## 14.5 Conclusies – 167

## 14.6 Tips – 168

Het komt in het hoger onderwijs regelmatig voor dat tentamens mondeling worden afgenomen, vooral in opleidingen met weinig studenten. Het is voor docenten vaak meer werk om een goede schriftelijke toets op te stellen en na te kijken, dan om het geringe aantal studenten dat de opleiding volgt een mondeling tentamen af te nemen. Meestal heeft een docent dan snel in de gaten of de student de stof kent en heeft begrepen. Uit enkele vragen kan bijvoorbeeld blijken dat studenten alleen de hoofdlijnen hebben geleerd en de stof eigenlijk onvoldoende beheersen. Maar dat is slechts één kant van de zaak: het geven van een oordeel. De andere kant is dat beoordelaars hun oordeel 'hard' moeten kunnen maken. Bij de accreditatie zal zeker naar een verantwoording worden gevraagd: toetsen en toetsuitslagen moeten bijvoorbeeld transparant en eerlijk zijn.

Naast het mondeling toetsen zijn er ook andere situaties waarin de mondelinge communicatie een rol speelt bij de beoordeling van studenten. Bij toetsvormen als (groeps)projecten, (groeps)opdrachten en portfolio's vormt de mondelinge communicatie een essentieel onderdeel van de te toetsen competenties. In dit hoofdstuk gaat het over het mondeling tentamineren; een ander hoofdstuk, ► hoofdstuk 16, behandelt het toetsen van mondelinge presentaties.

## 14.1 Overwegingen voor mondeling toetsen

Dit hoofdstuk vat een mondelinge toets op als een situatie waarin twee docenten (tevens beoordelaars) vragen stellen aan één student en de student deze vragen beantwoordt. De docenten proberen op deze wijze na te gaan of de student voldoende competent is om te kunnen slagen voor de toets. De toets heeft daarmee een summatieve functie. Over het algemeen is het niet erg doelmatig om een mondelinge toets te hanteren als een formatief instrument en op die manier het toekomstig studiegedrag te sturen. Weliswaar is het mogelijk om er via bevraging achter te komen van welke onderwerpen studenten nog te weinig weten, en op die manier het toekomstige studiegedrag te sturen (zie de toetsdialoog in ► hoofdstuk 8), maar een schriftelijke toets bestaande uit vragen die de hele leerstof dekken, kan deze functie beter vervullen dan een mondelinge waarbij slechts een beperkt aantal onderwerpen aan bod kan komen.

In ► hoofdstuk 1 is gesteld dat de keuze voor een toetsvorm afhankelijk is van de doelstellingen van het onderwijs dat men wil toetsen. Wanneer het doel elementen bevat als 'het mondeling kunnen uitleggen ...', ligt het voor de hand een mondelinge toetsvorm te kiezen. Maar zo helder is het vaak niet. Er zijn onderwijssituaties waarin de doelen geen mondelinge toets impliceren en waarin docenten er toch toe besluiten. De volgende overwegingen spelen daar doorgaans een rol bij.

### ■ Doelmatigheid

In situaties waarin zeer weinig studenten (vijf of minder) moeten worden getoetst, lijkt het uit doelmatigheidsoverwegingen aantrekkelijk om mondeling te toetsen. De tijdsinvestering voor docenten is dan doorgaans niet meer dan de tijd voor het gesprek, welke in geen verhouding staat tegen de investering in een schriftelijke toets (het construeren van

vragen, het samenstellen van de toets, het organiseren van tijdstip en ruimte, het nakijken van de antwoorden en de kwaliteitscontrole).

- **Specifieke verlangens van de toetsdeelnemers**

Voor sommige studenten geeft het lezen en/of schrijven van teksten problemen. Slechtzenden en dyslectici zijn bekende voorbeelden. Maar ook studenten die de Nederlandse taal slecht, of niet, machtig zijn, komen in problemen wanneer zij een toets in het Nederlands moeten maken. Een mondelinge toets is voor deze studenten vaak een goede oplossing. Uiteraard geldt dat de doelstellingen zich moeten lenen voor de mondelinge toetsvorm.

- **'Genade'-behandeling**

Iedere opleiding kent ze wel, studenten die steeds maar weer zakken voor een bepaalde toets, of deze telkens hebben uitgesteld en inmiddels toch bijna aan het eind van hun studie zijn beland. De opleiding zit met dergelijke studenten in de maag. Het gaat er niet om dat deze studenten de studie niet aankunnen – ze hebben niet voor niets de eindstreep bijna gehaald – maar kennelijk zijn de toetsen voor het betreffende vak voor de student om een of andere reden onoverkomelijk. De verantwoordelijke docent wordt dan soms onder druk gezet om ervoor te zorgen dat de student slaagt. Een andere toetsvorm – bijvoorbeeld een mondelinge toets – kan in zo'n geval uitkomst bieden, ook al is dit niet de ideaalste oplossing.

Deze overwegingen rechtvaardigen niet altijd een mondelinge toets; tenslotte was in de normale toetssituatie een andere toetsvorm geïndiceerd. Een mondelinge toets heeft echter ook (onderwijskundige) voordelen heeft.

## 14.2 Voordelen

---

- **(Her)formuleren**

Bij een mondeling is er directe communicatie tussen docenten en student. Studenten kunnen bijvoorbeeld aangeven dat ze de vraag niet begrijpen. Soms kunnen docenten dat zelf afleiden uit de reactie van de student. Docenten hebben dan het hulpmiddel om de vraag te herformuleren in bewoordingen die de student wel begrijpt.

- **Doorvragen**

Een voordeel van de interactie is de mogelijkheid van docenten om het antwoord van studenten 'door te prikken'. Via doorvragen komen docenten erachter of studenten er werkelijk iets van begrijpen of dat ze slechts woordelijk herhalen wat in het boek staat.

- **Hoog cognitief niveau**

Het formuleren van vragen die verder gaan dan het bevragen van reproductiekennis is niet makkelijk. Een mondelinge toets biedt docenten de mogelijkheid, naast de hiervoor genoemde, na te gaan of de student de vraag begrijpt en aan de hand van het antwoord van de student te beoordelen of de bedoeling van de vraag correct is overgekomen. Is dat niet

het geval, dan kunnen docenten de bedoeling uitleggen zodat de student de mogelijkheid kan aangrijpen om het antwoord het gewenste niveau te geven.

- **Geen beroep op schriftelijke vaardigheden**

Het feit dat er geen beroep wordt gedaan op schriftelijke vaardigheden van de student, maakt de mondelinge toetsvorm geschikt voor specifieke groepen studenten, bijvoorbeeld motorisch gehandicapten, dyslectici of anderstaligen.

- **Studiegedrag**

Het valt niet te ontkennen dat van een mondeling een bijzondere uitstraling uitgaat. Studenten weten dat de toets doorgaans niet over feitjes of rijtjes gaat. Dat vergt een andere toetsvoorbereiding. Zij zullen zich wat meer moeten voorbereiden op vragen die begrip en inzicht toetsen. Onderwijskundig gezien heeft dat voordelen: kennis die op deze wijze is vergaard, vergeet de student minder snel.

- **Feedback**

Hoewel de feedbackmogelijkheden op het eerste gezicht bij een mondelinge toets afwezig zijn – er is immers meestal geen sprake van een gedetailleerde uitslag aan de hand waarvan studenten hun sterke en zwakke plekken kunnen ontdekken – komen veel studenten terug van een mondeling met de verzuchting: ‘Ik ben wel gezakt, maar ik heb veel geleerd’. Slechts zelden krijgen studenten, gezien de massaliteit van het hoger onderwijs, de gelegenheid een soort leergesprek aan te gaan met docenten. Wanneer een student op een mondelinge toets een vraag niet (helemaal) correct beantwoordt, kunnen docenten het meestal niet laten het de student (nogmaals) uit te leggen.

### 14.3 Kwaliteitseisen voor de beoordeling en haalbaarheid

---

Indien docenten hebben gekozen voor het beoordelen met een mondelinge toets, zal ook deze toetsvorm moeten voldoen aan de algemene eisen die aan ieder toetsinstrument worden gesteld (zie ► hoofdstuk 2). In het onderstaande staat aangegeven in hoeverre de mondelinge toetsvorm hieraan voldoet of kan voldoen.

#### 14.3.1 Inhoudsvaliditeit

---

Inhoudsvaliditeit (of representativiteit) betreft de mate waarin de toetsvragen een juiste weerspiegeling zijn van de doelen van het te toetsen onderwijs. Een docent moet van iedere vraag kunnen aangeven op welke doelstelling deze betrekking heeft. Vervolgens moet de docent ervoor zorgen dat iedere doelstelling als gespreksonderwerp ook daadwerkelijk in de toets wordt bevraagd. In het algemeen geldt: hoe meer vragen in de toets, des te beter de toets de doelen weerspiegelt. Wanneer een docent besluit veel vragen in het gesprek te stellen, is het in beginsel mogelijk dat de toets als geheel representatief is ten opzichte van de doelen.

Waar schriftelijke toetsen vaak twee tot drie uur duren, is de tijdsduur van een mondelinge doorgaans niet meer dan één uur. Dat heeft onder andere te maken met het optreden van vermoeidheid. Het is voorstelbaar dat zowel docenten als student de toets na één uur intensieve bevraging willen beëindigen. Het is echter de vraag of er in die periode voldoende vragen en antwoorden aan bod zijn gekomen op grond waarvan de docenten beslissingen kunnen nemen. In tegenstelling tot schriftelijke vragen zijn mondelinge vragen vaak niet *to-the-point* gesteld, en beantwoord. Het gevaar daarbij is dan dat in een gesprek van één uur minder onderwerpen aan de orde komen dan in een schriftelijke toets. Dit vormt een bedreiging voor de inhoudsvaliditeit.

#### ■ Oplossingen

Zonder tekort te doen aan het beoordelende karakter van de mondelinge toets, kunnen docenten van tevoren enkele gespreksonderwerpen op papier zetten die tijdens de toets aan de orde komen. Belangrijk geachte onderwerpen krijgen meer aandacht dan minder belangrijke. De docenten moeten ervoor zorgen dat de vragen als geheel de doelen representeren. Om de representativiteit en vergelijkbaarheid over verschillende mondelinge toetsen heen te garanderen dienen docenten ook voor een mondelinge toets een toetsmatrix te maken. Daarbij geven zij aan hoeveel tijd aan elk onderwerp mag worden besteed om te voorkomen dat alle tijd op gaat aan de bespreking van een enkel onderwerp. Op deze wijze is de mondelinge toets representatief of inhoudsvalide.

Een andere oplossing is het aanleggen van een kaartenbak met vragen. Hieruit trekken studenten blindelings en willekeurig vragen die zij op het mondeling hardop lezen en proberen te beantwoorden. Blindelings, omdat ze niet alleen vragen mogen uitkiezen die ze kunnen beantwoorden, en willekeurig, om ervoor te zorgen dat de toetsvragen representatief zijn voor de doelen. Wil dat laatste lukken, dan dient de bak aanzienlijk meer vragen te bevatten over belangrijke onderwerpen. Deze methode van mondeling toetsen wordt ook wel de ‘hoge-hoedmethode’ genoemd.

### 14.3.2 Betrouwbaarheid en objectiviteit

---

De (on)betrouwbaarheid van een toets is afhankelijk van veel factoren. Toegesplitst op de mondelinge toetssituatie zijn de belangrijkste:

- objectiviteit van het oordeel
- vermoeidheid
- emotionele spanning
- externe omstandigheden (temperatuur, licht, ventilatie)
- duidelijkheid van de vragen
- interactie tussen student en docent ten gevolge van het verschil in persoonlijkheid, geslacht en/of ras
- lichamelijke gezondheid

Hoewel een aantal van de genoemde factoren zeker niet specifiek is voor een mondelinge toetssituatie, is de literatuur eensluidend over wat het grote probleem is van mondelinge

toetsen: de objectiviteit van het oordeel. De persoon van de beoordelaar, de docent dus, blijkt erg bepalend te zijn voor de waardering van de antwoorden en daardoor voor het toe te kennen cijfer. Op zichzelf is dat zo gek nog niet. Immers, docenten zijn de eerstverantwoordelijken voor de toets en daarmee ook voor het cijfer. Maar daar ligt het probleem niet. Het probleem is dat studenten die hetzelfde kennisniveau hebben, lang niet altijd een gelijk cijfer verkrijgen wanneer ze mondeling worden getoetst. Docenten hebben te maken met allerlei beoordelaareffecten (zie ► bijlage). Zo blijkt dat docenten die een mondelinge toets afnemen bij hun beoordeling op meer punten letten dan uitsluitend op het kennisniveau van studenten. Bijvoorbeeld op:

- de wijze waarop studenten zich presenteren;
- de wijze waarop studenten hun kennis onder woorden brengen;
- het geslacht;
- het uiterlijk.

Bij een toets gaat het er echter om te meten in hoeverre studenten de doelen hebben bereikt. Een toetsuitslag moet als het ware ‘waardevrij’ zijn. De genoemde punten hebben daar doorgaans niets mee te maken.

#### ■ **Waarom twee docenten?**

Het lijkt erg ondoelmatig: twee docenten. Moet dat echt? Weet een enkele docent er dan te weinig van af? Onwetendheid bij een docent is niet aan de orde. Een eis voor twee docenten heeft met andere redenen te maken. Net zoals in het voortgezet onderwijs moeten één op één gesprekken tussen docent en student (leerling) zo veel mogelijk worden vermeden. Tussen docent en student is een te groot machtsverschil. Het wordt in deze tijd als niet passend opgevat wanneer zo'n gesprek plaatsvindt achter gesloten deuren. Bovendien, en wellicht belangrijker, bieden twee inhoudsdeskundigen tijdens een mondelinge toets grote voordelen. Wanneer beide docenten zijn betrokken bij het geven van het cijfer, is dat oordeel haast per definitie objectiever dan wanneer één docent oordeelt. Bovendien is het zeer goed denkbaar dat twee docenten verschillende dingen opmerken tijdens het mondeling. Het uiteindelijk oordeel heeft dan betrekking op een breder geheel en is daarom meer valide. Een van de eisen aan iedere toets is dat studenten een reële mogelijkheid moeten hebben om tegen de uitslag in beroep te gaan. Wanneer de toets heeft plaats gevonden in een één op één situatie, zijn er verder geen getuigen aanwezig die de student kunnen ondersteunen. Een beroepszaak tegen de uitslag van een mondeling blijft echter problematisch, omdat het moeilijk voor te stellen is dat de twee docenten ‘elkaar afvallen’. Toch zijn hiervoor oplossingen.

De beide docenten kunnen vooraf een bepaalde rolverdeling afspreken, bijvoorbeeld ten aanzien van wie de gespreksleiding neemt, wie het meest gaat notuleren en wie waar op gaat letten. Na afloop moet er zeker een nabespreking plaatsvinden tussen de beide docenten om vragen te beantwoorden als: ‘Wat ging goed? Wat zou beter hebben gekund? Hadden we moeten ingrijpen in het betoog van de student, of juist niet?’

Bovendien oordelen verschillende docenten verschillend over niet-relevante punten, zodat het voor studenten veel uitmaakt door welke docent zij mondeling worden getoetst. Vandaar de eis van twee docenten. Toch schuilt hier een addertje onder het gras. Do-



centen in een mondelinge toets opereren namelijk niet onafhankelijk van elkaar en zien bijvoorbeeld elkaars reacties op de antwoorden. Twee beoordelaars is wel in het belang van de student in het geval deze in beroep zou willen gaan tegen de uitslag (zie ► verder) en, bovendien, twee beoordelaars komt de objectiviteit ten goede.

Doordat docenten de mogelijkheid hebben door te vragen, wat een voordeel is, is de mondelinge toets niet voor iedere student gelijk. Resultaten op schriftelijke toetsen zijn per toets met elkaar vergelijkbaar omdat iedere student dezelfde toets heeft gemaakt. De toetsresultaten van de hele groep studenten worden vaak gehanteerd bij het bepalen van de zak/slaaggrens. Indien docenten bijvoorbeeld voorafgaand aan de toets een grens hebben bepaald, maar bij toepassing ervan blijkt dat er naar hun idee te veel studenten zijn gezakt, passen zij de grens aan (zie ► hoofdstuk 7). Bij mondelinge toetsen kan dat niet omdat de uitslag meestal direct na de afname door docenten bekend wordt gemaakt. Docenten passen dan een puur absolute wijze van normering toe. Sommige docenten besluiten om de uitslagen pas na een reeks mondelinge toetsen te bepalen; zij zijn dus in principe in staat studenten te vergelijken en dit tot uiting te brengen in de uitslag. Maar de vraag is wat ze dan vergelijken. Iedere mondelinge toets is anders. Het vergelijken van de prestaties van studenten op mondelinge toetsen is daardoor niet mogelijk.

Een ander probleem dat zich kan voordoen bij mondelinge toetsen is het zogenaamde sequentie- en halo-effect. Wanneer docenten verschillende mondelinge toetsen achter elkaar afnemen, is het denkbaar dat een eerder mondeling van invloed is op de waardering van dat erna. Wanneer bijvoorbeeld een mondelinge toets naar grote tevredenheid van docenten verliep, en de volgende wat minder, kan het cijfer voor deze tweede toets oneerlijk laag zijn omdat het verwachtingspatroon hoog was. Andersom kan dit effect ook optreden.

#### ■ Oplossingen

De genoemde bedreigingen van de objectiviteit en betrouwbaarheid hebben zowel betrekking op de toetsomgeving als op de wijze van toetsen. Maatregelen op het gebied van de wijze van toetsen komen vooral tegemoet aan de psychometrische eisen. Een rustige toetsomgeving zorgt ervoor dat de student zo min mogelijk wordt afgeleid en zich maximaal kan richten op de formulering van de antwoorden. Ook die maatregelen komen ten goede aan de kwaliteit van de toets.

Gedacht kan worden aan:

- Een aparte ruimte voor mondelinge toetsen, waarin temperatuur, licht en ventilatie goed geregeld zijn en waarin bovendien de toetsafname niet wordt gestoord door binnenlopende personen of telefoontjes.
- De toets laten beginnen met een globaal overzicht van de onderwerpen die aan de orde komen. Dat heeft als voordeel dat docenten worden gedwongen de mondelinge toets inhoudelijk gezien planmatig te laten verlopen. Een bijkomend voordeel kan zijn dat het overzicht studenten geruststelt. Het geeft hun immers het idee dat de docenten niet zullen 'doorzagen' over een enkel onderwerp. Het overzicht kan echter ook een averechts effect hebben op het gemoed van studenten: er kunnen bijvoorbeeld onderwerpen bijzitten waar de student zich minder op heeft voorbereid.

- Aan studenten vragen welke onderwerpen zijzelf aan de orde willen stellen. Dat zullen uiteraard onderwerpen zijn waar ze veel van afweten. Als deze passen in het vooraf gemaakte overzicht van onderwerpen, is er veel voor te zeggen met dat onderwerp te beginnen om studenten op hun gemak te stellen – ook al vinden de docenten het geen belangrijk onderwerp. De docenten kunnen besluiten de antwoorden niet in het cijfer te betrekken.
- De toets met relatief eenvoudige vragen laten beginnen is een methode om studenten zo veel mogelijk op hun gemak te stellen.
- De toetsduur tot maximaal één uur beperken.
- De uitslag pas vaststellen na discussie hierover tussen de beide docenten.

Het afnemen van een mondelinge toets vergt specifieke deskundigheden van docenten: het hanteren van een specificatietabel, het duidelijk formuleren van vragen, het wegen van antwoorden, het alert zijn op ontwijkende antwoorden, het op hun gemak stellen van studenten, enzovoort. Een opleiding zou docenten hiervoor zo nodig moeten scholen. In ieder geval mag van de opleiding worden verwacht dat er richtlijnen zijn waaraan docenten zich bij mondeling toetsen moeten houden.

### 14.3.3 Transparantie

---

Transparantie van een mondelinge toets heeft betrekking op de (on)duidelijkheid bij studenten over wat ze kunnen verwachten. Een mondelinge toetsituatie is doorgaans minder gestandaardiseerd dan een schriftelijke. Studenten zullen veel vaker schriftelijke toetsen hebben gemaakt en weten daardoor wel ongeveer wat ze kunnen verwachten. Bij een mondelinge toets ligt dat anders; de studenten zijn waarschijnlijk onbekend met:

- de toetsomgeving (de kamer van een van de docenten);
- de wijze waarop docenten vragen stellen (bondig of breedvoerig), de mate van doorvragen (gaan docenten de diepte in of nemen zij genoeg met het eerst gegeven antwoord?), de inhoud van de vragen (hebben docenten stokpaardjes waarop zij altijd terugkomen?), de beoordeling van de antwoorden;
- de mate waarin docenten belang toekennen aan elementen in de toetsituatie die weliswaar minder betrekking hebben op de toets zelf, maar wellicht toch een rol spelen bij de beoordeling (kleding, taalgebruik, wellevendheid enzovoort).

Het zijn juist deze onbekende factoren die in het geruchtencircuit van een opleiding vaak sappige verhalen opleveren. Maar het is ook een indicatie voor een opleiding om er wat aan te doen. Als het de studenten voldoende duidelijk is wat van hen wordt verwacht, kunnen zij zich hierop voorbereiden. Is dat niet het geval, dan kan onzekerheid opkomen, waardoor de student onnodig zenuwachtig wordt of zelfs geblokkeerd raakt.

#### ■ Oplossingen

Om transparantie te waarborgen, is het noodzakelijk dat de opleiding de studenten informeert. Dat kan bijvoorbeeld met een toetsfolder, waarin een docent van tevoren precies

aangeeft waarover de toets gaat, wat voor soort vragen worden gesteld (enkele voorbeeldvragen illustreren dit nog beter), wat van de studenten wordt verwacht met betrekking tot hun kennis- en vaardigheidsniveau, bijvoorbeeld dat studenten blijk moeten geven kennis te bezitten van de hele leerstof, of dat onvoldoende kennis over één onderwerp kan worden gecompenseerd door meer dan voldoende kennis over een ander onderwerp, en wat de uitslag van de toets kan zijn (geslaagd, gedeeltelijk geslaagd, bepaalde onderwerpen overdoen, of gezakt: alles overdoen).

### 14.3.4 Bruikbaarheid

---

De bruikbaarheid van een toets is sterk gerelateerd aan de doelmatigheid. De beslissing voor een bepaalde toetsvorm zal doorgaans afhangen van praktische argumenten: is er een ander toetsmiddel voorhanden dat dezelfde functies kan vervullen, maar om uiteenlopende redenen te prefereren valt (sneller, goedkoper, meer valide, betrouwbaarder)? Het antwoord op die vraag valt meestal in het nadeel uit van de mondelinge toets, behalve uiteraard wanneer er geen alternatieven voorhanden zijn die de doelstellingen tevens kunnen toetsen. Mondelinge toetsen zijn tijdrovend, en daardoor kostbaar en kwalitatief vaak onder de maat; zonder extra inspanningen is kwaliteit althans niet gegarandeerd. Daardoor zijn mondelinge toetsen in feite alleen bruikbaar als er maar weinig studenten aan een studieonderdeel deelnemen. Bij meer dan vijf<sup>1</sup> te toetsen studenten slaat de balans om naar een schriftelijke toetsvorm; bij vijf studenten is de tijdsinvestering ongeveer gelijk. De oplossing om bij een groter aantal studenten de mondelinge toets in groepsvorm af te nemen, is sterk af te raden. Een dergelijke toets zal dan weliswaar doelmatiger zijn qua tijdsbesteding, maar de betrouwbaarheid neemt af naarmate er meer personen aan deelnemen. Bovendien is hierbij de inhoudsvaliditeit, waarbij de participerende studenten een goede en even grote kans krijgen om hun kennis te demonstreren op het gebied waarover de toets gaat, twijfelachtig.

#### ■ Oplossing

De bruikbaarheid van een mondelinge toets is beperkt en dit probleem is niet oplosbaar in een toetsituatie waarin vijf of meer studenten moeten worden getoetst. Als de doelstellingen een mondelinge toetsvorm indiceren en er geen alternatieven zijn, is het geen goede oplossing om meer studenten tegelijkertijd te toetsen, maar zit er niets anders op dan de tijdrovende organisatie ter hand te nemen en individuele mondelinge toetsen af te nemen.

## 14.4 Beroepsmogelijkheden

---

Wanneer studenten het oneens zijn met de uitslag van een toets, hebben zij het recht in beroep te gaan bij het College van Beroep voor de Examens. Bij een mondelinge toets waarin twee docenten en één student tegenover elkaar zitten en een gesprek voeren over

---

1 Overigens bestaat over dit aantal geen eensluidendheid. Ter Burg & Kempers-Warmerdam (2013) gaan uit van een breakeven point van zeventien deelnemers.

de bestudeerde onderwerpen, waarin de docenten de uitslag bepalen, waarin noch een schriftelijke verslaglegging plaatsvindt, noch getuigen (anderen dan de betrokken docenten) aanwezig zijn, is het instellen van beroep illusoir. Wat zich heeft afgespeeld in een dergelijke situatie valt onvoldoende te reconstrueren. Zo'n toetssituatie is zeer onwenselijk: voor de student, indien deze meent ten onrechte een onvoldoende te hebben gekregen, maar ook voor docenten, die niet kunnen aantonen dat de student terecht is gezakt (net zomin als de student kan aantonen een voldoende prestatie te hebben geleverd).

### 14.4.1 Oplossingen

Het basisprobleem bij het beroepsrecht naar aanleiding van een mondelinge toets is dat doorgaans bewijsstukken ontbreken. Oplossingen voor dit probleem moeten daarom worden gezocht in het aandragen van bewijsmateriaal. Bijvoorbeeld:

- De docenten nemen het gesprek met de student op. Misschien dat het plaatsen van een recorder tussen docenten en student op het eerste gezicht wat geforceerd overkomt, maar de aanwezigheid van het apparaat blijkt al gauw te worden vergeten.
- Docenten noteren tijdens de toets de gestelde vragen (dit kan ook vooraf gebeuren zodat de inhoudvaliditeit meer wordt gegarandeerd) en maken een korte samenvatting van de door de student gegeven antwoorden. Aan het eind van het gesprek lopen docenten en student deze schriftelijke notities samen door en de student parafeert ze (eventueel na correcties te hebben aangebracht). Vervolgens geven de docenten hun oordeel.
- Docenten laten de te toetsen studenten voorafgaand aan het mondeling een notitie inleveren waaruit moet blijken dat ze de studiestof ten minste op rudimentair niveau hebben begrepen. Bijvoorbeeld: een samenvatting van de bestudeerde stof, enkele dwarsverbanden tussen onderwerpen, inhoudelijke kritiekpunten, een recensie. Voorwaarden hierbij zijn wel dat de student een gangbare schrijftaal machtig is en de notitie wordt beoordeeld op punten die als basis dienen voor het gesprek en niet op bijvoorbeeld schrijfvaardigheid. Pas wanneer de notitie voldoende kwaliteit bezit, mag de student een mondelinge toets doen, waarbij de notitie als uitgangspunt dient. De student verdedigt dus min of meer het schriftelijke stuk. Het gegeven dat de notitie al van tevoren is beoordeeld, draagt er zorg voor dat de student praktisch gesproken is geslaagd voor de toets en dat de rol van de mondelinge verdediging er een is van differentiatie in de eindbeoordeling: voldoende, ruim voldoende, goed. Wanneer blijkt dat studenten totaal niet in staat zijn de notitie te verantwoorden c.q. te verdedigen – misschien omdat zij de notitie niet zelf hebben geschreven? – met inachtneming van de voor studenten niet alledaagse situatie met de daarmee soms gepaard gaande gespannenheid, is een onvoldoende beoordeling verantwoord. Natuurlijk moet er wel een schriftelijke of elektronische neerslag zijn van de mondelinge verdediging. In hoeverre hier echter nog sprake is van een mondelinge toets, is twijfelachtig.

Zie  tabel 14.1 voor een overzicht van het besprokene.

■ **Tabel 14.1** Toetseisen, haalbaarheid en oplossingen bij mondelinge toetsen.

Toetseis	Haalbaarheid	Oplossing(en)
inhouds-validiteit	moeilijk	<ul style="list-style-type: none"> <li>– maak vooraf een lijst met onderwerpen en zorg voor een spreiding van de onderwerpen over de verschillende doelstellingen</li> <li>– stel meer vragen over belangrijke onderwerpen</li> <li>– laat studenten uit een kaartenbak willekeurig en blindelings vragen trekken</li> </ul>
betrouwbaarheid en objectiviteit	moeilijk	<ul style="list-style-type: none"> <li>– neem de toets af in een daarvoor geëigende ruimte</li> <li>– laat studenten zelf ook met onderwerpen komen</li> <li>– stel makkelijke vragen aan het begin</li> <li>– volg een training over het afnemen van mondelinge toetsen</li> <li>– twee beoordelaars die overleggen over het cijfer</li> </ul>
transparantie	haalbaar	<ul style="list-style-type: none"> <li>– maak een toetsfolder</li> </ul>
bruikbaarheid	alleen bij kleine aantallen studenten	
beroeps-mogelijkheden	moeilijk	<ul style="list-style-type: none"> <li>– maak een geluidsopnamen van de toets</li> <li>– maak aantekeningen van de vragen en antwoorden, laat studenten die paraferen en geef dan pas het cijfer</li> <li>– laat studenten vooraf een schriftelijke notitie indienen</li> </ul>

## 14.5 Conclusies

Mondelinge toetsvormen komen regelmatig in het onderwijs voor. Een mondelinge toets is ondoelmatig wanneer het aantal te toetsen studenten meer dan vijf bedraagt en kan slechts in beperkte mate voldoen aan de eisen die men aan iedere toetsvorm moet stellen. Vooral de betrouwbaarheid en objectiviteit blijken een probleem te zijn. Hiervoor zijn wel oplossingen te bedenken, maar die zijn vaak zo gekunsteld (en tijdsintensief) dat in feite niet meer kan worden gesproken van een mondelinge toets. Bijvoorbeeld studenten voorafgaande aan het mondelinge een schriftelijke notitie laten inleveren om deze te ‘verdedigen’ tijdens de mondelinge toets.

Toch zullen er situaties zijn waarin zowel praktische als inhoudelijke argumenten zijn aan te voeren die docenten doen besluiten om tot een mondelinge toets over te gaan. In dit hoofdstuk is geprobeerd aan te geven welke problemen kunnen spelen bij mondelinge toetsen en hoe deze zo goed mogelijk zijn te ondervangen. De belangrijkste tips staan in navolgend schema.

## 14.6 Tips

### Tips voor het inrichten van een mondelinge toets

1. Stel de studenten vooraf schriftelijk op de hoogte van algemene zaken (bijvoorbeeld met een toetsfolder), met informatie over de tijdsduur, het gewenste soort antwoord, procedure, cijfervaststelling, en dergelijke.
2. Standaardiseer de toetssituatie door maatregelen te treffen om de factoren die onbedoeld de score bepalen zo veel mogelijk te reduceren.
3. Neem de toets af met twee docenten die beide het cijfer bepalen.
4. Zet voorafgaand aan de toets de vragen of onderwerpen op papier en zorg ervoor dat ze voldoende gespreid zijn over de studiestof, of laat studenten uit een kaartenbak vragen trekken.
5. Besteed aan belangrijke onderwerpen meer tijd.
6. Stel de studenten op hun gemak door hen in eerste instantie gemakkelijke vragen te stellen; laat studenten eventueel zelf onderwerpen aandragen.
7. Vraag niet door als de studenten een antwoord niet weten, maar ga over op een ander onderwerp.
8. Vraag door bij onduidelijke antwoorden of bij onvoldoende niveau van het antwoord.
9. Geef aan het eind, voor het bepalen van de uitslag, een samenvatting van het gesprek en vraag of de student het ermee eens is.
10. Maak korte aantekeningen van het gegeven antwoord en laat studenten die aan het eind paraferen voordat ze een cijfer krijgen.
11. Begin iedere mondelinge toets met een schone lei c.q. wees onbevooroordeeld.
12. Stel vast wanneer een antwoord afdoende is gegeven en maak een aantekening van de waardering vóórdat de volgende vraag wordt gesteld.
13. Oordeel uitsluitend over die uitspraken van de student die zijn gedaan tijdens de toets en die vallen binnen de doelen van het studieonderdeel.
14. Stel de vragen duidelijk en licht ze eventueel even toe.
15. Zorg dat de studenten antwoord geven op uw vraag, en er niet 'omheen kletsen'.

# Toetsen met voortgangstoetsen

*Arno Muijtjens en Wynand Wijnen*

## **Inleiding**

Waarom voortgangstoetsing?

Wat is de rol van de vakdocent?

Wat is te bereiken met voortgangstoetsing?

Wat zijn de voor- en nadelen?

Welke voorwaarden gelden voor succesvolle toepassing?

- 15.1 De rol van de docent bij voortgangstoetsing – 170
- 15.2 Een voorbeeld: de interuniversitaire Voortgangstoets Geneeskunde (iVGT) – 171
- 15.3 Mogelijkheden van voortgangstoetsing – 176
- 15.4 Voor- en nadelen – 177
- 15.5 Voorwaarden – 179
- 15.6 Ten slotte – 180

Een voortgangstoets is een veelomvattende toets die gericht is op cognitieve doelstellingen van het onderwijsprogramma, net zoals een eindexamen. Het verschil met een eindexamen is dat een voortgangstoets niet alleen wordt afgenomen bij studenten in het laatste jaar van de studie, maar bij alle studenten van alle jaargroepen. Een verschil in kennisniveau komt dan tot uitdrukking in het niveau van de toetsscores: de scores van oude-rejaarsstudenten zullen gemiddeld hoger liggen dan die van jongerejaars. De ontwikkeling (groei) in kennis over de studiejaar wordt dus gemeten. Om voortgang in de studie te kunnen meten, dient een voortgangstoets meerdere keren te worden afgenomen. Bij een volgende gelegenheid wordt weliswaar niet exact dezelfde toets gebruikt (vanwege het gevaar van herkenning van vragen), maar wel één die vergelijkbaar is qua inhoud en moeilijkheidsgraad. Latere scores van een en dezelfde student zullen als het goed is in het algemeen hoger zijn dan eerdere in verband met de in de loop van het curriculum vergaarde kennis. Voortgangstoetsing biedt de mogelijkheid om vast te stellen hoe ver een student qua kennisniveau verwijderd is van het einddoel, en ook welke kennisgroei (of gebrek aan kennisgroei) zich in een bepaalde periode heeft voorgedaan. De gerichtheid op de doelstellingen van de gehele opleiding en de daaruit voortvloeiende veelomvattendheid van de voortgangstoets maken het vrijwel ondoenlijk voor studenten om de toets specifiek voor te bereiden. Daardoor wordt toetsgericht studeren tegengegaan en een continue wijze van studeren bevorderd (Wijnen, 1991).

## 15.1 De rol van de docent bij voortgangstoetsing

---

De veelomvattendheid van de voortgangstoets heeft consequenties voor de rol van een vakdocent als toetsers. Bij vakgebonden toetsing is het doorgaans de docent die niet alleen het onderwijs verzorgt maar ook zorg draagt voor de toetsing van het betreffende onderdeel van het onderwijsprogramma. Vaak gaat dat om het gehele traject van toetsontwerp tot toetsanalyse (zie ► hoofdstuk 2). Vanwege het vakoverstijgende karakter van een voortgangstoets dient noodzakelijkerwijs een aantal van die activiteiten centraal te worden uitgevoerd. Individuele vakdocenten leveren hun bijdragen dan aan het gezamenlijke product – de voortgangstoets – in de eerste plaats door voor hun vak relevante vragen aan te leveren. De overige activiteiten worden vervolgens uitgevoerd door een instantie die is te karakteriseren als een centrale toetsorganisatie die wordt aangestuurd door een toetscommissie bestaande uit vakdocenten en toetsdeskundigen. Een groot deel van de verantwoordelijke taken van docenten op het gebied van toetsen is dus in de praktijk van voortgangstoetsen overgeheveld naar een centraal niveau. Docenten blijven primair verantwoordelijk voor het construeren van toetsvragen.



## 15.2 Een voorbeeld: de interuniversitaire Voortgangstoets Geneeskunde (iVGT)<sup>1</sup>

Toen in 1974 de Medische Faculteit Maastricht startte met probleemgestuurd onderwijs (PGO), bleken toetsing en onderwijsdoelen elkaar danig in de weg te staan. Het PGO beoogde zelfgestuurd, ontdekkend leren in onderwijsgroepen op basis van geïntegreerd studeren in plaats van vakgericht studiegedrag. De aan het eind van elk onderwijsblok afgenomen summatieve bloктоetsen bevorderden echter toetsgestuurd studeren en specifiek toetsgerichte voorbereiding. Het volgen van individuele leerpaden werd erdoor ontmoedigd en het ‘van buiten leren’ werd bevorderd ten koste van het verwerven van inzicht, begrip en beklijvende, functionele kennis. De oplossing werd gezocht in het doorbreken van de directe relatie tussen het onderwijsprogramma en de toetsing, wat uiteindelijk leidde tot de ontwikkeling van de Maastrichtse voortgangstoets (Wijnen & Van der Vleuten, 1985). Onafhankelijk van de ontwikkeling in Maastricht ontstond ook aan de University of Missouri het idee van een voortgangstoets (Arnold & Willoughby, 1990). Sindsdien heeft deze longitudinale op meerkeuzevragen gebaseerde toetsvorm zich wereldwijd verspreid (Freeman, van der Vleuten, Nouns & Ricketts, 2010; Wrigley, van der Vleuten, Freeman & Muijtjens, 2012).

### *Blauwdruk*

Viermaal per jaar wordt bij alle geneeskundestudenten een voortgangstoets bestaande uit tweehonderd meerkeuzevragen afgenomen.<sup>2</sup> Elke vraag hoort bij een bepaald onderwerp (categorie) en bij een vakgebied (discipline). Voorbeelden van categorieën zijn ademhalingsstelsel, spijsverteringsstelsel, preventieve gezondheidszorg; voorbeelden van disciplines zijn fysiologie, interne geneeskunde en medische psychologie. Bij kruising van de categorieën met disciplines ontstaat een matrix. In elke cel van die matrix staat het aantal vragen dat de toets voor de betreffende combinatie van categorie en discipline dient te bevatten. Deze matrix dient als blauwdruk voor elke nieuwe voortgangstoets. De verdeling van de vragen over de blauwdruk is als het ware een operationalisatie van de doelstellingen van een opleiding Geneeskunde. Door steeds dezelfde blauwdruk aan te houden wordt gewaarborgd dat de opeenvolgende toetsen qua inhoud vergelijkbaar blijven.

### *Constructieproces*

De vakdocenten construeren de toetsvragen die vervolgens worden verzameld in een vragenreservoir. Iedere vraag bevat een ‘weet niet’-optie (zie verder).

- 1 Sinds het academisch jaar 1999/2000 wordt in een samenwerkingsverband tussen de geneeskundefaculteiten van de universiteiten van Groningen, Nijmegen en Maastricht gezamenlijk een interuniversitaire Voortgangstoets Geneeskunde (iVGT) samengesteld en afgenomen bij alle studenten geneeskunde van de drie faculteiten. In 2006 is de geneeskundefaculteit van de universiteit van Leiden toegetreden tot dit samenwerkingsverband en sinds 2007 wordt de toets ook afgenomen bij geneeskundestudenten van de VU Amsterdam.
- 2 Voor september 2005 was de voortgangstoets samengesteld uit juist/onjuist-vragen. De overgang van juist/onjuist-vragen naar meerkeuzevragen werd gemotiveerd door overwegingen van meettechnische en inhoudelijke aard. Meerkeuzevragen leveren in het algemeen een meer betrouwbare score en een kleinere kans om het correcte antwoord te raden dan juist/onjuist-vragen. Daarnaast is deze vraagvorm flexibeler dan een juist/onjuist-vraag en biedt daardoor meer mogelijkheden om een probleem adequaat te bevragen.

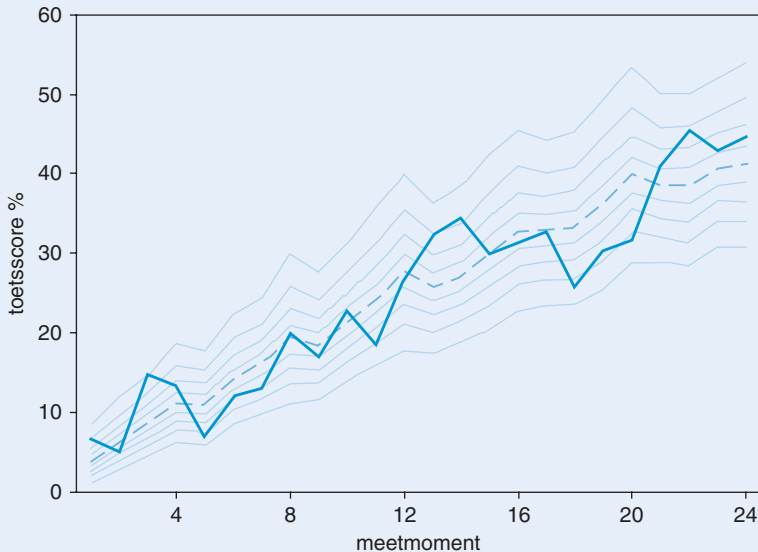
Een commissie beoordeelt de vragen op vorm en inhoud alvorens ze worden goedgekeurd voor opname in de toets. Niet-goedgekeurde vragen worden voorzien van commentaar en eventuele wijzigingsvoorstellen teruggestuurd naar de vraag-constructeur. Deze herconstrueert de vraag, eventueel na overleg met de commissie. De constructeur kan ook besluiten de vraag te laten vervallen. Goedgekeurde vragen worden in een vragenbank opgeslagen. Naast de nieuwe vragen bevat de vragenbank ook vragen die in een eerdere toets zijn gebruikt. Gebruikte vragen blijven een periode van drie jaar geblokkeerd, waarna ze weer selecteerbaar zijn voor een nieuwe toets. Geselecteerde gebruikte vragen worden altijd opnieuw door de commissie beoordeeld: nieuwe inzichten kunnen ertoe leiden dat de antwoordsleutel inmiddels niet meer correct is.

#### *Samenstelling toets*

De definitieve samenstelling van een hele toets is de verantwoordelijkheid van de commissie. De commissie ziet bij de toetsamenstelling toe op het voorkomen van eventuele hinderlijke relaties tussen vragen (geeft de informatie bij vraag x het antwoord weg op vraag y).

#### *Kwaliteitscontrole*

Bij elke toetsvraag in de vragenbank is opgenomen in welk boek (en op welke pagina) van de adviesboekenlijst het antwoord is terug te vinden. Na afloop van de toetsafname krijgen de studenten een formulier uitgereikt waarop de antwoordsleutel per vraag en de bijbehorende literatuurverwijzing staat. Aan de hand daarvan kunnen studenten hun score berekenen, nagaan waar de fouten zaten en verifiëren in de literatuur waarom. Mochten zij daarbij stuiten op vermeende onjuistheden of onduidelijkheden (bijvoorbeeld tegenstrijdige informatie in verschillende boeken van de adviesboekenlijst) dan melden zij die aan de toetscommissie. Aan de hand van deze studentcommentaren stelt de commissie vast of de antwoordsleutel van sommige vragen dient te worden gewijzigd en/of er vragen dienen te vervallen. Na afname van de toets vindt een itemanalyse plaats (zie ► hoofdstuk 2). Een hoog percentage weet-niet (bij zesdejaars) kan bijvoorbeeld een aanwijzing zijn dat de vraag onvoldoende relevant is of dat de vraag zo vaag geformuleerd is dat veel studenten denken deze niet te kunnen beantwoorden. Vragen met een hoog percentage fout zijn mogelijk misleidend geformuleerd, er is sprake van controversie in de literatuur of de antwoordsleutel is fout. Een lage item-testcorrelatie ten slotte houdt in dat de 'goede' studenten de vraag overwegend fout beantwoorden en 'zwakke' studenten juist relatief veel correcte antwoorden geven. Vragen die 'verdachte' waarden hebben voor een of meer van deze itemparameters worden, net als de op grond van studentcommentaar geselecteerde vragen, nader onderzocht door de toetscommissie. In deze nabewerkingsfase stelt de toetscommissie voor elk van de geselecteerde vragen vast of deze komt te vervallen, de sleutel gewijzigd wordt of dat de vraag ongewijzigd gehandhaafd blijft. In die fase komt de definitieve versie van de toets tot stand. Vervolgens worden de definitieve resultaten berekend en gerapporteerd aan studenten, docenten en onderwijsorganisatie.



■ **Figuur 15.1** Voortgangstoetsresultaten (percentage correct minus aftrek vanwege incorrecte antwoorden) bij de faculteit Geneeskunde van de Universiteit Maastricht in de jaren 2005/2006-2011/2012.

### Scoring en uitslag

De toets bestaat uit tweehonderd meerkeuzevragen met maximaal vijf alternatieven waarvan er één correct is. Om gokken te ontmoedigen en voor gokken te corrigeren is een 'weet niet'-optie toegevoegd en worden strafpunten in rekening gebracht bij een fout antwoord. Door te kiezen voor de 'weet niet'-optie kan een student deze 'boete' ontlopen. De score wordt berekend als een percentage van het totaal aantal vragen in de toets, waarbij een correct antwoord wordt beloond met +1, een fout antwoord wordt bestraft met  $-1/(\text{aantal alternatieven} - 1)$  en een 'weet-niet' levert 0 punten op. Deze scoringsmethode staat in de literatuur aangeduid als *formula scoring* (Lord, 1963; Rowley & Traub, 1977). Een belangrijke reden om deze methode toe te passen is gelegen in het feit dat veel jongerejaars een groot deel (40%-80%) van de vragen, dat gericht is op het eindniveau van de opleiding, niet kunnen beantwoorden. Zonder formula scoring zou het antwoord op deze vragen blind worden gegokt, met als gevolg het ontstaan van veel ruis en dus een onbetrouwbare meting.

De einduitslag voor een student wordt bepaald door de verkregen toetsscore te vergelijken met een of meerdere normen. Bij de iVGT worden per jaargroep twee normen gehanteerd (zie verder) en de einduitslag wordt verkregen in de vorm van een van de kwalificaties 'onvoldoende', 'voldoende' of 'goed'.


■ **Figuur 15.1** geeft een samenvatting van de voortgangstoetsresultaten voor de scores (percentage correct minus aftrek voor incorrecte antwoorden) behaald in zeven academische jaren (2005/2006-2011/2012). Met vier toetsen per jaar in een onderwijsprogramma van zes jaar is er sprake van 24 opeenvolgende meetmomenten. De toets in september resulteert in scores op 6 meetmomenten namelijk 1, 5, 9, 13, 17 en 21. De toets

in december levert scores op meetmomenten 2, 6, 10, 14, 18 en 22 enzovoort. Per meetmoment (1 tot en met 24) is er dus een verzameling scores beschikbaar afkomstig van zeven bijbehorende toetsen in de periode 2005/2006-2011/2012. Voor de verzameling scores per meetmoment zijn de percentielen (10, 20, ..., 90) bepaald; het tiende percentiel, bijvoorbeeld, is het niveau waarvoor geldt dat tien procent van de scores gelijk of lager is. Vervolgens zijn alle punten behorend bij eenzelfde percentiel door een lijn verbonden. De resulterende waaier van lijnen geeft aan waar de uitersten van de scoreverdeling zich bevinden (de breedte van de scoreverdeling). Daarnaast is ook aangegeven waar het centrum van de verdelingen zich bevindt. Dit wordt aangegeven door de onderbroken lijn die de mediaanscores van de verdelingen volgt: vijftig procent van de scores bevindt zich beneden de lijn en vijftig procent erboven. De onderbroken lijn is dus te beschouwen als de gemiddelde kennisgroei over de 24 meetmomenten.

Bij meetmoment 1 is het gemiddelde niveau gelijk aan vier procent en eindigt bij eenenveertig procent voor meetmoment 24. Duidelijk is dat de spreiding van de verdeling toeneemt met oplopend meetmoment. Dit komt ook terug in de standaarddeviatie van de scoreverdeling die bij meetmoment 1 gelijk is aan drie procent en oploopt op tot negen procent bij meetmoment 24. Dat geeft aan dat het kennisniveau van de studenten steeds meer uiteenloopt met het vorderen van de studie. Een gemiddelde toetsscore van veertig aan het eind van de studie lijkt wellicht op het eerste gezicht laag, maar indien wordt afgezien van de 'weet-niet'-optie, komt een score van veertig procent goed-fout overeenkomt met een score van zeventig procent correct bij gewone stellingvragen.

Er is ook een voorbeeld van een groeicurve van een individuele student ingetekend (dikke doorgetrokken lijn). Deze vertoont eveneens groei, maar laat een veel grilliger verloop zien dan de gemiddelde kennisgroei (onderbroken lijn). De onregelmatigheden in de individuele curve illustreren tekortkomingen in de kennismeting. De score op een toets wordt voor een (hopelijk groot) deel verklaard door wat je wilt meten: het kennisniveau van een student in het beoogde domein. De gemeten score wordt daarnaast echter ook beïnvloed door storende factoren (deze student weet toevallig meer over deze vragen, de vorm van de dag, enzovoort). Storende factoren zijn in het algemeen wisselend van aard per toetsafname en voegen daardoor grillige patronen toe aan de gemeten groei.

#### *Feedback*

Behalve de einduitslag krijgt elke student ook gedetailleerde feedback over het niveau van zijn/haar kennis op deelgebieden. Het percentage goed en fout beantwoorde vragen voor de subsets van vragen per categorie en per discipline staan in tabellen gepresenteerd.  Tabel 15.1 toont een voorbeeld van percentages goed en fout per categorie voor de individuele student en daarnaast, ter vergelijking, ook voor de gehele jaargroep.

Aangegeven is of studenten vergeleken met hun jaargroep een gelijke score behalen, een duidelijk lagere (– of --) dan wel een hogere (+ of ++).

De feedback biedt de student de mogelijkheid om een sterkte-zwakke analyse te maken door de eigen resultaten te vergelijken met de gemiddelde prestatie van de

**■ Tabel 15.1** Voorbeeld van de feedback die aan studenten geneeskunde van de Universiteit Maastricht wordt verstrekt.

Resultaten per categorie (procentueel)						ID-nummer: xxxxxx   Jaargroep: 6   Voortgangstoets (M11102): December 2012							
omschrijving	aantal vragen	individueel				studiejaargroep (n=279)							
		goed	fout	?	g-f	goed	std	fout	std	?	std	g-f	std
1 ademhalingsstelsel	14	79 ++	14 --	7	73 ++	53	14	37	14	10	12	40	18
2 bloed- en lymfstelsel	9	56 +	33	11	45 +	48	15	36	17	16	16	35	18
3 spier- en skeletst.	13	62	15 -	23	56 +	58	11	25	12	17	12	48	14
4 geestelijke gez. zorg	10	50	40	10	35	53	15	35	15	12	14	41	19
5 voortplantingsstelsel	11	64	27	9	53 +	56	16	28	14	16	15	42	21
6 hart- en vaatstelsel	15	47	47 ++	7 -	23 --	53	13	25	12	22	15	41	15
7 hormonen en metabol.	10	80 ++	10 --	10 -	77 ++	52	14	27	14	22	17	41	16
8 huid en bindweefsel	9	33 --	33	33 ++	22 -	50	13	35	16	15	14	38	17
9 pers. en mtsch. aspecten	10	20 --	40 +	40 ++	7 --	50	13	29	13	21	15	37	17
10 spijsverteringsstelsel	12	67	17	17 +	60	69	15	22	13	9	11	61	19
11 nieren en urinewegen	14	36 -	29	36 +	23	44	15	32	16	24	18	31	19
12 zenuwst. en zintuigen	19	47	37	16	34	48	13	35	12	17	14	35	15
13 mol. en cellulaire aspecten	9	56 +	22 -	22	48 ++	41	15	34	18	26	20	29	18
14 wetenschapsleer en methode	9	22 -	44 +	33	6 -	39	18	29	16	32	20	25	21
15 levensfasen	7	57 --	29 ++	14	48 --	75	14	14	13	11	13	71	17
16 kennis over vaardigheden	15	67	27	7	58 +	61	13	30	12	9	10	49	17
17 prev. gezondheidszorg	6	50 -	17	33 +	44	63	18	16	16	20	18	54	23
<b>Totaal</b>	<b>192</b>	<b>53</b>	<b>29</b>	<b>18</b>	<b>42</b>	<b>53</b>	<b>7</b>	<b>29</b>	<b>8</b>	<b>17</b>	<b>11</b>	<b>42</b>	<b>8</b>

-- / - / + / ++ / laag resp. hoog in vergelijking met totale groep

| De norm voor deze toets is: 33.09 | Uw procentuele goed-fout-score is: 42.23 |

Dit resultaat is: voldoende | Niet-opgenomen vragen: 11, 15, 18, 41, 46, 94, 103, 109 | Sleutelwijzigingen: 136 (a), 171 (c) |

jaargroep (referentiegroep). Dat kan door het resultaat van bijvoorbeeld de recentste toets te inspecteren (momentane informatie), maar het kan ook door het verloop van het kennisniveau te bestuderen (longitudinale informatie), waarbij dat laatste uit oogpunt van betrouwbaarheid natuurlijk de voorkeur verdient. Naast deze scoreresultaten op toets en subtoetsniveau krijgen de studenten na afloop van de toets ook een overzicht met literatuurverwijzingen per item. De verwijzingen betreffen altijd bronnen in de aanbevolen literatuurlijst en daarbij wordt aangegeven op welke bladzijde informatie over het correcte antwoord op de betreffende vraag gevonden kan worden.

#### *Vaststelling cesuur*

De interuniversitaire voortgangstoets hanteert een relatieve norm die bekend staat als de Wijnen-norm (Wijnen, 1971, zie ook ► hoofdstuk 7). Uitgangspunt daarbij is dat het niveau en de studie-inspanningen van de gemiddelde student voldoende moeten zijn wil de student kunnen slagen. Het referentiepunt voor de zak/slaagbeslissingen is daarom het groepsgemiddelde. Daarnaast wordt rekening gehouden met de meetfout. De methode hanteert daarom een onzekerheidsmarge overeenkomstig een 95% betrouwbaarheidsinterval en legt de zak/slaaggrens niet bij het groepsgemiddelde maar bij tweemaal de standaardmeetfout lager dan het groepsgemiddelde. Doorgaans is de betrouwbaarheid van de voortgangstoets om en nabij 0,75, en in dat geval ligt de norm bij het groepsgemiddelde minus eenmaal de standaarddeviatie van de scores in de groep. Uiteraard wordt deze norm apart berekend voor elk van de zes jaargroepen die aan een voortgangstoets deelnemen. Zoals elke norm heeft ook een relatieve norm bepaalde nadelen: de norm is niet vooraf bekend, er zakt altijd een deel van de deelnemers, de norm is beïnvloedbaar door afspraken tussen studenten en de norm is niet geschikt voor heterogene groepen. Die nadelen worden bij de interuniversitaire voortgangstoets deels ingedamd doordat de norm per jaargroep wordt bepaald op basis van de resultaten van geneeskundestudenten bij vier universiteiten. Voor een beschrijving van de verschillende methoden van normering zoals toegepast bij voortgangstoetsing, zie Muijtjens & Wijnen (2010) en Wrigley e.a. (2012).

### 15.3 Mogelijkheden van voortgangstoetsing

Voortgangstoetsing is een toetsingsprocedure die, naast het tegengaan van toetsgericht studeren, nog een aantal aantrekkelijke eigenschappen heeft (Van der Vleuten, Besseling, Tromp & Verbraeck, 1986). Omdat het een toets betreft die afgestemd is op de doelstellingen van de gehele opleiding (en niet op de onderwijsmodules die het curriculum vormen) biedt deze wijze van toetsing de mogelijkheid om het kennisniveau en de veranderingen daarin te meten, onafhankelijk van de mogelijk individueel verschillende leerpaden van studenten. De totaalscore geeft daarbij aan hoe dicht een student het beoogde eindniveau is genaderd.

Het verloop van de totaalscore voor opeenvolgende toetsen representeert de voortgang in kennisverwerving, ofwel de groei van kennis die relevant is voor de cognitieve doelstellingen. In feite reikt deze eigenschap van voortgangstoetsen nog verder. Indien er sprake is van groepen studenten die verschillende curricula volgen (bij verschillende instituten), kan een en dezelfde voortgangstoets als instrument worden gebruikt om de kennisverwerving te meten bij alle studenten. De enige voorwaarde is dat het curricula betreft die dezelfde doelstellingen hebben. De voortgangstoets is dus ook een instrument waarmee de effecten van verschillende curricula wat betreft kennisverwerving met elkaar kunnen worden vergeleken. Naast het aanwenden van dezelfde voortgangstoets als meetinstrument bij verschillende curricula, is het uiteraard ook goed mogelijk om bij de constructie van toetsen gebruik te maken van de inbreng van instituten met verschillende curricula. Naast een meer efficiënt gebruik van de beschikbare capaciteiten zou een dergelijke samenwerking ook leiden tot een meer universele inhoud van de toets, zie ook het hoofdstuk in dit boek over landelijke kennistoetsen (► hoofdstuk 4). Uiteraard heeft samenwerking alleen zin als er voldoende overeenstemming is over de interpretatie van de doelstellingen.

De voortgangstoetsresultaten bieden zicht op de groei van kennis op zowel geaggregeerd niveau als het niveau van de individuele student. De patronen op geaggregeerd niveau (bijvoorbeeld het gemiddelde verloop van de kennisgroei van 2005 tot en met 2012 voor de groep studenten die in 2005 is ingestroomd) bieden informatie betreffende de effecten van het onderwijsprogramma. De patronen op individueel niveau kunnen vroegtijdig laten zien of de voortgang van een student te wensen overlaat: bijvoorbeeld studenten die relatief hoog zijn begonnen, maar geen kennisgroei laten zien. Zij scoren nog steeds voldoende, maar zullen bij gelijkblijvende ontwikkeling in de problemen komen.

## 15.4 Voor- en nadelen

---

### ■ Voordelen

De belangrijkste voordelen van voortgangstoetsing zijn:

- *Toetsgericht studeren wordt tegengegaan.*  
Omdat het domein van de toets veelomvattend is, is het vrijwel ondoenlijk om de toets specifiek voor te bereiden. Dit bevordert continu studeren.
- *Iedere individuele studieactiviteit die relevant is voor de doelstellingen wordt beloond.*  
De toets is gericht op de doelstellingen in hun volle omvang; alle relevante studieactiviteiten worden in principe gelijk beloond.
- *Het instrument is gericht op het meten van functionele, beklijvende kennis.*  
De omvang van het toetsdomein brengt met zich mee dat met het snel van buiten leren van allerlei losstaande feiten geen goede toetsresultaten kunnen worden bereikt. Het is vooral de opgebouwde beklijvende, functionele kennis die bijdraagt aan een goed toetsresultaat. Immers, vragen over bijvoorbeeld vroeg aangeleerde (basis)kennis moeten ook in een later stadium van de studie nog goed kunnen worden beantwoord.

- *Dezelfde kennis wordt herhaald getest, iets wat de betrouwbaarheid bevordert.*  
In principe is iedere toets een herhaalde, onafhankelijke afname uit hetzelfde domein (of eigenlijk verzameling van domeinen). De gecumuleerde toetsresultaten hebben dus een hoge betrouwbaarheid.
- *Aparte herkansingen zijn niet nodig.*  
In feite zijn alle toetsen qua samenstelling vergelijkbaar en dus kan elke volgende toets fungeren als herkansing voor de voorafgaande toets.
- *Vroege detectie van studenten met uitzonderlijke hoge of lage scores.*  
Doordat de voortgangstoets toetst op eindniveau, kan al vroeg in de studie worden vastgesteld welke studenten in de buurt van het beoogde eindniveau presteren of welke studenten vergelijkenderwijs achterblijven. Het is dan mogelijk het onderwijsprogramma voor hen aan te passen.
- *De toetsresultaten vormen een rijke bron van informatie voor zowel studenten als docenten.*  
Toetsresultaten bieden zicht op het niveau en de ontwikkeling van door de studenten verworven kennis (de groei van kennis) in alle domeinen behorend bij de doelstellingen.
- *De toets biedt veel onderzoeksmogelijkheden.*  
Vanwege de volgende kenmerken biedt de toets(uitslag) veel mogelijkheden tot onderzoek: er is sprake van een gecombineerd transversaal en longitudinaal design; er zijn metingen op één tijdstip bij verschillende (groepen) personen en herhaalde metingen bij dezelfde persoon en de resultaten van subtoetsen kunnen worden vergeleken.
- *De curriculumonafhankelijkheid van de voortgangstoets.*  
De voortgangstoets is curriculumonafhankelijk. Dat biedt goede mogelijkheden voor een aanpak waarbij scholen of universiteiten met dezelfde doelstellingen (maar eventueel verschillende onderwijsprogramma's) de toets gezamenlijk construeren, afnemen en verwerken.

#### ■ Nadelen

Naast voordelen heeft voortgangstoetsing ook een aantal nadelen, waarvan de belangrijkste zijn:

- *De betrouwbaarheid van de toets is laag voor jongerejaars.*  
In het algemeen is maar een klein deel van de vragen te beantwoorden door de jongerejaars waardoor de effectieve lengte van de toets gering is. Daardoor zijn de scores in het lagere jaren minder betrouwbaar dan in de hogere jaren.
- *Het gebruik van formula scoring leidt tot bias.*  
De bereidheid om te antwoorden ook als je als student niet helemaal zeker bent dat je genoeg van het onderwerp weet, is van invloed op de score en leidt ertoe dat studenten die terughoudend zijn benadeeld kunnen worden (Muijtjens e.a., 1999; McHarg e.a., 2005).
- *De profielscores zijn onbetrouwbaar.*  
Omdat de toets veelomvattend is en dus veel verschillende onderwerpen betreft, zijn er per onderwerp relatief weinig vragen, wat nadelig is voor de betrouwbaarheid. Dit



nadeel kan worden ingedamd door aggregatie van de herhaalde metingen zoals dat bijvoorbeeld in een feedbacksysteem (ProF) gebeurt door gebruik te maken van de gecumuleerde scores (Muijtjens e.a., 2010). De gecumuleerde score is het gemiddelde van alle door een student behaalde scores op de reeks van voortgangstoetsen tot en met een bepaald meetmoment.

— *De toets vergt een lange testtijd.*

Om voldoende betrouwbaar te zijn, moet de toets veel vragen bevatten, waardoor er sprake is van een relatief lange afnametijd. Om een betrouwbaarheid van 0,80 te bereiken (bij zesdejaars) worden bij de interuniversitaire Voortgangstoets Geneeskunde tweehonderd vragen afgenomen. Met een testtijd van vier uur hebben studenten in het algemeen voldoende tijd om alle vragen te beantwoorden.

— *Er is een centrale organisatie vereist voor het ontwikkelen van de toets.*

Om het toetsconstructieproces aan te sturen en te begeleiden en de kwaliteit van de toets te waarborgen is een centrale toetscommissie nodig. Overigens, een centrale organisatie hoeft niet noodzakelijk te worden gezien als een nadeel; het leidt immers ook tot betere mogelijkheden voor kwaliteitsbewaking en efficiëntie (Van der Vleuten e.a., 1996). Bovendien ontlast het de vakdocent van allerlei administratieve en logistieke aangelegenheden.

— *Het is moeilijk om een voortgangstoets samen te stellen op basis van minder efficiënte vraagtypen.*

Bij een veelomvattende toets als de voortgangstoets is men aangewezen op het gebruik van juist/onjuist- of meerkeuzevragen, omdat minder efficiënte vraagtypen zouden leiden tot onacceptabele afnametijden en/of verwerkingsinspanningen.

## 15.5 Voorwaarden

---

Een voortgangstoetsprocedure kan in principe voor uiteenlopende vakgebieden worden ontwikkeld en toegepast, maar dan moet wel aan een aantal voorwaarden zijn voldaan.

De belangrijkste voorwaarden zijn:

- De (kern)inhoud behorend bij de doelstellingen moet goed gedefinieerd zijn. Er dient bij de staf een voldoende mate van consensus te bestaan ten aanzien van wat al of niet behoort tot de functionele kennis die relevant is voor de doelstellingen.
- De opleiding moet homogeen zijn: sterke specialisaties (zeker als ze vroeg in de opleiding ingezet worden) verhinderen succesvolle implementatie van een voortgangstoets.
- Men moet bereid zijn om te investeren in het opzetten van een centrale organisatie die een coördinerende en kwaliteitsbevorderende rol speelt bij voorbereiding en uitvoering van de voortgangstoetsprocedure.
- Er moet sprake zijn van een voldoende mate van acceptatie van de toetsingsprocedure bij zowel staf als studenten.

## 15.6 Ten slotte

---

De centrale beoordeling van de vragen en de samenstelling van de voortgangstoets vereisen een duidelijk zichtbare extra investering. Die prijs wordt onder meer betaald voor de bewaking van de kwaliteit van de vragen en het vakoverstijgende en daardoor niet-storende karakter van de toets. Daarnaast levert die aanpak als belangrijk voordeel op dat vakdocenten hun inspanningen kunnen richten op de inhoudelijke kant van de toets. Voorts dient men zich te realiseren dat bij vakgebonden toetsing de kosten vaak minder goed zichtbaar zijn vanwege het gedistribueerde karakter van de toetsactiviteiten. In hoeverre een voortgangstoets kan functioneren in een bachelor-mastermodel hangt vooral af van de mate waarin wordt voldaan aan de hiervoor beschreven voorwaarden. Bachelorprogramma's zullen veelal voldoende homogeen zijn om succesvolle toepassing van een voortgangstoets mogelijk te maken. Voor masterprogramma's, met meer uiteenlopende specialisaties, is voortgangstoetsing wellicht minder geschikt.

Een voortgangstoets toetst kennis. Naast kennis zijn er ook andere aspecten (vaardigheden, probleemoplossingsvermogen, professioneel gedrag enzovoort) die een rol spelen bij het ontwikkelen van competentie voor een bepaald beroep. Om ook de voortgang van die aspecten te toetsen, zijn andere instrumenten nodig. Het toetsstelsel voor een opleiding zal dus dienen te bestaan uit een verzameling van instrumenten en procedures die alle aspecten beslaat. Daarbij biedt de voortgangstoetsingsprocedure uitstekende mogelijkheden voor het tot stand brengen van een niet met het onderwijs interfererende vorm van kennistoetsing.

# Toetsen met performance assessment methodieken

*Gerard Straetmans*

## Inleiding

Performance assessment: waar gaat dat over?

Waarom geniet performance assessment ineens de belangstelling van opleiders?

Wat is het belangrijkste kwaliteitsaspect van een performance assessment?

Waarom staan betrouwbaarheid en validiteit op gespannen voet bij performance assessments?

Wat is een assessmentdomein en waarin verschilt het van een criteriumdomein?

Hoeveel prestaties zijn er nodig om beheersing van een leerdoel vast te stellen?

Wat zijn de vier belangrijkste vormen?

Waarom is het verstandig om assessoren te trainen?

- 16.1 Inleiding – 182
- 16.2 Argumenten voor de inzet – 183
- 16.3 Knelpunten bij het gebruik – 184
- 16.4 Enkele cruciale ontwerp vragen – 187
- 16.5 Ten slotte – 192

Performance assessment is een verzamelterm voor een groot aantal methoden waarmee gedrag kan worden uitgelokt bij studenten met als doel vast te stellen in hoeverre hogere orde doelstellingen zijn verworven. De kwaliteit van een performance assessment wordt op de eerste plaats bepaald door de representativiteit van de assessmenttaken ten opzichte van zowel het assessment- als het criteriumdomein. Het vaststellen van de beheersing van complexe doelen, bijvoorbeeld competenties, vraagt om een assessmentprogramma dat op systematische wijze dient te worden ontworpen.

## 16.1 Inleiding

Assessment is een woord dat in toenemende mate wordt gebruikt in onderwijskringen. Via de selectiepsychologie, waar assessment en assessmentcenter vanaf de jaren tachtig in de vorige eeuw centrale begrippen zijn geworden, is het woord ook definitief doorgedrongen in het onderwijs. Wie weleens Engelstalige literatuur leest over de toepassing van assessment in het onderwijs, weet dat die term daar vaak is gekoppeld aan een andere term, zoals bij: *alternative assessment*, *authentic assessment* en *performance assessment* (PA).

Alternative assessment is een term die aangeeft dat het gaat om een reactie op een heersende praktijk inzake het beoordelen van prestaties van onderwijsdeelnemers. In de Verenigde Staten was het onderwijs tot de jaren negentig van de vorige eeuw in de greep van de gestandaardiseerde meerkeuzetoetsen die in het kader van ‘accountability’ verplicht waren gesteld door de staat. Op het moment dat werd vermoed dat het gebruik van deze toetsen leidde tot een verschraling van het onderwijs, als gevolg van een verschijnsel aangeduid als ‘teaching-to-the test’, werd naarstig omgezien naar alternatieve vormen van assessment. En daaronder verstond men alle beoordelingsmethoden die verschillen van de conventionele pen-en-papier toetsen (McMillan, 2001).

De alternatieve toetsmethoden waar men naar op zoek was werden nu eens aangeduid met de term *authentic assessment*, dan weer met de term PA. Palm (2008) onderzocht het gebruik van de terminologie en concludeerde dat er geen overeenstemming te vinden was over de definiëring van beide begrippen:

» ... the concepts of performance assessment and authentic assessment have been given so many different meanings that the terms themselves practically no longer possess any meaning at all, although they are frequently used in the literature as if they had a well-defined meaning. «

Ondanks Palm's ontmoedigende conclusie wordt in dit hoofdstuk toch een poging ondernomen om een definitie te geven van het begrip performance assessment die in de context van het hoger onderwijs bruikbaar is. De definitie die een aantal Amerikaanse belangenorganisaties op het gebied van onderwijskundig meten opgesteld hebben, is aantrekkelijk vanwege zijn eenvoud:

» Performance assessments require students to complete a process or produce a product in a context that closely resembles real-life situations.' (AERA & NCME, 1999). «

■ **Tabel 16.1** Kenmerken van PA en traditionele kennistoetsen vergeleken.

Traditionele kennistoets	Performance assessment
exclusief gericht op product	gericht op product én proces
reproductie van kennis	toepassen van kennis
mentale taken	handelingstaken
abstracte taken	taken in context
veel korte opgaven per toets	één of enkele taken
één correct antwoord per opgave	meerdere oplossingen mogelijk
geheime opgaven	taken zijn niet per se geheim
groepsafname	individuele afname
weinig en vertraagde feedback	veel en onmiddellijke feedback
machinale scoring (bij gesloten toets)	waardering van prestaties door assessor
ver doorgevoerde standaardisatie	weinig standaardisatie

In gewijzigde vorm overgenomen uit McMillan (2001).

Zwak punt hierin is de nadruk op de dichte benadering van levensechte situaties. Dit zou er bijvoorbeeld toe leiden dat de kunstmatige beademing van een dummy niet benoemd zou mogen worden als een PA, aangezien dat een te grote abstractie van de werkelijkheid is. Toch is het gebruik van dummy's om levensreddende handelingen te oefenen en te toetsen onmisbaar voor onderwijs op dit gebied. Een meer acceptabele definitie luidt als volgt:

➤ **Van een performance assessment is sprake als de kandidaat onder zo natuurgetrouw mogelijke omstandigheden een zo realistisch mogelijke taak (al dan niet expliciet opgedragen) uitvoert.**

Deze definitie verduidelijkt voor een deel het begrip PA, maar is niet bepaald informatief over wat het precies behelst. Wat zijn de kenmerken van een toets die te kenschetsen is als een PA? Om hier zicht op te krijgen staat in ■ tabel 16.1 het PA gekarakteriseerd door dat te vergelijken met de beter bekende traditionele kennistoets.

## 16.2 Argumenten voor de inzet

In de literatuur worden diverse motieven genoemd voor de inzet van PA's. Nitko & Brookhart (2007) zijn bijvoorbeeld van mening dat PA's:

- de betekenis van complexe leerdoelen verhelderen;
- passen bij de uitgangspunten van de moderne (constructivistische) leertheorie;
- mogelijkheden bieden voor het beoordelen van de processen en strategieën die kandidaten volgen om taken uit te voeren.

Maar het belangrijkste argument is toch wel dat met een PA opleidingsdoelen kunnen worden getoetst die meer eisen van studenten dan het zich herinneren of begrijpen van leerstof. Het kan daarbij gaan om vaardigheden of competenties. Tegenwoordig zijn dit belangrijke doelen in alle onderwijssectoren, maar voor studenten die worden opgeleid tot professional zijn ze van primair belang. Des te vreemder is het dat nog al eens wordt verondersteld dat studenten competente professionals (dat wil zeggen dat ze de kerntaken van een beroep of functie adequaat kunnen vervullen) zijn als ze (a) alle kennistoetsen hebben gehaald en (b) de stage en (c) de afstudeeropdracht met een voldoende hebben afgesloten. Maar de relatie tussen genoemde prestaties en de beheersing van de kerntaken is vaak niet overtuigend aan te tonen:

- Traditionele schriftelijke toetsen zijn alleen geschikt om vast te stellen of studenten de voorwaardelijke kennis beheersen, niet of ze in staat zijn die toe te passen in authentieke taaksituaties.
- Stages worden doorgaans te weinig geregisseerd door de opleiding (Lampe & Hendriks, 2010; Straetmans, Roelofs & Peters, 2011). Het gevaar daarvan is dat de stagiair 'klussen' uitvoert in plaats van een goed doortimmerd stageplan dat kan informeren over de beheersing van de kerntaken door de student.
- Afstudeerscripties, ten slotte, vormen de schriftelijke neerslag van uitgevoerd (literatuur)onderzoek en zijn in de eerste plaats bedoeld om het hbo- of wo-niveau van de kandidaat aan te tonen.

Een PA daarentegen confronteert studenten met taken die sterk lijken op de taken die ze als beginnend professional moeten kunnen uitvoeren. Studenten ervaren PA's als stimulerend omdat ze het nut van de verworven kennis duidelijk maken. Een ander voordeel van PA's is dat ze studenten ertoe aanzetten om 'anders' te gaan studeren. De taken die in een PA worden aangeboden, vergen van de student een diepgaand inzicht in de relevante kennis en bekwaamheid om die toe te passen volgens onderwezen (of zelf ontwikkelde aanpassingen van) procedures. Omdat studenten al gauw beseffen dat oppervlakkige bestudering van de leerstof niet volstaat, zullen ze hun studiegewoontes aanpassen om de leerstof op een meer diepgaande wijze te verwerken. Dit zogeheten 'backwash effect' van toetsing is in de literatuur regelmatig beschreven (zie onder andere Crooks, 1988; Watkins, Dahlin & Ekholm, 2005).

### 16.3 Knelpunten bij het gebruik

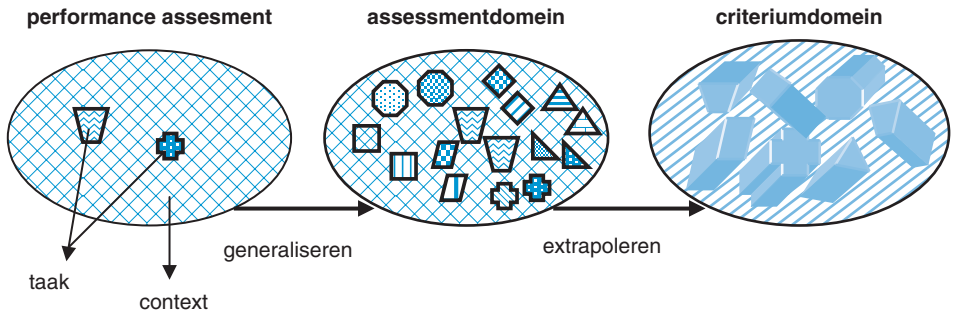
De argumenten voor de inzet van PA's klinken overtuigend, maar daar staan enkele problemen tegenover.

#### ■ Het dilemma van de representativiteit

Representativiteit is volgens Fitzpatrick & Morrison (1971) het belangrijkste kwaliteitsaspect van een PA. Dit kwaliteitsaspect heeft twee dimensies: volledigheid (*comprehensiveness*) en natuurgetrouwheid (*fidelity*). Het volgende voorbeeld helpt om te begrijpen waarom beide nodig zijn om valide uitspraken te doen over datgene wat men met een PA beoogt te beoordelen.

Het prepareren (boren) en restaureren (vullen) van tanden en kiezen is een van de belangrijkste competenties die tandheelkunde studenten moeten verwerven. Ze doen dat door in de prekliniek te oefenen met kunststof tanden en kiezen, gemonteerd in een kunststof kaak die op zijn beurt weer in een zogeheten fantoomkop is geplaatst. Stel dat de competentie op het gebied van prepareren en restaureren zou worden beoordeeld aan de hand van één zo'n in de prekliniek vervaardigd werkstuk. Daar zou men ernstige bedenkingen tegen kunnen hebben. In de eerste plaats omdat één werkstuk wel erg mager is. Niet alleen omdat je een keer pech of juist geluk kunt hebben, maar vooral omdat één werkstuk een slechte afspiegeling is van het assessmentdomein. Het assessmentdomein omvat alle verschillende taken die in het assessment hadden kunnen worden aangeboden. Het maakt nogal wat uit of je een gat moet boren in een tand, kleine kies of grote kies. En ook of dat gat in het kauwvlak zit, tussen de tanden, aan de tong- of aan de wangkant. Ten slotte is de positie in de kaak van invloed want als het element in de bovenkaak zit, moet er worden gewerkt vanuit de spiegel en dat is een complicerende factor. Kortom, de assessmentprestatie op één werkstuk kan nooit volledig representatief zijn voor de assessmentprestatie op alle werkstukken die in principe aangeboden hadden kunnen worden in een assessment. Dit betekent dat het prestatieniveau afhankelijk kan zijn van de taak of taken die de kandidaat in het assessment moet uitvoeren. Linn & Burton (1994) geven aan dat het probleem van de taakspecifieke prestaties in alle domeinen speelt. Taakspecifieke prestaties zijn een gevolg van het feit dat kennis en vaardigheden niet zonder meer overdraagbaar zijn van de ene toepassingscontext naar de andere (Eraut & du Boulay, 2000).

Maar het is nog gecompliceerder, want eigenlijk gaat het niet over prestaties die studenten leveren onder assessmentcondities. Het gaat uiteindelijk om de prestatie in de beoogde vervolgsituatie, bijvoorbeeld in de beroepsuitoefening, ook wel aangeduid als het criteriumdomein. Dat is in dit specifieke geval de verzameling van verschillende preparerende en restaurerende handelingen, uit te voeren bij echte patiënten, met elk hun eigen tandheelkundige problematiek. In dat licht bezien is het een ernstig manco van het hierboven besproken PA dat de competentie wordt beoordeeld op basis van een taak die tamelijk ver af staat van de werkelijkheid. Een kleine kies of een grote kies van kunststof is altijd hetzelfde, maar de kleine kies bij patiënt A kan heel anders van vorm zijn dan bij patiënt B. Bovendien voelt boren in kunststof totaal anders aan dan boren in echt tandmateriaal. En ten slotte is het een wereld van verschil of je boort in een stuk plastic of in een echte tand in de mond van een patiënt van vlees en bloed, die waarschijnlijk niet voor het plezier in de stoel zit. ■ **Figuur 16.1** geeft het probleem van het trekken van conclusies over competentie op basis van assessmentprestaties schematisch weer. Te zien is dat de uitgevoerde assessmenttaken geen goede afspiegeling zijn van het assessmentdomein (tekortschietende 'volledigheid'), waardoor generalisatie van de assessmentprestatie over alle taken in het assessmentdomein problematisch wordt. Bovendien blijken de taken in het assessmentdomein naar inhoud en context aanzienlijk te verschillen van de taken in het criteriumdomein (tekortschietende 'natuurgetrouwheid'). Daarmee wordt het erg riskant om de prestatie op het beschreven PA te extrapoleren naar prestaties in het criteriumdomein en een conclusie te trekken over het al dan niet beheersen van de competentie 'prepareren en restaureren van tanden en kiezen'.



■ **Figuur 16.1** Tekortschietende representativiteit van een PA (Straetmans, 2006).

*‘Basically, you can’t win’* concludeerde Kane (1992) nadat hij had geconstateerd dat het, uitgaande van een beperkte hoeveelheid tijd en geld die aan een PA mag worden besteed, doorgaans niet mogelijk is om zowel de volledigheid als de natuurgetrouwheid van een PA te bevorderen.

#### ■ Beoordelen is zelden objectief

Een PA confronteert een kandidaat met een taak. In hoeverre de te meten eigenschap (vaardigheid of competentie) is verworven, moet blijken uit de kwaliteit van de taakuitvoering (proces) en het uiteindelijke resultaat daarvan (product). Meestal moet een ter zake deskundige worden ingezet om de kwaliteit van de geleverde prestatie in te schatten. Helaas is de afgelopen decennia vaak aangetoond dat de vakinhoudelijke deskundigheid van assessoren geen garantie biedt voor een deugdelijke beoordeling. Onafhankelijk van elkaar gegeven oordelen door meerdere assessoren komen maar zelden volledig overeen (Hofstee, 1999). In een poging de objectiviteit van de oordelen te vergroten wordt dikwijls voorgeschreven op welke aspecten van het te beoordelen proces en/of product het oordeel zich moet richten. Soms schiet men daarin door en probeert men prestatiecriteria te formuleren die zo min mogelijk hoeven te worden geïnterpreteerd door een assessor. Jammer genoeg beperkt men zich daarmee tot eigenschappen en kenmerken die gemakkelijk waarneembaar zijn en dat zijn lang niet altijd de aspecten die het verschil maken tussen competent en incompetent handelen. Neem bijvoorbeeld de zangprestatie van een conservatoriumstudent. Met de moderne ICT-middelen is het vrij eenvoudig om een lied op te nemen en met gespecialiseerde software te analyseren op kenmerken als toon, pitch, timbre, volume. Met enige inspanning lukt het ook om voor een specifiek lied per passage de gewenste meetwaarden ten aanzien van die kenmerken te beschrijven, waarmee de weg naar een objectieve beoordeling van een creatief product zou zijn vrijgemaakt. Gelukkig willen de meesten niet zover gaan; het staat veel te ver af van waar het bij een lied eigenlijk om moet gaan, namelijk of het mooi klinkt en of je er door wordt geraakt. Maar vanwege het subjectieve karakter wordt dit in het geval van een formele beoordeling meestal geen goed idee gevonden. Er zal daarom een tussenweg moeten worden gevonden, een die de te beoordelen competentie niet ‘kapot analyseert’, maar tegelijkertijd zo veel steun biedt aan de assessor dat de beoordeling niet teveel mening en te weinig feitelijk is.

Dit voorbeeld maakt duidelijk dat volledig objectieve beoordeling van competenties geen doel is dat koste wat kost moet worden nagestreefd. De wellicht hogere



betrouwbaarheid zal ten koste gaan van de betekenis die men aan die scores kan geven; van de validiteit dus. Belangrijker is het dat de beoordeling zich toespitst op aspecten die kwaliteitsverschillen in processen en producten bewerkstelligen.

#### ■ Kunnen is iets anders dan doen

Bij een PA gaat het erom dat de te beoordelen personen hun kennis en vaardigheden inzetten om een bepaalde prestatie te leveren. Vaak zal een assessmentprestatie uitmonden in een te beoordelen product, maar ook de gevolgde handelwijze, het proces, is beoordeelbaar. Een probleem daarbij is dat de aanwezigheid van een assessor het te beoordelen gedrag kan beïnvloeden. Het is in het voordeel van de kandidaat als die zijn gedrag afstemt op wat de assessor graag wil zien. Maar daarmee wordt gedeeltelijk het zicht ontnomen op hoe de kandidaten zouden handelen in een reële werksituatie, als er niemand op hun vingers kijkt. Dat dit een zwak punt is van PAs blijkt bijvoorbeeld uit het agressieve rijgedrag van een groep weggebruikers. Ongetwijfeld hebben alle agressieve rijders tijdens hun rijexamen laten zien dat ze in staat zijn om defensief te rijden, maar kennelijk is ‘kunnen’ iets anders dan ‘doen’.

Barnes & Morgeson (2007) beschrijven dit verschijnsel met de begrippenparen ‘obtrusive versus unobtrusive observation’ en ‘maximum versus typical performance’. Bij obtrusive observation zijn de te beoordelen personen zich ervan bewust dat zij worden geobserveerd en beoordeeld. Deze kandidaten zullen proberen om een zo goed mogelijke prestatie te leveren en het is dan ook niet ondenkbaar dat ze zich bewust positiever gedragen dan in een reële werksituatie. Maar het kan ook dat er juist een slechtere prestatie wordt geleverd. Bijvoorbeeld omdat een persoon, door zenuwen geplaagd, apathisch of juist hyperactief reageert op de assessmenttaak. Bij unobtrusive observation is de te beoordelen persoon zich er niet van bewust dat er wordt beoordeeld. De persoon gedraagt zich ‘naturel’. Uit dat gedrag valt ook iets af te leiden over de (beroeps)houding en motivatie. Bij obtrusive observation wordt waargenomen waartoe iemand in staat is (maximum performance), bij unobtrusive observation wat iemand werkelijk doet in een bepaalde taaksituatie (typical performance). Een voorbeeld van unobtrusive observation is het werken met een ‘mystery guest’. Een mystery guest is een assessor die zich voordoet als een reguliere ‘klant’. Aan de hand van vooraf opgestelde richtlijnen en beoordelingscriteria evalueert de mystery guest bepaalde aspecten van het gedrag van degene die wordt beoordeeld. Op deze manier beoordeelt de NS bijvoorbeeld of de conducteurs hun werk naar behoren uitvoeren. Het hoger onderwijs maakt nauwelijks gebruik van deze methode terwijl de omvangrijke stageperiodes daar voldoende gelegenheid voor bieden. Het belangrijkste argument dat pleit voor toepassing in het onderwijs is dat deze methode beter zicht geeft op de attitude van de kandidaat. Zeker als het gaat om competentiebeoordeling kan dat een zinvolle aanvulling vormen op het bewijs dat via andere methoden is verkregen (Van der Kleij, Wools & Straetmans, 2011).

## 16.4 Enkele cruciale ontwerp vragen

Deugdelijke beslissingen nemen over kandidaten op grond van toetsscores is alleen mogelijk als die zo veel mogelijk vrij zijn van meetfouten (betrouwbaar zijn) en eenduidig te

interpreteren en te gebruiken zijn voor het beoogde doel (valide zijn). Om betrouwbare en valide scores te kunnen genereren, moet een toets zo veel mogelijk (gegeven de beschikbare tijd) items, vragen of opgaven bevatten. Beheersing van competentie moet daarom worden vastgesteld op basis van een assessmentprogramma in plaats van op basis van één of enkele uitgevoerde assessmenttaken (zie onder andere Stiggins, 1987; Baartman, e.a., 2006; Straetmans, 2006). Een systematisch ontwikkelproces biedt de beste waarborg voor de ontwikkeling van assessmentopdrachten en -programma's waarmee deugdelijke beslissingen over studenten zijn te nemen (zie onder andere Mueller, z.j.; Straetmans, in voorbereiding). Daarbij zijn aantal beslispunten cruciaal.

#### ■ Vaardigheid versus competentie

Bij een PA voert een kandidaat een taak uit onder meer of minder gecontroleerde condities. De prestatie die daarvan het gevolg is, wordt gebruikt om een uitspraak te doen over de beheersing van een bepaald leerdoel. Hoeveel prestaties er nodig zijn, hangt in de eerste plaats af van de vraag hoe cruciaal de te nemen beslissing is. Bij zwaarwegende beslissingen (selectie of diplomering) gelden hogere kwaliteitseisen en dienen daarom meer prestaties te worden geleverd dan bij voortgangsbepalingen. In de tweede plaats hangt het aantal prestaties af van het type leerdoel. Gaat het om een vaardigheid of om een competentie? Personen zijn vaardig als ze in staat zijn tot het efficiënt en doelgericht uitvoeren van geleerde procedures. Die procedures kunnen heel divers van aard zijn. Het kan gaan om manuele, perceptuele, communicatieve en redeneervaardigheden. Personen zijn competent als ze kunnen omgaan met slecht-gedefinieerde problemen in complexe situaties. Dit houdt in dat competentie geen generieke eigenschap is, maar slechts geldingskracht heeft voor een bepaalde klasse van taaksituaties. De nadruk op slecht-gedefinieerde taaksituaties betekent dat er pas sprake is van competentie als de kandidaat ook in staat is om adequaat te handelen in taaksituaties die niet op een standaardwijze kunnen worden benaderd.

Een voorbeeld uit de ambulancezorg kan helpen om dit toe te lichten. Een ambulanceverpleegkundige moet vaardig zijn in het kunstmatig beademen. Dit wordt doorgaans getoetst door een zogeheten dummy te laten beademen. Beheersing van de vaardigheid moet blijken uit het strikt en adequaat uitvoeren van het daarvoor geldende protocol. Na één of hooguit twee keer het protocol onder toetscondities te hebben uitgevoerd, laat zich wel een conclusie trekken over de beheersing van de vaardigheid, maar niet over de competentie om deze levensreddende handeling uit te voeren. Competentie op dit terrein houdt meer in dan het adequaat uitvoeren van het voorgeschreven protocol. De gevallen waarin kunstmatige beademing noodzakelijk is, zijn divers en wijken niet zelden af van de doorsnee casus waarop het protocol is gebaseerd. De ambulanceverpleegkundige moet dan ook in staat zijn om het protocol daarop af te stemmen door bijvoorbeeld stappen over te slaan, aan te passen of in een andere volgorde uit te voeren. Daarvoor is vaak kennis vereist van het specifieke ziektebeeld. Een uitspraak over de beheersing van de betreffende competentie kan daarom niet gebaseerd worden op één of twee uitgevoerde casussen.

#### ■ Keuze voor assessmentvorm

'Gedrag voorspelt gedrag.' Dat is een uitspraak die wordt gebruikt om assessmentcenters te promoten. Men wil ermee zeggen dat naarmate assessmentopdrachten sterker lijken op

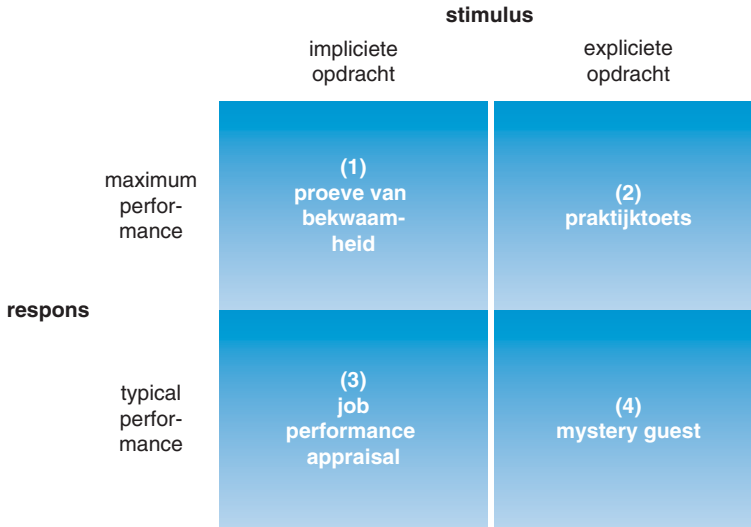
die in de criteriumsituatie (de vervolgsituatie waarvoor het assessmentprogramma de geschiktheid vaststelt) er meer valide voorspellingen over het functioneren in de criteriumsituatie kunnen worden gedaan. Dat klinkt aannemelijk. Als je wilt weten of iemand een auto kan besturen, dan moet je die persoon een stukje laten rijden. Wil je weten of iemand pannenkoeken kan bakken, proef dan de pannenkoeken die deze persoon heeft bereid. Een principekeuze voor 'natuurgetrouwe' assessmentopdrachten wil echter niet zeggen dat andere, minder authentieke assessmentopdrachten nooit zouden moeten worden ingezet. Er zijn diverse redenen waarom men er bij een assessment toe over zou kunnen gaan om de werkelijkheid na te bootsen.

- *Gevaar.* Als de kandidaat en/of diens omgeving door foute handelingen gevaar loopt of hinder ondervindt, is het zaak om terughoudend te zijn met beoordelingen die plaatsvinden in reële werksituaties. Noodlandingsprocedures worden om die reden geoefend en beoordeeld in een vluchtsimulator.
- *Beschikbaarheid van geschikte problemen.* Gevorderde EHBO'ers bijvoorbeeld moeten een 'triage' (een selectie van gewonden) kunnen uitvoeren. Maar calamiteiten met grote aantallen gewonden zijn uiteraard niet op afroep beschikbaar.
- *Kosten.* Bepaalde processen in de chemische industrie bijvoorbeeld zijn te duur om er in het kader van een beoordeling risico mee te willen lopen. Dure producten die maar eenmalig gebruikt kunnen worden, komen evenmin in aanmerking voor gebruik in een performance assessment.

In de literatuur worden tientallen assessmentvormen onderscheiden (zie Jaspers & Van Zijl, 2011; Van Berkel & Bax, 2006), maar vaak verschillen ze niet wezenlijk van elkaar. In feite zijn er niet meer dan vier grove methodieken te onderscheiden waarmee het competente handelen van personen kan worden beoordeeld. De vier methodieken ontstaan door combinatie van de stimulus- en responsecategorieën van assessmentvormen (zie [figuur 16.2](#)). De stimulus bestaat uit de opdracht(en) en de verdere relevante informatie, waaronder de instructies die een kandidaat krijgt. Een stimulus kan expliciet of impliciet zijn. Een stimulus is expliciet als de kandidaat een opdracht krijgt om een bepaalde taak uit te voeren. Een impliciete stimulus houdt in dat de kandidaat geen opdracht(en) krijgt anders dan om aan het werk te gaan.

De responsecategorieën die onderscheiden worden, verwijzen naar het eerder genoemde verschil tussen maximum en typical performance.

In cel 1 van [figuur 16.2](#) vallen assessmentvormen waarbij de kandidaat functioneert in een reële werksituatie, maar weet dat er wordt beoordeeld. De kandidaat krijgt geen expliciete opdrachten; deze voert gewoon de taken uit die zich in de werksituatie aandienen. Het is aan de kandidaat om tijdens zijn werkzaamheden stimuli te signaleren en daarop te responderen. Het voordeel daarvan is dat zo kan worden getoetst of de kandidaat het inzicht heeft om zelfstandig te beslissen of en hoe er moet worden gereageerd in bepaalde situaties. De gebruikelijk benaming voor dergelijke assessments is 'proeve van bekwaamheid'. Een voorbeeld is de LIO-stage waarin een student gedurende enkele maanden de volledige verantwoordelijkheid heeft voor een groep leerlingen. De assessments in deze categorie kenmerken zich door een hoge natuurgetrouwheid, maar ook door een zekere inefficiëntie aangezien een assessor soms erg lang moet observeren voordat alle benodigde informatie is vergaard.



■ **Figuur 16.2** Indeling van assessmentvormen op basis van gedrag van de kandidaat en aard van de stimulus.

In cel 2 is de praktijktoets ondergebracht. Dat zijn assessmentvormen waarbij kandidaten onder meer of minder gecontroleerde omstandigheden werken aan meer of minder authentieke taken. De kandidaat weet dat er tijdens de uitvoering van de assessmenttaak wordt beoordeeld. Anders dan bij de proeve van bekwaamheid wordt door middel van gerichte opdrachten het gewenste gedrag (en het daaruit voorkomende resultaat) uitgelokt. De meeste sturing valt daarbij te geven in de hands-off variant en de minste bij de hands-on variant (Straetmans, 2006). Van de hands-off variant is bijvoorbeeld sprake bij de 'situational judgment test'. Dat is een praktijktoets die bestaat uit een aantal beschreven of gefilmde kritische beroepssituaties. Van elk daarvan moet de kandidaat aangeven wat daarop de beste reactie is. Dat gebeurt door de beste keuze te maken uit een reeks alternatieve vervolgacties. Zie bijvoorbeeld ► <http://www.vandermoesen.nl/ecartoon/nl/>. Een bekend voorbeeld van de hands-on variant is het praktisch deel van het rijexamen. De examiner geeft opdrachten over de te volgen route of uit te voeren verrichtingen, maar heeft geen controle over de omstandigheden waaronder kandidaten hun taken moeten verrichten. Denk aan verkeersintensiteit, weersomstandigheden, onverwacht gedrag verkeersdeelnemers, en dergelijke.

Het voordeel van een praktijktoets ten opzichte van een proeve van bekwaamheid is de grotere efficiëntie. De expliciete opdrachten lokken gedrag uit wat het mogelijk maakt de benodigde informatie sneller te vergaren. Een nadeel is dat deze methodiek geen antwoord geeft op de vraag of de kandidaat de wil/bereidheid en het inzicht heeft om in taaksituaties op eigen initiatief tot actie over te gaan. Als de beoordeling zich richt op een vaardigheid is dat geen probleem, maar bij de beoordeling van competenties is het ook van belang om te weten of de kandidaat in een taaksituatie de signalen kan detecteren die om interventie vragen.

De assessmentvormen in cel 3 worden vaak gebruikt bij de beoordeling van het functioneren van werknemers, al dan niet in opleiding. Er is een onderscheid tussen objectieve

en subjectieve methoden. Objectieve methoden zijn die waarbij de nadruk ligt op de opbrengst van een functionaris in termen van geld of producten. Dat klinkt heel vertrouwenwekkend, maar opbrengsten kunnen maar zelden volledig worden toegeschreven aan de prestaties van een individuele functionaris. Subjectieve methoden maken gebruik van assessoren om het functioneren van werknemers te waarderen. Denk bijvoorbeeld aan instrumenten voor 360 graden feedback. Assessmentvormen in cel 3 kenmerken zich door de vrijheid die de assessor heeft om zelf te bepalen op welke observatie(s) de beoordeling wordt gebaseerd. De assessoren proberen zich het functioneren van de kandidaat in diverse situaties voor de geest te halen en beoordelen dit aan de hand van criteria. Er is dus geen sprake van directe observatie van het functioneren. Assessmentvormen in cel 3 leveren daardoor informatie over de typical performance van kandidaten. Als de assessoren met zorg geselecteerd zijn en werkelijk zicht hebben op het functioneren van de kandidaat dan kan de validiteit van de beoordelingen groot zijn. Van de andere kant kan het informele karakter van de informatieverzameling en het sterke beroep dat wordt gedaan op de herinnering van assessoren gemakkelijk leiden tot beoordelingen die sterk zijn ‘gekleurd’ door enkele opvallend goede of zwakke prestaties van de betreffende kandidaat. Dat zijn namelijk de prestaties die de assessoren zich het makkelijkst zullen herinneren.

Bij de assessments in cel 4 is weer sprake van expliciete opdrachten. Dit keer niet aan de kandidaat, maar aan een persoon – de mystery guest – die tot taak heeft het gedrag waarnaar de interesse uitgaat, uit te lokken. Hier zijn de voordelen van de methoden uit cel 1 (dichte benadering van de reële werksituatie) en cel 2 (hoge efficiëntie door gerichte uitlokking van gedrag) gecombineerd. Nadelen van deze methode schuilen in de hoge voorbereidings- en uitvoeringskosten en in ethische vragen. Is het wel acceptabel om personen te observeren en te beoordelen zonder dat ze daarvan op de hoogte zijn?

De assessmentvorm in cel 1 is in principe te verkiezen boven de altijd wat kunstmatige assessmentvormen uit cel 2. De assessmentvormen uit cel 3 en 4 zijn in onderwijssituaties vooral van belang om zicht te krijgen op de attitude als component van de te beoordelen competentie en op wat mensen daadwerkelijk doen (dus niet alleen ‘kunnen’) in taaksituaties. Elk van de vier methoden heeft sterke en zwakke eigenschappen. Het is dan ook raadzaam om een bewijsdossier op te bouwen waarin de bewijzen tot stand zijn gekomen op basis van verschillende methoden, de zogenoemde ‘methodemix’ (Straetmans, 2006).

#### ■ Waarborgen van de beoordelingskwaliteit

Bij een PA voeren kandidaten taken uit. Zowel de taakuitvoering zelf als de resultaten daarvan dienen te worden beoordeeld om zicht te krijgen op de mate van beheersing.<sup>1</sup> In de literatuur is veel bewijs voorhanden waaruit blijkt dat mensen slechte beoordelaars zijn (zie Hofstee, 1999). De inzet van rubrics kan bijdragen aan een betrouwbaardere beoordeling (zie ► hoofdstuk 17 voor uitvoerige informatie over toepassing van rubrics voor beoordelingsdoeleinden).

Het gebruik van rubrics leidt niet vanzelf tot betere beoordelingen. Training van assessoren is aan te bevelen. De feedback die een trainee krijgt, draagt eraan bij dat de niet te vermijden interpretaties van de criteria zo veel mogelijk in dezelfde richting gaan als die

1 Soms zijn proces en product niet afzonderlijk te beoordelen. Met name is dit het geval bij prestaties die leiden tot een ‘vluchtig’ product zoals zang, dans, toneel, spreekvaardigheid en dergelijke.

van de collega's. Training helpt ook om assessoren zich bewust te laten zijn van hun feilbaarheid waardoor ze eerder zullen zijn geneigd om maatregelen te treffen die nodig zijn om 'personal-bias errors' te voorkomen. Dat zijn aan de persoon gebonden zwakheden die de beoordelingen van de assessor systematisch doen afwijken van het gemiddelde oordeel van andere beoordelaars (zie de bijlage van dit boek).

Training van assessoren leidt vaak tot betere beoordelingen, in die zin dat de interbeoordelaarsbetrouwbaarheid van getrainde assessoren vaak groter is dan die van ongetrainde assessoren. De onderzoeksresultaten op dit gebied zijn echter wisselend, met name wordt getwijfeld aan het beklijven van het trainingseffect (Lumley & McNamara, 1995). Anders gezegd: houdt het positieve effect stand bij andere (ten opzichte van de training) te beoordelen eigenschappen, andere rubrics en het verstrijken van de tijd? Vooralsnog lijkt het verstandig om met meerdere assessoren te werken als het gaat om het beoordelen van prestaties op basis waarvan belangrijke, onomkeerbare beslissingen over de kandidaat worden genomen.

## 16.5 Ten slotte

---

Performance assessment is de aangewezen weg om na te gaan in hoeverre studenten in staat zijn om in bepaalde situaties vaardig en/of competent te kunnen handelen. De 'directheid' van deze methode is een aantrekkelijke eigenschap. Toch schuilen hier ook gevaren. Het uitvoeren van een authentieke assessmenttaak lijkt als vanzelf te leiden tot valide scores, maar soms is dit niet meer dan schijnvaliditeit (*face validity*). Als bijvoorbeeld een kandidaat tijdens het praktisch deel van het rijexamen door de drukte zo vaak stilstaat dat er maar weinig tijd resteert om de rijvaardigheid te kunnen beoordelen, dan is de zak/slaagbeslissing in alle redelijkheid niet als valide te beschouwen. Representativiteit behelst meer dan een hoge natuurgetrouwheid. Om die reden vraagt valide beoordeling van competentie om een goed doordacht en systematisch ontworpen assessmentprogramma.

# Toetsen met rubrics

*Paul van den Bos, Carel Burghout en Desirée Joosten-ten Brinke*

## Inleiding

Wat is een rubric?

Waarin onderscheidt een rubric zich van beoordelingscriteria?

Welke verschillende typen zijn er?

Wanneer gebruik je welk type?

Hoe gebruik je een rubric formatief?

Hoe gebruik je een rubric summatief?

Hoe construeer je een rubric?

## 17.1 Definities en soorten – 194

17.1.1 Definitie – 194

17.1.2 Holistische en analytische rubrics – 196

17.1.3 Taakspecifieke en generieke rubrics – 197

17.1.4 Beoordelingsniveaus – 198

17.1.5 De indicatoren – 199

## 17.2 Rubrics voor formatief én summatief toetsen – 199

17.2.1 Rubrics en de summatieve toetsfunctie – 199

17.2.2 Rubrics en de formatieve toetsfunctie – 200

## 17.3 Curriculumbrede inzet – 201

## 17.4 Het beoordelen van de kwaliteit – 202

## 17.5 Een stappenplan voor de constructie – 202

Om de meer complexe vaardigheden zichtbaar te maken, bijvoorbeeld op het terrein van kritisch denken, samenwerken, presenteren, kunnen reflecteren op het eigen handelen, een gesprek kunnen voeren, of een professionele houding vertonen, is het niet mogelijk om gebruik te maken van gestandaardiseerde schriftelijke toetsen met een antwoordmodel. Veelal wordt voor deze vaardigheden gebruikgemaakt van alternatieve toetsvormen, zoals performance assessments of authentieke assessments. Een van de lastigste aspecten van dit type toetsen is te komen tot een betrouwbare manier van beoordelen en – wanneer deze toetsvormen summatief worden gebruikt – te komen tot een normering en becijfering (zie ook de hoofdstukken 16 en 19).

Rubrics kunnen hierbij goed van pas komen. Een rubric, ook wel beoordelingsrubriek genoemd, is een instrument om op een valide en betrouwbare manier aan de hand van een beschrijving van criteria en verwachtingen informatie geven over de prestatie van een student. Het belangrijkste voordeel van rubrics is dat ze inhoudelijke informatie geven over de geleverde leerprestatie. Hierdoor kunnen ze worden ingezet voor formatieve en summatieve toetsing, peer- en zelfbeoordeling. In de Angelsaksische landen worden rubrics als beoordelingsinstrument al veel gebruikt; in Nederland is de bekendheid met rubrics in het hoger onderwijs de laatste jaren groeiend.


De rubric maakt voor studenten én docenten de relatie inzichtelijk tussen cursusdoelen en de verschillende beheersingsniveaus op weg naar deze cursusdoelen. Het is een bruikbaar instrument om te bepalen hoe ver studenten zijn gevorderd in het beheersen van de doelen. De rubric geeft daardoor inzicht in de kloof tussen gewenst beheersingsniveau en het huidige niveau en geeft de student gelegenheid gerichte feedback te vragen of om zelf te reflecteren over hoe die kloof te overbruggen. Daardoor zal het gebruik van rubrics bijdragen aan het vergroten van de zelfstandigheid van studenten en het verkrijgen van een beter inzicht in het eigen leerproces.

## 17.1 Definities en soorten

---

### 17.1.1 Definitie

---

Een rubric is een beoordelingsmatrix (ofwel scoringslijst, antwoordmodel, lijst met criteria), bestaande uit twee dimensies (zie  figuur 17.1). De eerste dimensie bevat de toetscriteria en de tweede dimensie de beoordelingsniveaus. In de rubric worden de beoordelingsniveaus zo goed mogelijk inhoudelijk beschreven in de vorm van indicatoren. Deze indicatoren zijn beschreven op verschillende niveaus.

Bij het indelen van de niveaus kan worden gekozen voor een indeling van een lage naar een hoge score van links naar rechts, of andersom, van de hoogste naar de laagste score van links naar rechts. Voordeel van dit laatste is dat de best mogelijke prestaties meteen zichtbaar zijn, direct na de eerste kolom. Deze weergave gaat uit van een positieve benadering. In het gezaghebbende handboek van the American Council on Education, *Educational Measurement*, worden de hoogst te behalen scores in een rubric ook als eerste genoemd (Brennan, 2006). Een korte verkenning van op het internet aanwezige rubrics laat echter zien dat beide worden gehanteerd. In een cultuur waarin het de gewoonte is om van de



		niveaus			
		A	B	C	D
criteria	1				
	2				
	3				
	---				
	n				

■ **Figuur 17.1** Het format van een lege rubric met verticaal in de eerste kolom de criteria en horizontaal in de tweede kolom de verschillende niveaus.

■ **Tabel 17.1** Voorbeeld van een analytische rubric (onderdeel van een rubric voor het beoordelen van een presentatie; afkomstig uit een hbo-opleiding Economie).

	Onvoldoende	Voldoende	Goed	Uitstekend
<b>Afstemming van inhoud en doel van de presentatie</b>	de inhoud van de presentatie komt niet overeen met het doel en de informatie is niet volledig	de inhoud van de presentatie komt overeen met het doel en bevat de benodigde informatie	de inhoud van de presentatie komt overeen met het doel en bevat de benodigde informatie	inhoud en doel van de presentatie vormen een samenhangend geheel
<b>Gebruik van bronnen</b>	er zijn geen bronnen geraadpleegd	er zijn enkele relevante bronnen geraadpleegd	er zijn meerdere relevante bronnen geraadpleegd	de gebruikte bronnen ondersteunen duidelijk de gegeven informatie
<b>Gebruik van voorbeelden</b>	er worden geen voorbeelden gegeven	er wordt een voorbeeld gegeven	er wordt een adequaat voorbeeld gegeven	er wordt gebruikgemaakt van meerdere adequate voorbeelden

laagste score naar de hoogste score te werken (denk aan Likertschalen bij vragenlijsten) ligt een indeling andersom, met eerst links een kolom met lage score, meer voor de hand.

Rubrics zijn geschikt bij alle open toetsvormen. Arter & Chappuis (2006) bespreken vier *typen doelstellingen* waarbij rubrics als beoordelingsinstrument kunnen dienen:

- kennisdoelen op het niveau van inzicht en toepassing;
- hogere orde cognitieve vaardigheden zoals analyseren, vergelijken, synthese, probleem oplossen, kritisch denken;
- ‘performance skills’ zoals een presentatie geven, een taal spreken, laboratoriuminstrumenten correct gebruiken, groepswork, een instrument bespelen (zie ■ tabel 17.1);

■ **Tabel 17.2** Onderdeel van een rubric voor het beoordelen van een ontwerpanalyse (afkomstig uit een opleiding Architectuur).

	Slecht	Onvoldoende	Voldoende	Goed	Uitstekend
<b>Analyse en uitgangspunten</b>	minder dan 4 onderdelen geanalyseerd	alle 4 onderdelen geanalyseerd en bevatten schetsen; conclusies niet aanwezig of niet duidelijk, zodat ook uitgangspunten voor ontwerp nog niet kloppen	alle 4 onderdelen geanalyseerd en bevatten duidelijke, schematische schetsen; hieruit zijn duidelijke conclusies getrokken; eerste aanzet voor ontwerpuitgangspunten beschreven	alle 4 onderdelen geanalyseerd en bevatten duidelijke, schematische schetsen; hieruit zijn duidelijke conclusies getrokken die leiden tot ontwerpuitgangspunten	alle 4 onderdelen geanalyseerd en bevatten uitgesproken en persoonlijke schetsen en conclusies; ontwerpuitgangspunten vloeien hier op logische manier uit voort

- ‘creating products’ zoals onderzoeksrapporten, grafieken maken, kunstobjecten maken (zie ■ tabel 17.2).

Er is een onderscheid te maken tussen rubrics: holistische versus analytische rubrics, en taakspecifieke versus generieke rubrics.

### 17.1.2 Holistische en analytische rubrics

Het eenvoudigste criterium om een oordeel te geven over een product is: ‘werkt het?’ Als een student de opdracht heeft gekregen voor het ontwikkelen van een werkend computerprogramma, dan kan de beoordeling bestaan uit het draaien van het computerprogramma. Levert het programma het gewenste resultaat op dan is er een positief oordeel, zo niet dan volgt een negatief oordeel. Dit is een heel holistische manier van beoordelen. Bij *holistische* rubrics worden complexe vaardigheden zo veel mogelijk als één geheel beoordeeld. Dit type rubrics geeft een algemene impressie over de uitvoering van een taak. Of het programma werkt is waarschijnlijk maar een van de aspecten waaraan een goed product moet voldoen. Als er in een keer een aantal criteria samen worden genomen om een oordeel te geven over het product, dan wordt gesproken van een holistische beoordeling (zie ■ figuur 17.2 waarin is weergegeven hoe analytische rubrics verschillen van holistische rubrics). Bijvoorbeeld bij een presentatie geeft de beoordelaar een overall indruk van de kwaliteit met behulp van drie niveaus: zwak, voldoende of goed, die elk uitgebreid worden omschreven.

In een *analytische* rubric (zie ■ tabel 17.1) wordt juist gestreefd naar zo eenduidig en enkelvoudig mogelijke criteria zodat alle relevante beoordelingsaspecten afzonderlijk van een kwalificatie worden voorzien.

**a**

	1	2	3	4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				

**b**

	1	2	3	4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				

**c**

	1	2	3	4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				

**d**

	bv. kwaliteit van het portfolio
1. uitstekend	
2. voldoende	
3. onvoldoende	

■ **Figuur 17.2** Van een analytische rubric naar een holistische rubric.

Beide vormen van rubrics hebben voor- en nadelen. Holistische rubrics hebben het voordeel dat ze snel tot een scoring leiden en dat de betekenis van het geheel behouden blijft. Bij een portfolio, een creatief product, een onderzoeksverslag en dergelijke kan immers bij te veel versnippering het totaalbeeld op de achtergrond raken. Een te uitgebreide 'afvinklijst' kan ertoe leiden dat er teveel naar de vinkjes toe wordt gewerkt. Met andere woorden: de rubric weerspiegelt in dat geval niet meer het onderliggende onderwijsdoel maar is een doel op zich geworden. Bij holistische rubrics geldt juist: het geheel is méér dan de som der delen. Holistische rubrics laten veel ruimte over voor interpretatie. Training van docenten op het gebruik van deze rubrics en dan vooral op de overeenstemming in interpretatie is noodzakelijk.

Het voordeel van de meer analytische rubric is dat de kans dat meer beoordelaars tot hetzelfde eindoordeel komen groter is, en daarmee ook de betrouwbaarheid van de beoordeling. Bij summatieve toetsing is dat belangrijk. Een analytische rubric geeft de student meer inzicht in de veelzijdigheid en complexiteit van de taak dan een holistische rubric. Een nadeel van analytische rubrics is echter dat ze meer tijdrovend zijn dan holistische rubrics. Het is bij de keuze voor een holistische of een analytische benadering dus zaak een goede balans te vinden tussen behoud van betekenis (validiteit) enerzijds en de betrouwbaarheid van de beoordeling anderzijds.

Een compromis kan worden gevonden in het wat meer clusteren van de diverse te beoordelen deelaspecten die inhoudelijk met elkaar samenhangen (zie ■ tabel 17.2). Hierdoor kunnen wel verschillende typen criteria worden onderscheiden, maar de aspecten binnen het criterium blijven inhoudelijk een eenheid. Het maakt het mogelijk om in de beoordeling een gradatie van diepgang aan te brengen.

In ■ figuur 17.2 is weergegeven hoe van een analytische rubric een meer holistische rubric kan worden gemaakt. Van links naar rechts worden steeds meer criteria samengevoegd.

### 17.1.3 Taakspecifieke en generieke rubrics

Een tweede onderverdeling in rubrics is te maken tussen taakspecifieke en generieke rubrics. Taakspecifieke rubrics zijn te gebruiken voor specifieke opdrachten. In zo'n rubric

Onvoldoende		Voldoende		Goed	
25%		50%		75%	
100%					
Beginnend		Gevorderd		Geoefend	
Startbekwaam					
0	3	6	8	10	
Laag		Gemiddeld		Hoog	
Expert					
Beginner		Gevorderd		Expert	
Excellent					
Novice		Apprentice		Practitioner	
Expert					
Kennis		Inzicht		Toepassing	
Analyse					
Ok		Good		Very good	
Done!					
Ongestructureerd	Uni-structureel	Multi-structureel	Relationeel		Generaliserend
Het begin is er ...		Werk in uitvoering		Af!	

■ **Figuur 17.3** Voorbeelden van indelingen in beoordelingsniveaus.

komen dan expliciet onderwerpen aan bod die cruciaal zijn voor dit probleem. Bijvoorbeeld bij een rubric voor het ontwerpen van een computerprogramma om kansberekeningsprobleem op te lossen kan als taakspecifiek criterium opgenomen worden: ‘Met het programma is het mogelijk om zowel met als zonder teruglegging te toetsen.’

Generieke rubrics zijn geschikt voor een breder doel. Het kan dan bijvoorbeeld wel gaan om het ontwerp van een computerprogramma, maar het soort problemen dat opgelost moet worden maakt niet uit. De rubric is hierdoor breder inzetbaar, maar geeft studenten op een minder detailniveau feedback.

De keuze voor een taakspecifieke of een generieke rubric is afhankelijk van het soort leerdoelen dat wordt getoetst en of de rubric een formatief of summatief doel heeft (zie ► paragraaf 17.2). Bij specifieke kennis die voor alle studenten geldt, zijn de taakspecifieke rubrics aan te raden. Deze rubrics kunnen dan zowel door docenten – voor het geven van feedback en het scoren – als door studenten – voor zelfbeoordelingen en als communicatiemiddel – worden gebruikt. De generieke rubrics zijn ook voor dezelfde doelen geschikt, maar minder voor het beoordelen van specifieke kennis. Daar staat weer tegenover dat generieke rubrics geschikt zijn voor het toetsen van de ontwikkeling van een (deelaspect) van een competentie over een langere periode.

### 17.1.4 Beoordelingsniveaus

Bij de genoemde indelingen in ► paragraaf 17.1.2 en 17.1.3 gaat het vooral om de formulering van de criteria (de eerste dimensie van een rubric). Over de inrichting van de tweede dimensie (de beoordelingsniveaus) is ook een breed scala aan mogelijkheden in de literatuur terug te vinden. In ■ figuur 17.3 is een aantal mogelijkheden beschreven. Daarin is te lezen dat een rubric ook goed bruikbaar is om tot een wat globalere en meer inhoudelijke kwalificatie te komen, naast een exacte becijfering.

De keuze voor het type beoordelingsniveaus en het aantal beoordelingsniveaus wordt bepaald door de functie die men de rubric toekent. De verschillende niveaus moeten goed van elkaar te onderscheiden zijn en het onderscheid moet nuttig zijn om de ontwikkeling van de studenten te stimuleren. Zo kan de opbouw van de niveaus worden gerelateerd

aan verschillende taxonomieën. Taxonomieën van Bloom (1956), Biggs & Collis (1982) en Anderson & Kratwohl (2001) bijvoorbeeld gaan ervan uit dat een hoger niveau in de taxonomie de lagere niveaus veronderstelt. Als een gewenste prestatie dan op het niveau van analyse is gesitueerd, kunnen de onderliggende beoordelingsniveaus worden aangeduid met 'kennis van' of 'inzicht in'. Als studenten dan aantonen het niveau van 'analyse' te beheersen, worden 'kennis van' en 'inzicht in' ook verondersteld aanwezig te zijn.

### 17.1.5 De indicatoren

---

De indicatoren zijn de beschrijvingen van de verschillende niveaus bij de gestelde criteria. Het zijn de observeerbare en meetbare eenheden. Het spreekt voor zich dat bij het formuleren van de indicatoren helderheid en eenduidigheid is gewenst. Ze moeten ontwikkelingsgericht zijn en een voorbeeld geven van gewenst gedrag. Zeker bij formatief gebruik is het verder van belang dat het taalgebruik in de indicatoren voor studenten begrijpelijk is en dus geen jargon of vage begrippen bevat. Het uitgangspunt KISS (Keep It Smart and Simple) is zeker van toepassing bij het formuleren van indicatoren. Vanuit motivationeel oogpunt verdient het voorkeur de indicatoren zo veel mogelijk in positieve bewoordingen te formuleren. Dat wil zeggen dat ook bij het laagste kwaliteitsniveau zo goed mogelijk moet worden benoemd wat de student wél kan. Het gebruiken van de ik-vorm stimuleert het 'eigenaarschap' bij de student.

Om te komen tot goede indicatoren, kan een docent gebruikmaken van bestaand werk van studenten en zich afvragen waarom een bepaalde prestatie zeer goed, gemiddeld of onvoldoende is. Het antwoord op die vraag is input voor de indicatoren. Daarnaast wordt de kwaliteit van de indicatoren hoger als deze tot stand gekomen zijn uit een dialoog met collega's en/of studenten.

## 17.2 Rubrics voor formatief én summatief toetsen

---

De keuze voor meer of minder criteria, aard en aantal van de beoordelingsniveaus en inhoud van de indicatoren, wordt vooral bepaald door de functie die men de rubric wil geven in het onderwijs.

### 17.2.1 Rubrics en de summatieve toetsfunctie

---

Wanneer men een rubric gebruikt om een cijfer voor geleverd werk of een performance te geven, dient het voor zowel docenten als studenten duidelijk te zijn hoe dat cijfer tot stand komt. Hierbij is in eerste instantie het validiteitscriterium van belang: meet je met de rubric wat je wil meten? Daarnaast is het betrouwbaarheidscriterium ook belangrijk. De criteria moeten helder zijn en het aantal beoordelingsniveaus moet het hele spectrum van onvoldoende tot '10' mogelijk maken, er moet een goede spreiding mogelijk zijn en er moet onderscheid worden gemaakt tussen de studenten die de prestatie goed hebben

uitgevoerd en studenten die dat niet hebben gedaan. Het moet duidelijk te zien zijn hoe de indicatoren afgeleid zijn van de ‘standaarden’ of te behalen doelen.

Methoden om de cesuur te bepalen, hangen af van de aard van de rubric. Bij een holistische rubric wordt er één oordeel gegeven op basis van de inschatting van de gezamenlijke criteria. Het is bij zo’n oordeel van groot belang dat er voor elk niveau voorbeelden zijn om het oordeel aan te ijken. De ijkpunten (benchmarks) kunnen komen van eerder beoordeeld werk. Het is van ook van belang om oordelen te vergelijken en de voor het ijken gekozen voorbeelden regelmatig te herwaarderen.

Dit geldt eveneens bij het scoren met een analytische rubric. Bij zo’n rubric kan het maximaal te verkrijgen aantal punten op 100% worden gesteld en kan een gewicht worden gegeven aan een hele kolom met indicatoren (wat krijg je als je overal in kolom 2 hebt gescoord, bijvoorbeeld). Gewicht toekennen aan domeinen is ook mogelijk. Dat betekent dat voor bepaalde rijen in de rubric percentages of cijfers worden gegeven, of zelfs per ‘vakje’. Dat geeft de mogelijkheid om per domein af te wegen hoe belangrijk een bepaald aspect van het werk is ten opzichte van andere aspecten.

Het gevaar van te ver doorvoeren van wegenen is dat de rubric steeds meer als een soort afvinklijst gaat fungeren. Studenten kunnen zich gaan richten op onderdelen of domeinen waar een hoog percentage van het cijfer te behalen is en andere onderdelen van het werk verwaarlozen, docenten die cijfers geven kunnen zich verliezen in details en voorbijgaan aan de kwaliteit in samenhang. Om ervoor te zorgen dat bepaalde criteria buiten beschouwing worden gelaten door de studenten, kan een voorwaardelijke eis worden gesteld dat op alle criteria in ieder geval boven een bepaald niveau moet worden gescoord. De beste manier om ongewenste effecten bij het cijfergeven met gebruik van een rubric te voorkomen is om dit met meerdere beoordelaars te doen.

### 17.2.2 Rubrics en de formatieve toetsfunctie

---

Een van de sterke punten van rubrics is dat ze geschikt zijn om precieze, inhoudelijke informatie te geven over de geleverde leerprestatie. Ze stellen docenten en studenten in staat om vast te stellen wat er tot nu toe is geleerd, tot op welk niveau en wat er nog te leren is. Rubrics beantwoorden daarmee goed aan de ‘feed forward’-functie van toetsing (Hattie & Timperley, 2007) en kunnen dus goed worden ingezet in het leerproces.

Doordat de indicatoren in de rubric laten zien wat er nog moet worden gedaan om tot een betere prestatie te komen, kunnen studenten en docenten de leervorderingen zichtbaar maken. Hiermee is de rubric zeer geschikt als didactisch instrument en als instrument om de zelfregulatie van studenten te bevorderen.

Docenten kunnen studenten vragen om ter voorbereiding op onderwijsbijeenkomsten of tijdens onderwijsbijeenkomsten een rubric te laten invullen. Tijdens de les wordt deze dan besproken. Op deze manier zet de docent de rubric in als een zelfbeoordelingsinstrument. Aangezien de zelfbeoordelingen van studenten niet altijd betrouwbaar zijn, is het belangrijk dat de docent de studenten feedback geeft op hun eigen inschatting in de rubric: ‘Waarom kan je zien dat je op een bepaald niveau functioneert?’

Studenten kunnen ook elkaar aan de hand van een rubric beoordelen. Door de beschikbaarheid van de rubric en het expliciet bespreken van de rubric met de studenten, is de interpretatie van de criteria voor alle studenten redelijk gelijk. Studenten kunnen deze criteria dan ook hanteren voor een oordeel over het werk van hun medestudenten.

Met name voor deze formatieve functie van de rubric is het wenselijk de indicatoren zo veel mogelijk in positieve bewoordingen te formuleren. De leerprestatie is immers een punt 'op weg' in het leerproces. Wanneer een docent aan het begin van een cursus een rubric introduceert en uitlegt hoe de rubric werkt, hebben de studenten meteen een goed inzicht in de doelen en het gewenste eindniveau, bijvoorbeeld voor het maken van een werkstuk, het bijhouden van een portfolio of het schrijven van een essay of analyse. De docent biedt zo de mogelijkheid om aan de hand van de rubric studenten te laten oefenen, hun stand van zaken in het leerproces of bij de leerprestatie vast te laten stellen (de zogenaamde 'feed up'-fase van feedback) en elkaar feedback te laten geven.

#### ■ Het bevorderen van zelfregulatie

Rubrics maken door hun structuur transparant wat er wordt getoetst en welke kwaliteitscriteria daarbij horen. Door de stapsgewijze opbouw van de beoordelingcriteria wordt er per criterium in zekere zin ook een leerproces zichtbaar. Studenten kunnen aan de hand hiervan zelf vaststellen op welk punt ze zich bevinden in het leerproces en wat de volgende stap is. Daardoor is een rubric geschikt bij self- en peerassessment, het vergroot de zelfstandigheid van studenten en levert een bijdrage aan het proces van leren leren.

Waar dat kan is het aanbevolen om de rubric geheel of gedeeltelijk samen met de studenten te ontwikkelen. Het gevoel van mede-eigenaarschap motiveert en draagt mogelijk bij aan een kritische houding ten opzichte van de studieprestatie. Ervaringen met studenten die zelf meewerken door de rubric te (helpen) construeren, laten zien dat ze niet alleen bewuster omgaan met feedback, maar ook bewuster zijn van het belang van transparante, objectieve en betrouwbare beoordeling. Studenten die werken met rubrics krijgen beter inzicht in leerdoelen, en ook op de complexiteit van de taak en dit op zich kan het leren bevorderen (Panadero & Jonsson, 2013).

### 17.3 Curriculumbrede inzet

---

Wanneer er in een curriculum een doorlopende leerlijn is, dient een goed ontworpen rubric verschillende doelen. Het doel van de leerlijn is om de studenten van een ingeschat beginniveau naar een gewenst eindniveau te brengen. Door middel van een rubric is het voor studenten bij de aanvang van de leerlijn al duidelijk wat het eindniveau is. Ze weten dan wat ze al wel kunnen en wat ze nog moeten doen. De eerste functie van de rubric is in dit geval het bieden van een raamwerk waarmee de docent vooraf doelen bespreekt, de 'feed up'-functie van goede feedback (Hattie & Timperley 2007).

De belangrijkste functie bij het herhaaldelijk gebruiken van eenzelfde rubric voor gelijksoortig werk in een doorlopende leerlijn is de 'feed forward'-functie. Deze houdt in dat de docent de studenten met behulp van de rubric vertelt welk niveau van beheersing is behaald en wat er nog moet worden gedaan om tot het volgende of zelfs tot het uiteindelijk gewenste niveau te komen. Bij het schrijven van bijvoorbeeld werkstukken kan worden

aangegeven dat voor jaar 1 de prestatie voldoende is, maar dat in jaar 2 op een volgend beheersingsniveau moet worden gepresteerd. Als de student vanaf het begin vertrouwd is met de criteria en de indicatoren van een rubric en de docent ook consequent naar de rubric verwijst, kan de student zelf de 'feed forward'-functie gaan hanteren en zichzelf sterker gaan sturen.

Door curriculumbreed te werken met een rubric voor vergelijkbare prestaties is het niet meer nodig dat docenten in elk jaar opnieuw de criteria gaan bepalen. Over de jaren heen kunnen criteria worden opgesteld; de uitwerking in de niveaus en indicatoren maakt het dan mogelijk om in een hoger jaar terug te verwijzen naar de criteria van afgelopen jaar. Doordat ook de feedback van elk jaar gebaseerd is op dezelfde rubric, kunnen docenten uit hogere jaren van de studenten vragen om de feedback (en dan vooral ook de feed forward) van het jaar ervoor erbij te pakken en dat als startpunt te gebruiken voor de nieuwe prestatie op een iets hoger niveau. Dit vereist dan wel een curriculumbreed overleg tussen docenten op vergelijkbare thema's.

#### 17.4 Het beoordelen van de kwaliteit

---

De basiscriteria voor een goede rubric zijn dat het de juiste inhoud dekt, dat de criteria duidelijk zijn, het aantal niveaus overeenkomt met het doel en dat de niveaus duidelijk zijn gedefinieerd. Om kwaliteit van rubrics te beoordelen is in [tabel 17.3](#) een rubric opgenomen.

#### 17.5 Een stappenplan voor de constructie

---

Op grond van de afwegingen voor het gebruik van rubrics volgt hieronder een stappenplan voor de constructie van een rubric.

1. *Reikwijdte en context bepalen.*

Allereerst wordt bepaald of er sprake is van een generieke of een taakspecifieke rubric en wat de precieze context is van de rubric (opleiding, jaar, vak, type vaardigheid, enzovoort). Zie ► par. 17.1.3.

2. *Doelstelling bepalen.*

Vervolgens wordt een expliciete verbinding gelegd met de (deel)competentie of de doelstelling waarvoor de rubric de toetsvorm is. Dit versterkt de validiteit van de rubric en draagt bij aan de transparantie van deze toetsvorm. Zie ► par. 17.1.1, 17.2.1 en 17.2.2.

3. *Toetscriteria opstellen.*

De doelstellingen vormen de bron voor de toetscriteria. Het is hier van belang te bepalen in welke mate criteria kunnen worden geclusterd of niet. Zie hiervoor de opmerkingen in ► par. 17.1.2 over holistische en analytische rubrics.

4. *Toetsfunctie bepalen.*

Om te komen tot een verdere detaillering van de rubric wordt bepaald op welke manier men de rubric wil gaan gebruiken: summatief of formatief. Zie ► par. 17.2.1 en 17.2.2.



■ **Tabel 17.3** Rubric voor het beoordelen van rubrics.

	Sterk	Gemiddeld	Zwak
<b>Rubric dekt juiste inhoud</b>	inhoud representeert het domein, is in lijn met de leerdoelen; rubric geeft op alle punten inzicht of een student presteert in lijn met de verwachting	inhoud representeert domein grotendeels, is grotendeels in lijn met de leerdoelen; aantal belangrijke thema's/onderwerpen ontbreken of irrelevante doelen worden benoemd	onduidelijk welke leerdoelen met rubric beoordeeld moeten worden; leerdoelen dekken voornamelijk minder belangrijke thema's/onderwerpen
<b>Criteria zijn duidelijk</b>	criteria zijn begrijpelijk; aantal criteria sluit aan bij moeilijkheid leerdoel; uit criteria blijkt welke onderdelen belangrijker zijn dan andere; criteria zijn onafhankelijk van elkaar	enkele criteria moeten opgesplitst of samengevoegd worden; nadruk op sommige criteria is te zwaar of te licht	rubric is te holistisch of te analytisch; criteria worden met elkaar gecombineerd die niet bij elkaar horen; balans tussen criteria ontbreekt; criteria zijn afhankelijk van elkaar
<b>Aantal niveaus komt overeen met het doel</b>	voldoende niveaus om studenten hun voortgang te tonen; voldoende niveaus om onderscheid te maken tussen twee opeenvolgende niveaus	onduidelijk of een goed onderscheid tussen prestaties van studenten gemaakt kan worden	aantal niveaus is niet geschikt om criteria te beoordelen; te veel of te weinig niveaus om goed onderscheid te maken; te veel ruimte voor eigen interpretatie.
<b>Indicatoren zijn duidelijk omschreven</b>	bij elk criterium en niveau is een indicator; beschrijvingen zijn herkenbaar voor de prestaties van de studenten; onafhankelijke (getrainde en ervaren) gebruikers komen tot dezelfde score; indicatoren zijn beschrijvend, niet beoordelend.	alleen hoogste en/of laagste niveau is beschreven; niet alle beschrijvingen zijn even herkenbaar voor de prestaties van studenten; niet zeker of onafhankelijke (getrainde en ervaren) gebruikers komen tot dezelfde score; indicatoren zijn meestal beschrijvend, soms evaluatief	indicatoren zijn niet beschreven, er is alleen rangorde aangegeven; woordgebruik is vaag; onwaarschijnlijk dat onafhankelijke gebruikers tot eenzelfde scoring komen; indicatoren zijn voornamelijk evaluatief.

Gebaseerd op Arter & Chappuis, 2006.

##### 5. *Beoordelingsniveaus formuleren.*

De keuze voor een summatieve of formatieve toetsfunctie is vervolgens bepalend voor het bepalen van aantal, richting en titel van de beoordelingsniveaus. Zie

► par. 17.1.4.

##### 6. *Indicatoren formuleren.*


Per toetscriterium wordt vastgesteld wat het laagste en wat het hoogste beoordelingsniveau is. Daarna worden de tussenliggende beoordelingsniveaus ingevuld. Zie

► par. 17.1.5.

##### 7. *Normering bepalen.*

Met name bij summatief gebruik worden vervolgens becijfering, normering en cesuur bepaald. Zie ► par. 17.2.1.

8. *Betrouwbaarheid en 'fine tuning'.*

Rubrics moeten voor een belangrijk deel in het feitelijk gebruik hun kwaliteit bewijzen. Vooral wanneer én studenten én meerdere beoordelaars met de rubric werken, kunnen de toetscriteria, kwaliteitsniveaus en indicatoren steeds beter worden afgebakend en eenduidiger geformuleerd. De kwaliteit van rubrics kan met behulp van de rubric in  tabel 17.3 beoordeeld worden.

# Toetsen met EVC

*Ruud Klarus en Desirée Joosten-ten Brinke*

## **Inleiding**

Wat is EVC?

Wat is de rol van assessor bij EVC?

Wat is het verschil tussen EVC voor het onderwijs en EVC voor de arbeidsmarkt?

Welke beoordelingsinstrumenten worden gebruikt bij EVC?

Waarom is EVC ooit ontstaan?

Hoe ontwikkel je EVC-procedures?

- 18.1      Wat is EVC? – 206**
- 18.2      Toetsmethoden voor een EVC-procedure – 207**
- 18.3      Het ontwikkelen van een EVC-procedure – 208**
- 18.4      De kwaliteitscode EVC – 208**
- 18.5      Instrumenten – 209**
- 18.6      Voorbeelden van EVC-procedures – 211**
- 18.7      Aandachtspunten voor EVC – 215**
- 18.8      Ten slotte – 216**

Erkenning van Verworven Competenties (EVC) is aan het eind van de vorige eeuw in het Nederlandse onderwijs geïntroduceerd. Belangrijkste reden hiervoor was het beter benutten van menselijk kapitaal en de wens te erkennen dat competenties ook worden ontwikkeld door het opdoen van ervaring en door leren op de werkplek. Twee voorbeelden maken dit duidelijker. Wat te doen met de medewerker van een technisch constructiebedrijf met een mbo-diploma die ongeveer acht jaar ervaring heeft, verschillende cursussen op zowel technisch als management- en marketinggebied heeft gevolgd, is opgeklommen tot hoofd van de afdeling verkoop en zich vervolgens aanmeldt bij de opleiding technische bedrijfskunde met de vraag: 'Kan ik hier mijn hbo-diploma halen zonder dat ik op herhaling hoef?' Hoe te reageren op de vraag van het ministerie van Defensie aan de directeur van Opleidingskunde of de medewerkers die zij intern hebben opgeleid in uiteenlopende logistieke, technische en operationele functies, en die daarna zijn opgeleid tot instructeur, een opleiding kunnen krijgen waarin rekening wordt gehouden met de door hen verworven competenties?

Instellingen voor hoger onderwijs krijgen te maken met kandidaten die al veel kennis en kunde hebben opgedaan op andere plaatsen. Docenten proberen hiermee rekening te houden. Moeten de kandidaten echt aan elk onderdeel van een onderwijsprogramma meedoen omdat het in het programma staat – hoewel zij daar al competent in zijn - of kunnen zij een leerweg op maat krijgen? Hoe weet een docent of iemand competent is op een bepaald gebied? Voor het hoger onderwijs is het belangrijk om na te gaan in hoeverre dergelijke kennis en vaardigheden kunnen meetellen voor het behalen van het diploma.

## 18.1 Wat is EVC?

In een kennisintensieve samenleving wordt steeds meer kennis ontwikkeld en benut buiten het geïstitutionaliseerde onderwijs. Er wordt bijgevolg ook steeds meer buiten het instituut school geleerd. EVC is in het Nederlandse onderwijs geïntroduceerd op het moment dat competentiegericht opleiden zijn intrede deed (Klarus, 1998). Hoe en waar je geleerd hebt, is minder belangrijk dan wat je hebt geleerd. EVC is daarom gericht op het toetsen van leerresultaten, ongeacht de gevolgde leerweg; deze kan formeel, non- of informeel zijn. Het gaat vooral om de specifieke combinatie van context, activiteiten, individuele en begeleidingskenmerken, en of er al dan niet en effectief wordt geleerd. Formeel leren houdt in dat er wordt geleerd in een formele onderwijssetting, zoals universiteiten of hogescholen. Daar wordt doelbewust geleerd voor een diploma met een algemeen erkende waarde. Non-formeel leren is ook doelbewust en voorgestructureerd, maar het verstrekte certificaat heeft een minder grote civiele waarde dan bij formeel leren. Denk bijvoorbeeld aan deelname aan bedrijfstrainingen en conferenties. Er vindt ook heel veel informeel leren plaats. Deze manier van leren heeft niet een van tevoren beoogd doel, is niet voorgestructureerd en levert ook geen getuigschrift op. Voorbeelden hiervan zijn het uitvoeren van taken op het werk of het bezoeken van een museum (Colley, Hodgkinson & Malcolm, 2002). Bij EVC gaat het daarom om leerwegaafhankelijk toetsen. Maar hoe meet je of iemand de vereiste competenties op bijvoorbeeld hbo-niveau heeft?

Toetsen van competenties veronderstelt beoordelingsinstrumenten en technieken die passen bij de te beoordelen competentie. Op voorhand is geen enkele techniek

uitgesloten. Voorwaarde is dat de toetsvorm aansluit op het doel van de toets en valide is (zie ► hoofdstuk 19). Een belangrijke vraag is of de toetsprocedure bijdraagt aan het individuele, organisatie en maatschappelijke doel van toetsen, zoals persoonlijke ontwikkeling of civiel effect. Dat een toetsprocedure bijdraagt aan de beoogde resultaten (competenties ontwikkelen) is wellicht belangrijker dan de vraag of elke afzonderlijke toets wel betrouwbaar is. Dit sluit aan bij de verschuiving van het toetsen met de nadruk op momentopnames en voorspellende ambities (het meer psychometrisch toetsen) naar de nadruk op het volgen en sturen van ontwikkeling (het edumetrisch toetsen).

EVC is geschikt voor verschillende doelen, bijvoorbeeld het vergroten van de transparantie van het onderwijs of voor het verstrekken van vrijstellingen. De Onderwijsraad wees er in 2004 al op dat EVC de transparantie van het hoger onderwijs zou kunnen vergroten. ‘Voorwaarde is wel dat dit (EVC) het niveau van traditionele vrijstellingsprocedures (bij EVC-erkenning) en de handhaving van separate leeromgevingen (bij duale opleidingen) overstijgt.’ (p. 48) EVC is van meerwaarde voor het onderwijs als het meer is dan alleen een vrijstellingsprocedure en wordt ontwikkeld vanuit het besef dat effectief leren te maken heeft met de vraag of het leren betekenisvol is. Een belangrijke variabele voor betekenisvol onderwijs is de mate waarin rekening wordt gehouden met voorgaande leer- en werkervaringen. EVC-procedures kunnen zo de basis leggen voor individueel maatwerk. Een docent kan in deze EVC-procedures een rol spelen in de functie van assessor of begeleider.

## 18.2 Toetsmethoden voor een EVC-procedure

---

Beoordelen, ook competentiegericht beoordelen, is en blijft het zoeken naar een productieve balans tussen ‘je wilt wat’ en ‘je moet wat’. De spanning tussen persoonlijke doelen, opleidings- of organisatie doelen respectievelijk de eisen van een kwalificatiestructuur wordt door Mabry (1999) geschetst als een moeizame combinatie tussen twee beoordelingsparadigma’s: een persoonsgericht en een contextueel paradigma. Een persoonsgericht paradigma is een vorm van beoordeling waarbij inhoudsvaliditeit wordt gecombineerd en in balans is met persoonlijke relevantie. In een persoonsgericht paradigma worden de beoordelingscriteria door het individu zelf vastgesteld, zijn dus persoonsgerelateerd want ze moeten bijdragen aan het bereiken van persoonlijke doelen. Omdat ze individuen de gelegenheid bieden te laten zien wat ze kunnen en welke competenties ze hebben ontwikkeld, zijn de beoordelingscondities en de tijdsduur van een beoordeling afgestemd op de beoordeelde.

In beoordelingsprocedures binnen het contextuele paradigma geven niet de eisen van de kandidaat, maar die van de context c.q. de beroepspraktijk de doorslag. Deze hebben de volgende karakteristieken:

- de beoordelingsinhoud is niet volledig gestandaardiseerd maar in overeenstemming met de actuele eisen van de authentieke leer- en werkcontext;
- de context van het leren en werken is in overeenstemming met de context van het beoordelen, door kandidaten zo veel mogelijk in hun eigen leer- en werkcontext te beoordelen;

- docenten worden ondersteund en opgeleid om hun eigen professionele oordeelsvermogen te benutten in de beoordelingen;
- kandidaten kunnen ook zichzelf beoordelen, en beoordelingsresultaten worden niet alleen gebruikt om anderen te informeren, maar juist ook om kandidaten te ondersteunen bij de planning van hun eigen ontwikkeling en bij het inrichten van persoonlijke leerwerkplannen;
- beoordelingscriteria zijn gerelateerd aan de eisen uit de beroepspraktijk en kunnen – in tegenstelling tot normgerelateerde beoordeling – niet worden bijgesteld op basis van beoordelingsresultaten van individuen. Het contextuele paradigma impliceert dus criteriumgerelateerde beoordeling.

### 18.3 Het ontwikkelen van een EVC-procedure

---

Om competenties te beoordelen is het van belang om te weten in welke context (domein en niveau) en naar welke erkenning wordt gestreefd. Voor het ontwikkelen van een EVC-procedure bestaat geen eenduidig stappenplan dat gegarandeerd tot succes leidt. Het ontwikkelen van EVC-procedures is daarvoor te afhankelijk van de combinatie van doelen, contexten en beschikbare voorwaarden en middelen. Dit neemt niet weg dat er wel vragen in een bepaalde volgorde moeten worden gesteld en beantwoord om ten minste een minimale garantie voor succes te hebben. In [tabel 18.1](#) staan de belangrijkste ontwikkelvragen. Per vraag zijn, niet uitputtend, een of meer aspecten genoemd die het verschil in doel, kwaliteit en consequenties van de procedure bepalen.

### 18.4 De kwaliteitscode EVC

---

EVC is een begrip waarop geen patent staat. Dit betekent dat wanneer een organisatie activiteiten verricht met als doel het erkennen van competenties, deze activiteit EVC mag worden genoemd. De afgelopen jaren is er landelijk echter onderzocht op welke manier wildgroei aan procedures kan worden voorkomen en hoe EVC-procedures een kwaliteitsstempel kunnen krijgen. Dit kwaliteitsstempel is de ‘Kwaliteitscode EVC’. In deze code staan uitgangspunten ten aanzien van vijf onderdelen, te weten het doel van EVC, de rechten van de kandidaat, het EVC-onderzoek, de assessoren en begeleiders en de kwaliteitszorg (zie [tabel 18.2](#)). Als een organisatie werkt volgens deze EVC-code, kan er een erkenning worden aangevraagd en bij een positief resultaat worden de erkende procedures in het EVC-register opgenomen (Kenniscentrum EVC, 2011). Nederland kent ongeveer honderd organisaties die door het kenniscentrum zijn erkend als EVC-aanbieder; dit betekent dat hun procedure voldoet aan de kwaliteitscode EVC.

Aan de onderdelen van de code zit een normering gekoppeld op basis waarvan organisaties hun EVC-procedure kunnen laten beoordelen door beoordelende instanties. De code heeft enerzijds geleid tot meer kwaliteitsborging voor EVC gericht op erkenning van verworven competenties met een civiel effect. Anderzijds blijken dezelfde eisen te leiden tot meer bureaucratie met als gevolg dat ze aanbieders van EVC-procedures afschrikt met als gevolg dat er steeds minder aanbieders zijn.

■ **Tabel 18.1** Eerst tot tien tellen: tien vragen bij het ontwikkelen van een EVC-procedure.

1. wat: visie & doel	waardering en zelfbeeldverheldering met het oog op loopbaankeuzes, vrijstellingen, certificering en diplomering; bewaking van de congruentie tussen doel, visie en alle volgende keuzes in het ontwerp
2. waarom: belangen	wie heeft het meeste profijt van de EVC-procedure; wie is de eigenaar?
3. wie: betrokkenen en verantwoordelijken	doelgroep/kandidaten; uitvoering; opleidingsmanagement/directie/bestuur; docent-beoordelaar: duobeoordeling gewenst, maar ook haalbaar? (triangulatie); student/kandidaat; werkveld/praktijkbeoordelaar; opleiders/docenten; begeleiding kandidaten
4. waar	aanmelding; centraal EVC-loket, EVC-loket per faculteit, per opleiding of per studierichting of vakgroep; beoordeling in instituut in de beroepspraktijk (waar) of beide; triangulatie?
5. wanneer: procedure	ritme van de opleiding versus ritme van de praktijk; planbaarheid versus improvisatie; welke activiteiten wanneer: voorlichting, intake, beoordeling, uitslag, voortgang, meerdere toetsmomenten haalbaar?; draaiboeken voor ontwerpen, onderhouden en bijstellen van de procedure en voor de toetsprocedure zelf; uitsplitsen naar kandidaten, beoordelaars en administratie
6. hoe: keuze van instrumenten	standaard, kwalificatiestructuur of beroepscompetentieprofielen en toetsindicatoren; welk soort competenties en welke bewijzen zijn daarvoor adequaat; congruentie tussen type bewijzen, type instrumenten en type competentie; wat voor materiaal (informatie, opdrachten, scoringsformats, instructie, draaiboek, enzovoort) is noodzakelijk voor de: 1 kandidaat, 2 voorlichter, 3 intaker, 4 instituutbeoordelaar, 5 beoordelaar in de praktijk, 6 administratie, 7 examencommissie, 8 kwaliteitsmedewerker, 9 externe kwaliteitsbeoordelaar (inspectie, vbi, auditor)
7. kwaliteitsborging en ontwikkeling	aantoonbare, gerealiseerde kwaliteit: kwaliteitsonderzoek; kwaliteitsontwikkeling versus controle (ink-model)
8. consequenties	maatwerk naar inhoud en planning, verkorte route (bijvoorbeeld vrijstelling propedeuse), vrijstellingen onderdelen/modules/blokken/taken, diplomering
9. kosten	procedure onderdeel van de opleiding, voorafgaand aan de opleiding, op zichzelf staand; kosten beroepsprocedure
10. invoering, risico's en weerstanden	hoe, waar en wanneer zorgen we voor succesvolle eerste ervaringen en EVC-ambassadeurs

## 18.5 Instrumenten

In EVC-procedures is het gebruikelijk om verschillende instrumenten in te zetten die het mogelijk maken om de competenties van de lerenden te meten. De meest gebruikte zijn een portfolio-assessment en het criteriumgericht interview. Deze twee worden hierna toegelicht. Ook andere instrumenten zoals een performance assessment (zie ► hoofdstuk 16) of zelfs een kennistoets zijn mogelijk, afhankelijk van de standaard die wordt gebruikt.

### ■ Het portfolio-assessment

In het portfolio-assessment beoordelen de assessoren (beoordelaars) of de kandidaat bewijzen voor competent handelen heeft aangeleverd in het portfolio en deze heeft onderbouwd conform de eisen van de standaard. Voor het portfolio is meestal een vaste opzet vereist. Onderzoek naar portfolio-assessment bij EVC heeft aangetoond dat kandidaten

<span style="color: #0070C0;">■</span> <b>Tabel 18.2</b> De kwaliteitscode EVC (Kenniscentrum EVC, 2011).	
1. doel EVC	Het doel van EVC is het zichtbaar maken, waarderen en erkennen van individuele competenties. Het erkennen van verworven competenties heeft een waarde in zichzelf en draagt bij aan employability. EVC leidt in veel gevallen tot verdere loopbaangerelateerde persoonlijke ontwikkeling.
2. rechten kandidaat	EVC beantwoordt aan de behoefte van het individu. Rechten en afspraken zijn duidelijk verwoord en geborgd.
3. onderzoek	Procedures en instrumenten zijn betrouwbaar en gebaseerd op erkende standaarden. Vertrouwen is een sleutelbegrip. Vertrouwen heeft te maken met goed gedefinieerde standaarden, civiel effect en duidelijke informatie over de manier waarop assessments worden uitgevoerd en op basis van welke argumenten conclusies zijn getrokken.
4. assessoren en begeleiders	Assessoren en begeleiders zijn competent, onafhankelijk en onpartijdig. Onafhankelijkheid en onpartijdigheid zijn cruciale factoren binnen de beoordeling en zijn ingebed in de rollen en verantwoordelijkheden van de betrokken assessoren. Het is van groot belang om onnodige vermenging van rollen te voorkomen. Onpartijdigheid kan worden versterkt door training en door deel te nemen aan lerende netwerken.
5. kwaliteitszorg	De kwaliteit van de EVC-procedure is geborgd en wordt continu verbeterd. De kwaliteit van de EVC-procedure en het bij de procedure gehanteerde instrumentarium zijn geborgd. Er vinden regelmatig evaluaties plaats. De resultaten daarvan worden verwerkt in verbeteracties.

het heel lastig vinden om op een goede manier een portfolio samen te stellen (zie bijvoorbeeld Colley e.a., 2002). De assessoren beoordelen de portfolio's aan de hand van criteria. Bewijzen moeten relevant, authentiek, valide, recent, voldoende en op het juiste niveau zijn. De combinatie van concrete bewijzen en een onderbouwing van de reden voor het gebruiken van die specifieke bewijzen levert voor assessoren veelal voldoende informatie op om een beslissing over erkenning van competenties te nemen (Joosten-ten Brinke, Sluijsmans & Jochems, 2010).

#### ■ Een criteriumgericht interview

Een criteriumgericht interview is een individueel vraaggesprek tussen een kandidaat en (bij voorkeur twee) assessoren. Het is gericht op de beschreven competenties uit de standaard en competentiebewijzen uit het portfolio. De assessoren onderzoeken op een gestructureerde manier in hoeverre een kandidaat voldoet aan de criteria. Dit maakt kennis en vaardigheden zichtbaar die niet direct uit het portfolio naar voren komen, zoals het toetsen van inzicht in beroepsdomein of het toetsen van beroepshouding. De STARRT-methode is geschikt om het gesprek goed te structureren. Hiermee worden achtereenvolgens Situatie, Taak, Actie, Resultaat, Reflectie en Transfer van een thema besproken waarmee kandidaten hun competenties aantonen. Bij dit interview is het van belang dat de kandidaten hun eigen activiteiten beschrijven (en dus in de 'ik'-vorm praten) en dat open vragen worden gesteld zodat duidelijk wordt of de kandidaat al dan niet een competentie beheerst.

#### ■ De rol van docenten

Zoals in de kwaliteitscode is aangegeven dienen de assessoren competent, onafhankelijk en onpartijdig te zijn. Deze competentie houdt in dat de assessoren inhoudelijk deskundig zijn en dat zij kennis hebben over EVC. Docenten kunnen op basis van hun inhoudelijke





■ **Figuur 18.1** Vijf varianten waarin EVC, of aan EVC verwante procedures worden gebruikt.

deskundigheid in de gehanteerde standaard (onderwijsvisie, inhoud, didactische werkvormen en niveau van de opleiding) worden aangesteld als assessor. Zij zullen dan in een training vooral nog moeten worden geschoold in de kennis over EVC. Dit houdt in dat zij kennis hebben van de doelstellingen, de kwaliteitscode, procedure en processen van EVC en dat zij de verschillen en overeenkomsten kennen in doelen en positionering van EVC ten opzicht van vrijstellingsbeleid en intake. Daarnaast kunnen docenten optreden als adviseur. Ook hierbij zijn beide deskundigheden noodzakelijk.

## 18.6 Voorbeelden van EVC-procedures

EVC-procedures zijn er in velerlei vormen en maten (ook: ► [www.kenniscentrumEVC.nl](http://www.kenniscentrumEVC.nl)). Hierna volgen twee voorbeelden. In het eerste voorbeeld is te zien dat EVC zich afspeelt op twee fundamenteel verschillende speelvelden (zie ■ figuur 18.1). Het ene is de arbeidsmarkt (in het voorbeeld de sector jeugdzorg) en het andere de onderwijssector (in het voorbeeld de hbo-opleidingen SPH en Pedagogiek). Dit levert conceptuele verschillen op, maar leidt ook tot verschillen in afstemming op doelgroepen, opdrachtgevers en in uitvoeringspraktijken. Door beide speelvelden tegelijkertijd te willen bespelen, dreigt EVC veel van haar oorspronkelijke doelen, hoe en waar dan ook verworven competenties erkennen, niet te realiseren.

De kwaliteitsborging in het onderwijs valt binnen de wettelijke kaders van de WHW (bij variant 4 in ■ figuur 18.1), EVC voor de poort van onderwijsinstellingen). Als het doel van EVC dan ook is om een groter aantal medewerkers tot hbo-niveau op te leiden, is de vraag of EVC buiten die instellingen moet plaatsvinden. EVC binnen de onderwijssector is bedoeld als erkenning, verzilvering van hoe en waar dan ook verworven competenties op basis van de landelijk vastgestelde kwalificatiestructuren (mbo), beroepsprofielen of standaarden (hbo). EVC-procedures in de onderwijssector moeten voldoen aan dezelfde kwaliteitseisen die voor de reguliere examinering worden gebruikt. Om die reden heeft het ministerie van OCW dan ook besloten het valideringssysteem aan te passen, zodat validering zowel vóór als ná de poort van het formele onderwijs plaatsvindt, afhankelijk van het loopbaandoel van de kandidaat (Stuurgroep EVC, 2013).

Op de arbeidsmarkt heeft EVC vooral het doel om verworven competenties te herkennen. Het beschrijven van de ervaringen en mogelijk daardoor verworven competenties is

daarmee voldoende. De eisen aan betrouwbaarheid en acceptatie van beoordelingen die leiden tot een erkenning en verzilvering in termen van een (branche)certificaat (variant 2) of diploma (variant 3) zijn strenger omdat er aan deze erkenning rechten kunnen worden ontleend. Rechten die te maken hebben met loopbaanontwikkeling en toelating tot opleidingen. Voor het branchecertificaat kan een door de branche aangewezen accreditatieorgaan toezien op de kwaliteit.

### Voorbeeld 1: Een EVC-procedure in het kader van jeugdzorg

Het Avans EVC-centrum organiseert EVC-trajecten voor drie specifieke branchestandaarden: 'vakvolwassen jeugdzorgmedewerker', 'vakbekwaam jeugdzorgwerker' en 'gevorderd beginner' (Van Bentum, 2012). Deze standaarden sluiten aan bij de behoefte van diverse jeugdzorginstanties aan het erkennen van de werkervaring van jeugdzorgmedewerkers. Het overgrote deel van de jeugdzorgwerkers is al in het bezit van een mbo-getuigschrift. Wanneer kandidaten voldoen aan de standaard 'vakvolwassen jeugdzorgmedewerker' worden zij opgenomen in een centraal register. Jeugdzorg NL heeft aangegeven dat een groot aantal werkgevers wil dat hun medewerkers een erkend hbo-diploma halen. EVC kan hierbij mogelijkheden bieden om ervoor te zorgen dat de medewerkers via een maatwerkprogramma het diploma Sociaal Pedagogisch Hulpverlener (SPH) of Pedagogiek, uitstroomprofiel jeugdzorg, kunnen behalen.

Avans Hogeschool werkt in dit traject samen met Inholland en de CHE. Aangezien Avans Hogeschool de enige partner is met een erkend EVC-centrum, is zij eindverantwoordelijke in het EVC-traject.

Om te beginnen met het EVC-traject zijn meetinstrumenten ontwikkeld op basis van de branchestandaarden. In dit geval worden de meetinstrumenten goedgekeurd door de examencommissie van de Academie voor Sociale studies. Voor de validering van het EVC-traject is het noodzakelijk dat er valide meetinstrumenten zijn. Hiervoor wordt een vertaalslag gemaakt van de branchestandaard (die taakgeoriënteerd is) naar een meetinstrument op basis van competenties. Hierbij wordt aangegeven op basis van welke beroepsproducten een kandidaat een competentie kan aantonen. Tevens kunnen kandidaten met een zelfevaluatie voor aanvang van het EVC-traject een inschatting maken van hun vermoedelijke niveau.

#### *Uitvoering van het EVC-traject*

Het EVC-traject wordt verzorgd door het Avans EVC-centrum. In tegenstelling tot reguliere EVC-trajecten wordt in dit traject gewerkt met zogenoemde werkveldassessoren. Dit zijn assessoren die binnen het werkveld zijn geworven. Zij treden altijd als tweede assessor op en moeten hiervoor een training volgen.

#### *Verzilvering van EVC*

De branchestandaarden zijn te vertalen naar de landelijke beroepscompetenties SPH en Pedagogiek, uitstroomprofiel jeugdzorgwerker, zodat na het doorlopen van de EVC-procedure bij een positief assessment de competenties erkend worden en door middel van vrijstellingen bij de opleiding kunnen worden verzilverd.

#### *Maatwerktrajecten*

Wanneer kandidaten via het EVC-traject hebben aangetoond dat zij voldoen aan de standaard vakvolwassen jeugdzorgwerker, kunnen zij deelnemen aan de ontwik-

kelde maatwerktrajecten. Bij de opleiding zullen dan nog wel een afstudeertraject en een aantal modules/minoren moeten worden gevolgd.

*Tips voor andere EVC-ontwikkelaars*


De conclusie van een evaluatie van zo'n EVC-traject in de jeugdzorg (Duvekot, 2012) is dat de uitgevoerde EVC-procedure heeft geresulteerd in ervaringscertificaten die het formatieve doel dienen: zicht krijgen op de individuele ontwikkelmogelijkheden ten aanzien van de standaard. De ervaringscertificaten hebben, aldus de assessoren van de opleiding, tevens gediend als raamwerk voor het zo goed mogelijk op werknemersniveau aansluiten bij mogelijke vrijstellingen op hbo-kwalificatieniveau.

Op basis van dit voorbeeld zijn er aanbevelingen voor andere opleidingen die aan de slag willen met EVC (Duvekot, 2012):

- EVC-kandidaten moeten goed voorbereid worden op het werken met een portfolio en verzamelen van passende bewijsstukken. Het gaat niet om de hoeveelheid van de bewijslast, maar om de relevantie en kwaliteit van de bewijslast.
- Blijf de rol van assessoren nadrukkelijk toespitsen op een positieve benadering van het portfolio: de fles is immers halfvol.
- Als assessoren na afloop van het gesprek ook een advies moeten geven, dan moet daar expliciet tijd voor ingeruimd worden.
- Zorg vooraf voor duidelijkheid naar werkgevers over de kloof tussen de resultaten van de EVC-procedure en de hbo-kwalificatie om teleurstelling te voorkomen.
- Een branchecertificaat kan een toegevoegde waarde hebben omdat zo'n certificaat het functioneren op het werkniveau zichtbaar maakt en eventuele extra scholingsbehoefte toont.

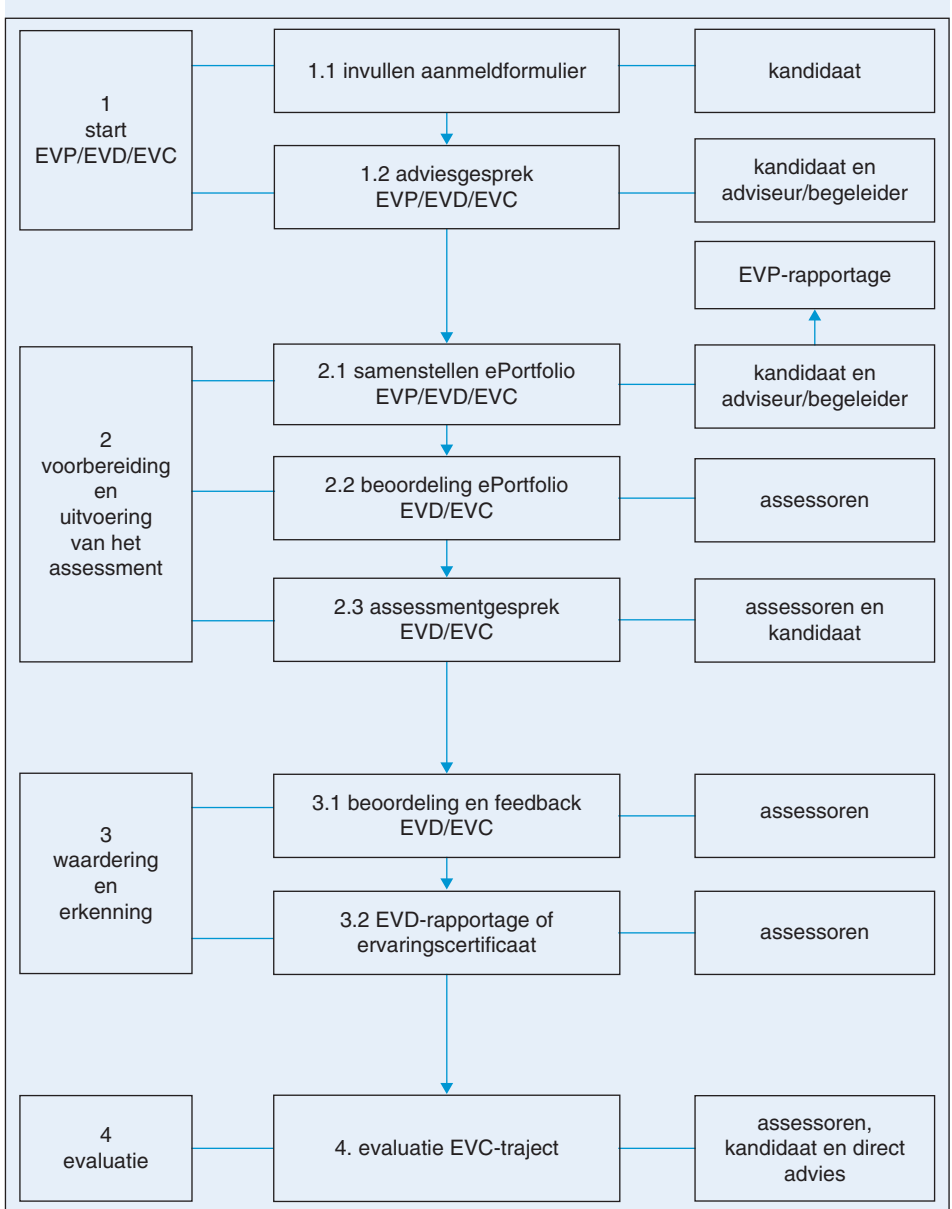
### Voorbeeld 2: EVC-procedure Servicecentrum Limburg Competent

Zuyd Hogeschool werkt samen met ROC Arcus en ROC Leeuwenborg om de kennis en vaardigheden van personen in kaart te brengen. Dit kan met behulp van EVC, maar ook met een procedure zonder dat er een assessment wordt afgenomen. Deze laatste procedure noemen zij een Ervaringsprofiel (EVP). Met betrekking tot de algemene hbo-competenties (afgeleid van de Dublin-descriptoren) is het ook mogelijk om een ervaringsdossier (EVD) op te bouwen. Hieraan is wel een assessment gekoppeld. Met dit dossier kunnen deelnemers hun algemene hbo werk- en denkniveau aantonen, wat nuttig is om in te zetten bij sollicitatieprocedures.

Bij EVC staan specifieke mbo- of hbo-opleidingen centraal. Aan de deelnemers wordt bij de aanmelding gevraagd waar behoefte aan is. In  figuur 18.2 staat de koppeling van verschillende procedures om kennis en kunde in kaart te brengen weergegeven.

De toetsmethoden en beoordelingsinstrumenten in deze EVC-procedure komen overeen met de beschrijving in ► paragraaf 18.5. Bijzondere kenmerken zijn de mogelijkheid van EVD en EVP direct aan het begin van het traject en dat er gewerkt wordt met een ePortfolio (Servicecentrum Limburg Competent, 2012).

Een ePortfolio is een digitale verzameling van allerlei tastbare leer- en werkervaringen. Het voordeel van deze digitale versie boven een papieren versie is dat personen



■ **Figuur 18.2** Een overzicht van de EVC-procedure met EVD en EVP.

de informatie mee kunnen nemen naar elke volgende leer- of werkfase. Aan de inhoud is te zien over welke competenties iemand beschikt, wat de feedback van betrokkenen bij het werk en leren is geweest en welke competenties de lerende bezig is te verwerven. Het ePortfolio heeft een ontwikkelings-, beoordelings- en presentatiefunctie. Het functioneert als cv en er staan resultaten in die zijn bereikt tijdens een opleiding of baan. Daarnaast is het een instrument dat het beoordelen en sturen van de eigen persoonlijke ontwikkeling mogelijk maakt.

De start van zo'n ePortfolio kan al in het onderwijs worden gemaakt, waarbij de docent studenten begeleidt bij het samenstellen van het portfolio. Na de opleiding zijn alle gegevens nog beschikbaar en kunnen worden aangevuld met nieuwe leerervaringen. Bij de EVC-procedure is dan alle informatie toegankelijk. De ePortfolio-eigenaar bepaalt welk deel van het portfolio wordt gebruikt voor een assessment en wie het ePortfolio mag inzien. Dit kunnen onderwijsinstellingen, arbeidsbemiddelaars en werkgevers zijn. De eerste ervaringen hiermee zijn positief. Zie voor meer informatie over deze EVC-procedure ► <http://www.limburgcompetent.nl/evc>.

## 18.7 Aandachtspunten voor EVC

In dit hoofdstuk zijn de belangrijkste eigenschappen van EVC beschreven. Het invoeren van EVC gaat niet zonder slag of stoot. Dat heeft enerzijds te maken met de verschillen tussen onderwijs en arbeidsmarkt waarin EVC een rol speelt en anderzijds met de kwaliteitseisen die worden gesteld aan EVC. Hieronder volgen van beide aspecten aandachtspunten.

### ■ Toegankelijkheid van onderwijs na EVC is niet vanzelfsprekend

Het doel van EVC is het vergroten van toegankelijkheid tot het onderwijs en ondersteuning van de ontwikkeling van individuele deelnemers. Door de hoge eisen die aan de kwaliteit van EVC-procedures gesteld worden, raakt het doel uit het vizier. Bij de start is de voorlichting over de procedure noodzakelijk om teleurstelling te voorkomen. Het doorlopen van een EVC-procedure kan immers ook leiden tot de conclusie dat iemand niet beschikt over bepaalde competenties. Het maken van leertrajecten op maat voor studenten die op basis van EVC meer of minder vrijstellingen hebben kunnen verwerven, kunnen niet altijd worden gerealiseerd. De oplossing hiervoor is ingewikkeld omdat deze te maken heeft met de wijze waarop onderwijs en curricula zijn georganiseerd. Bij het ontwerp van de opleiding moet al rekening worden gehouden met de mogelijkheid van EVC. De afstemming van de vragen van de arbeidsmarkt en het aanbod van onderwijsinstellingen staat of valt bij de mate waarin na de EVC-procedure het onderwijs in staat is haar aanbod aan te passen aan de vraag van deelnemers.

### ■ Het onderwijs accepteert de resultaten van een EVC-procedure niet (onzekerheidsreductie)

De kwaliteitsborging in het onderwijs vereist een hoge kwaliteit van de beoordeling tijdens de EVC. Als het onderwijsveld twijfelt over deze kwaliteit dan zullen zij de resultaten van een EVC-procedure niet accepteren. De oplossing hiervoor is deels te vinden in het verlenen van een accreditatiebevoegdheid aan één enkele neutrale instantie. Deze is verantwoordelijk voor het toezicht op de kwaliteitsborging van alle aanbieders van EVC-procedures buiten het erkende onderwijs. Daarbij is te denken aan het NVAO of aan de inspectie van het onderwijs. Hier hoort dan wel bij dat resultaten van EVC-procedures zonder meer worden overgenomen door onderwijsorganisaties, tenzij deze kunnen aantonen dat hun belang hiermee wordt geschaad. De aanpassing aan het EVC-stelsel om on-

derscheid te maken tussen een arbeidsmarktgerichte validering en een onderwijsgerichte validering (Stuurgroep EVC, 2013) zal hier positief aan bijdragen.

- **Aansluiting tussen resultaten van de EVC-procedure en de inhoud van een opleiding**

De aansluiting tussen de resultaten van de EVC-procedure in de vorm van adviezen voor een persoonlijk ontwikkelingsplan en de inhoud van de opleiding die daarop zou moeten volgen, zijn niet altijd goed op elkaar afgestemd. Zeker als de opleidingsdocenten niet betrokken zijn bij de EVC-procedure, bestaat het gevaar dat de toetsresultaten door hen niet zullen worden herkend of gehanteerd. Het gevolg is dat de opleiding inhoudelijk en qua programmering niet aansluit op de verwachtingen die met de EVC-procedure zijn gewekt.

- **Het samenstellen van een portfolio is lastig**

Kandidaten die wel beschikken over bepaalde competenties kunnen mogelijk minder goed in staat zijn om de bewijzen van deze competenties te leveren of te onderbouwen. Voorlichting en begeleiding zijn belangrijke voorwaarden om een te lage beoordeling als gevolg van het minder bedreven zijn in het samenstellen van een portfolio te voorkomen.

- **Intersubjectieve betrouwbaarheid**

Een EVC-beoordeling doet een groot beroep op intersubjectieve betrouwbaarheid. Deze is slechts ten dele te garanderen met behulp van procedures en soorten materiaal. Naast een goed professionaliseringstraject moet veel aandacht worden besteed aan onderhoud van adequaat beoordelaars- en toetsgedrag. In de opleiding van beoordelaars en in intervisiegesprekken waarin het eigen gedrag onderwerp van gesprek is, zal aandacht moeten worden besteed aan beoordelaarseffecten. Zie de bijlage voor een overzicht van de verschillende beoordelaarseffecten.

## 18.8 Ten slotte

---

Ondanks de verschillende speelvelden van onderwijs en arbeid kan EVC onder voorwaarden wel fungeren als de bezegeling van een goed werkende en voor beide partijen optimale latrelatie.

Voor de EVC-procedures gelden dezelfde kwaliteitseisen als voor de reguliere toetsprocedures in het hbo. Dit is logisch omdat de doelen van EVC in hoge mate sporen met de huidige ontwikkelingen in het beroepsonderwijs: flexibiliteit, individueel maatwerk, opleiden in authentieke beroepssituaties en de arbeidssector als medebeoordelaar. De grootste hobbel bij de ontwikkeling van EVC-procedures is veelal niet het beoordelingstechnische deel van de ontwikkeling. De perceptie van collega's bij toetsen, de connotaties ten aanzien van betrouwbaarheid en zekerheid die worden gezocht in gedetailleerde toetsindicatoren en criteria, vormen vaak een veel groter probleem.

Toetsen van verworven competenties – zo blijkt dan eens te meer – is niet een kwestie van optellen en aftrekken, maar een kwestie van veranderingsbereidheid, veranderingsbekwaamheid, omgaan met onzekerheid en zelfvertrouwen.

# Toetsen van competenties

*Judith Gulikers en Niek van Benthum*

## Inleiding

Wat is competentietoetsing?

Waarom zou ik het moeten toepassen?

Waar komt het vandaan?

Hoe moet ik het toepassen?

Wat zijn de valkuilen?

- 19.1**     **Wat is competentietoetsing? – 218**
- 19.2**     **De meerwaarde van competentietoetsing – 219**
- 19.3**     **Competentietoetsing vraagt ‘andersom denken’ – 219**
- 19.4**     **Het wat en hoe van competentietoetsing – 221**
- 19.5**     **Wat wordt getoetst? – 221**
- 19.6**     **Hoe wordt getoetst? – 223**
- 19.6.1     Integratie van kennis, houding en vaardigheden in professioneel gedrag – 223
- 19.6.2     Authentiek toetsen: competentietoetsing voor professioneel gedrag? – 224
- 19.6.3     Meer dan gedragstoetsing: toetsen van onderliggende processen – 226
- 19.6.4     Toetsing van ontwikkelvermogen – 226
- 19.6.5     Rubrics: handig hulpmiddel bij competentietoetsing – 227
- 19.7**     **Ten slotte – 227**

» Het succes van competentiegericht onderwijs zal in het hoger beroepsonderwijs voor een belangrijk deel afhankelijk zijn van de mate waarin onderwijsinstellingen erin slagen de toetsing ook competentiegericht te maken. Competentiegerichte toetsing impliceert een integrale toetsing van kennis, vaardigheden en houdingen.' (Van Merriënboer, Van der Klink & Hendriks, 2002, p. 29) «

Competentietoetsing heeft naast fervente voorstanders, ook hartgrondige tegenstanders. Tot de laatste groep behoren studenten die 'reflectiemoe' zijn en docenten die na verschillende pogingen competentietoetsing toe te passen, pleiten voor een teruggang naar kennistoetsing in het hoger onderwijs (Van Berkel, 2012). Uit een analyse van negen hbo competentiegerichte toetsprogramma's blijkt dat alle opleidingen weliswaar competenties hebben benoemd, maar dat het voor alle partijen vaak onduidelijk is hoe die nu concreet terugkomen in de toetsing (Baartman, Gulikers & Dijkstra, 2013). In het wetenschappelijk onderwijs komt competentietoetsing nog steeds maar mondjesmaat van de grond, vooral vanwege de associatie dat competentietoetsing ten koste gaat van academische kennisopbouw wat het wetenschappelijk onderwijs toch als de kern van haar taak ziet (Mulder, Gulikers, Biemans & Wesselink, 2009). Hieruit blijkt dat het nog niet voldoende duidelijk is wat competentietoetsing is, hoe dit zich verhoudt tot kennistoetsing en hoe competentietoetsing het beste kan worden uitgevoerd.

## 19.1 Wat is competentietoetsing?

Het hoger onderwijs bereidt studenten voor op een dynamische en veranderlijke arbeidsmarkt en leidt hen op voor een meer of minder concreet beroep (Sluijsmans, 2013), waarin zij moeten beschikken over een adequate set van kennis, vaardigheden en houdingen. Competente beroepsbeoefenaren zijn in staat om deze losse stukjes kennis, vaardigheden en houdingen te integreren (tot competenties), resulterend in adequaat professioneel gedrag in een variatie aan beroepsrelevante situaties (Van Berkel, 2012). Competenties krijgen dus pas betekenis in een concrete professionele context (Mulder, in druk), wat meteen laat zien dat competentietoetsing altijd nauw verbonden moet zijn met een beroepsrelevante situatie.

Informatie over wat bekwame afstudeerders moeten kennen, kunnen en kunnen laten zien, is terug te vinden in competentieprofielen die meestal afgeleid zijn van kwalificatieprofielen, opleidingsprofielen en/of Dublin-descriptoren. Als aanvulling hierop wordt door steeds meer onderzoekers een extra noodzakelijke eigenschap van iedere competente afgestudeerde benoemd: *ontwikkelvermogen* (Kloppenburger, 2011). Hierbij gaat het om het vermogen van iedere beroepsbeoefenaar om de eigen competenties voortdurend en doelbewust te blijven ontwikkelen en aan te passen aan de veranderende praktijk. Een competentie is immers ontwikkelbaar en is nooit 'af' (Van Merriënboer e.a., 2002).

Competentietoetsing geeft zicht op het niveau en/of de ontwikkeling van de competenties van studenten, ofwel op de mate waarin een student:

- kennis, vaardigheden en houdingen kan integreren in adequaat professioneel handelen in een beroepsrelevante situatie;



- dit professioneel gedrag flexibel kan aanpassen in verschillende beroepssituaties, ofwel *bewuste* beheersing van de competenties vertoont;
- zicht heeft op de 'mate van ontwikkeling' van een competentie en in staat is deze doelgericht verder te ontwikkelen. Met andere woorden, of er sprake is van *duurzame* competentiebeheersing.

## 19.2 De meerwaarde van competentietoetsing

---

Competentietoetsing is ontstaan om het gat tussen onderwijs en arbeidsmarkt te verkleinen. Een competentietoets brengt studenten in situaties die representatief zijn voor hun latere beroep en beoordeelt of een student in die situaties in staat is effectief te handelen. Doordat de toetsituatie een afspiegeling is van een echte beroepssituatie, verhoogt competentietoetsing de inhouds- en voorspellende validiteit van de toetsing ten opzichte van de toekomstige beroepspraktijk (Messick, 1994). Dit houdt dus in dat competentietoetsing in iedere opleiding aan bod zou moeten komen om een gefundeerde uitspraak te kunnen doen over de mate waarin de student daadwerkelijk in staat is goed te functioneren in een beroep. Hiervoor kan een opleiding zich niet baseren op alleen kennis- en vaardigheidstoetsen.

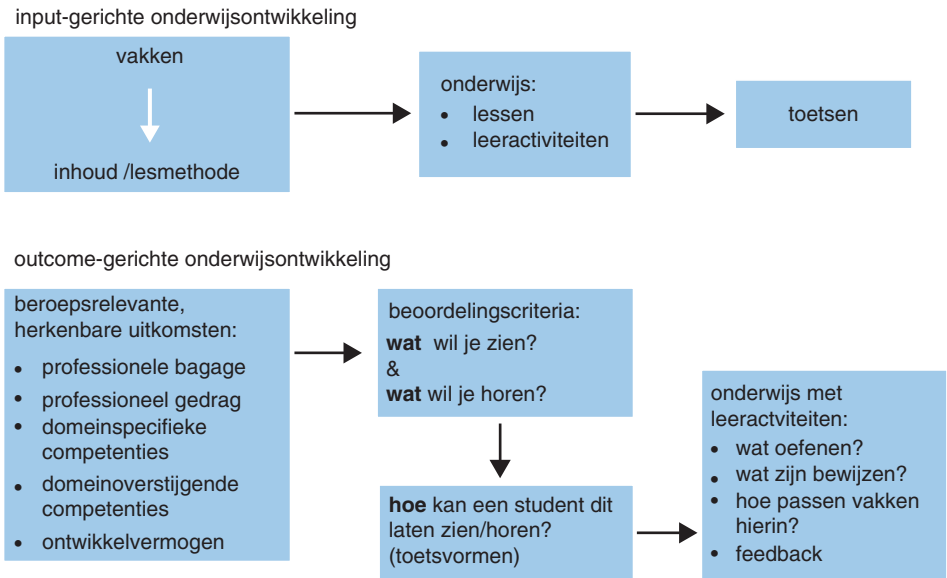
Door studenten te confronteren met kritische beroepssituaties waarin ze professioneel moeten handelen, geven competentietoetsen studenten bovendien een duidelijk beeld van het soort taken waarmee zij in hun toekomst zullen worden geconfronteerd. Via deze weg kunnen competentietoetsen een belangrijke bijdrage leveren aan de studiemotivatie (waar doe ik het allemaal voor?) maar bieden ze studenten vooral ook mogelijkheden om een beroepsbeeld en een beroepsidentiteit te creëren, wat erg belangrijk is voor de inzetbaarheid van afgestudeerden op de arbeidsmarkt (Holmes, 2013). Vooral in het wetenschappelijk onderwijs is deze beroepsvoorbereiding en beroepsbeeldvorming vaak nog een probleem.

Een minstens zo belangrijke motivatie voor competentietoetsing komt voort uit het inmiddels breed gedragen principe dat toetsing de sterkst sturende factor is achter het leren van studenten. Competentietoetsing is sterk gericht op het positief stimuleren en waarderen van relevante leerprocessen die iedere beroepsbeoefenaar, in een maatschappij van levenlang leren, zich zou moeten hebben eigen gemaakt (Boud & Falchikov, 2007). Het doel van competentietoetsing is hiermee niet langer alleen maar het beoordelen van het eindresultaat en het geven van cijfers, maar vooral ook het stimuleren en waarderen van de gewenste leerprocessen en de ontwikkeling richting gewenste leeruitkomsten (Boud & Falchikov, 2007; Wiliam, 2011).

## 19.3 Competentietoetsing vraagt 'andersom denken'

---

Uiteindelijk wil iedere opleiding competente beroepsbeoefenaren afleveren die klaar zijn voor de huidige en toekomstige arbeidsmarkt. Dat betekent dat het noodzakelijk is om te onderzoeken wat de arbeidsmarkt vraagt van afgestudeerden. Sluijsmans (2013) benoemt niet voor niets de 'analyse van de inhoud van de professie' als eerste belangrijke bouwsteen



■ **Figuur 19.1** Andersom denken: van input naar uitkomstgerichte onderwijs- en toetsontwikkeling.

voor professioneel toetsen. Dit veronderstelt een uitkomstgerichte manier van onderwijsontwikkeling (outcome-based education, Young, 2009):

» De inrichting van het onderwijs begint bij de (beroepsrelevante) uitkomsten die je uiteindelijk wilt bereiken. «

Dat wil zeggen bij de *wat*-vraag. Deze uitkomsten bepalen vervolgens als eerste de inhoud van het toetsprogramma om van daaruit terug te redeneren naar het onderwijs en gepaste lesmethoden, bronnen en activiteiten. Dit staat haaks op inputgerichte onderwijsontwikkeling waarbij vakken, bijbehorend studiemateriaal en methoden de basis vormen van waaruit lesinhouden en onderwijsactiviteiten worden benoemd en waarbij de toets pas aan het einde wordt ontwikkeld (Barreveld, 2012). Toetsing is in dit geval meestal een afspiegeling van de informatie die in een boek staat of op college is verteld. Deze toetsing is dan misschien wel valide ten opzichte van het onderwijs, maar niet ten opzichte van de beroepspraktijk wanneer dit onderwijs geen goede afspiegeling is van de beroepspraktijk (Messick, 1994). Voor competentietoetsing die professioneel handelen in kaart wil brengen, is een outputgerichte aanpak nodig. ■ Figuur 19.1 brengt dit ‘andersom denken’ in beeld: niet input-gerichte maar uitkomstgerichte onderwijsontwikkeling die start bij de *wat*-vraag: ‘Wat zijn de beroepsrelevante uitkomsten van onze opleiding?’

## 19.4 Het wat en hoe van competentietoetsing

Nadat de beroepsrelevante uitkomsten benoemd zijn, is de eerste stap richting de ontwikkeling van competentietoetsing het concreet en toetsbaar maken van deze uitkomsten binnen een bepaalde opleiding (zie ► hoofdstuk 5 en Van Berkel, 2012). De vaak generieke beroepsrelevante uitkomsten, zoals ‘probleemoplossend vermogen hebben’ moeten concreet worden vertaald in de context van een bepaalde opleiding. Hierbij vormen twee vragen de kern:

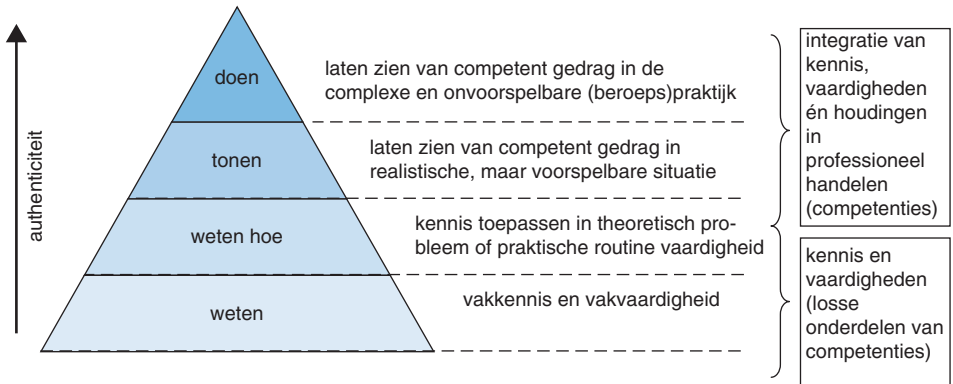
- *Wat* willen we toetsen? Of meer concreet:
  - Wat willen docenten van studenten *zien*?
  - Wat willen docenten van studenten *horen/lezen* om de benoemde beroepsrelevante uitkomsten te kunnen beoordelen?
- *Hoe* zouden wij onze studenten dan moeten beoordelen om met vertrouwen te kunnen zeggen dat ze competent kunnen handelen?

Het is cruciaal om eerst de *wat*-vraag te beantwoorden, voordat de *hoe*-vraag wordt gesteld. De *hoe*-vraag gaat namelijk over welke toetsvorm, of welke combinatie van toetsvormen (zie ook ► hoofdstuk 5) geschikt is om de doelen te beoordelen. Het kiezen van de juiste toetsvorm (*hoe*) op basis van aansluiting bij het doel (*wat*) wordt ook wel ‘fitness-for-purpose’ genoemd en is de kern van toetskwaliteit (Van der Vleuten, Schuwirth, Scheele, Driessen & Hodges, 2010). Het eerst stellen van de *hoe*-vraag zonder de *wat*-vraag te hebben beantwoord, leidt tot een instrumentele en weinig bewuste toetspraktijk. Dit komt echter wel veel voor in het onderwijs.

## 19.5 Wat wordt getoetst?

De piramide van Miller helpt bij het beantwoorden van de *wat*-vragen en het concreet maken van generieke beroepsrelevante uitkomsten voor studenten in een bepaalde opleiding (bijvoorbeeld: wat betekent ‘probleemoplossend vermogen’ voor studenten van onze opleiding diergeneeskunde?). Miller (1990) beschrijft beroepsbekwaamheid in vier lagen: 1) weten, 2) weten hoe, 3) tonen en 4) doen (► figuur 19.2; zie ook Baartman & Gulikers, in druk). De onderste twee lagen richten zich op losse kennis en/of vaardigheden en cognitieve/denkprocessen, terwijl de bovenste twee lagen daadwerkelijk professioneel gedrag aanspreken.

Voor het beoordelen van de bovenste twee lagen zijn competentietoetsen noodzakelijk en zullen toetsen een directe relatie met de beroepspraktijk moeten hebben, dat wil zeggen meer authentiek moeten zijn. De onderste twee lagen kunnen daarentegen met meer traditionele, schriftelijke toetsen worden getoetst. In een toetsprogramma van een opleiding én het daarop voorbereidende onderwijsprogramma moeten al deze lagen een plek hebben (zie ook ► hoofdstuk 5). Dit betekent aan de ene kant dat een opleiding niet kan volstaan met alleen toetsen die de ‘weten’- en ‘weten hoe’-lagen aanspreken, maar aan de andere kant ook dat *niet alle* toetsen in het hoger onderwijs competentietoetsen hoeven te zijn. Alleen *die* onderdelen die in de bovenste lagen van de piramide van Miller zijn gesitueerd



■ **Figuur 19.2** Gebaseerd op piramide van Miller (1990).

en die te maken hebben met bewust gedrag en de integratie van kennis, vaardigheden en houdingen, *die* onderdelen vragen om competentietoetsing. Bijvoorbeeld, een opleiding diergeneeskunde wil ten behoeve van de competentie 'probleemoplossend vermogen' van iedere student horen (of lezen) dat deze een relatie weet te leggen tussen ziektesymptomen en mogelijke onderliggende ziekten. Dit is natuurlijk cruciaal voor een competente dierenarts maar valt onder de bekwaamheidslaag 'weten hoe' van de piramide van Miller, waarvoor bijvoorbeeld een casustoets (en dus geen competentietoets) een geschikte toetsing is.

Dit bovenstaande suggereert wellicht dat een competentietoets gelijk is aan een dragstoets, een toets waarin de student professioneel gedrag laat zien. De twee wat-vragen in ■ figuur 19.1 geven echter al aan dat competentietoetsing niet volstaat met de vraag: 'Wat wil ik van studenten *zien*?' Om te beoordelen of een student *bewust* en *duurzaam* competent is, vraagt competentietoetsing ook om het beoordelen van onderliggende processen die leidend zijn tot bepaald gedrag. Deze zijn lastig te observeren. Om expliciet te krijgen wat er van onderliggende processen moet worden beoordeeld, beantwoorden docenten de tweede vraag: 'Wat willen we van studenten *horen/lezen*?' Processen waar een competentietoets inzicht in wil geven zijn: integratie (Baartman & De Bruijn, 2011), transfer (Salomon & Perkins, 1989) en zelfregulatie. Voor al deze processen is reflectie noodzakelijk (Baartman & De Bruijn, 2011; Schön, 1987; Van Berkel, 2012). Deze reflectie is te structureren aan de hand van drie concrete vragen:

1. Hoe heb ik een bepaalde taak uitgevoerd (knowing how)?
2. Waarom heb ik een bepaalde taak op een bepaalde manier uitgevoerd (knowing why)?
3. Waartoe heb ik een bepaalde taak op een bepaalde manier uitgevoerd (knowing who I am as professional/person)?

Bewuste beheersing van competentie waarbij studenten adequaat professioneel gedrag in een variatie aan beroepssituaties kunnen laten zien (transfer), vraagt in ieder geval een reflectie op vraag 1 en 2. Echter, duurzame beheersing van de competentie, ofwel *ontwikkelvermogen*, noodzaakt reflectie via vraag 3: dit laat zien dat studenten in staat zijn om een competentie in bepaalde richting te ontwikkelen omdat daar hun eigen leerbehoeften of

talenten liggen of omdat zij hiermee een bijdrage kunnen leveren aan de ontwikkeling van het beroep. Het kunnen reflecteren via vraag 3 is wat van een afgestudeerde in het hoger onderwijs wordt verwacht (Van Berkel, 2012).

Dus: *wat* wil een docent in competentietoetsen beoordelen? *Wat* wil een docent *zien* en *horen*?

- adequate uitvoering van professionele taken waarin een student kennis, vaardigheden en houdingen integreert (tonen/doen niveau van Miller);
- onderliggende processen die leiden tot dit vertoonde gedrag (bewust competent);
- of de student zicht heeft op eigen ontwikkeling van een competentie en in staat is deze doelgericht verder te ontwikkelen (ontwikkelvermogen, duurzaam competent).

## 19.6 Hoe wordt getoetst?

---

In principe kan competentietoetsing met allerlei toetsvormen plaatsvinden. Voorwaarde is echter dat ze passen bij de beoordelingsdoelen en beoordelingscriteria (zie ► hoofdstuk 5). Performance assessments, self-assessments, peer-assessment, reflectieverslagen, portfolio's kunnen allemaal passen binnen competentietoetsing, maar mogen geen doel op zich zijn. Ze moeten worden ingezet om zicht te geven op een of meer van de drie onderdelen van competentietoetsing: integratie in professioneel gedrag, onderliggende processen en/of ontwikkelvermogen. Hieronder wordt per aspect van competentietoetsing (het *wat*) concrete tips gegeven over de vormgeving van competentietoetsen (het *hoe*).

### 19.6.1 Integratie van kennis, houding en vaardigheden in professioneel gedrag

---

Een competentietoets dient dus in ieder geval een directe beoordeling van gedrag te bevatten, zoals een directe observatie of observatie via video. Bij competentietoetsing moeten studenten in dit gedrag *integratie* (van kennis, vaardigheden en houdingen) laten zien. Dit maakt competentietoetsing méér dan een vaardigheidstoets (zie ook ► hoofdstuk 16). Noodzakelijk hiervoor is ten eerste dat de criteria zodanig zijn geformuleerd dat ze daadwerkelijk deze integratie beoordelen, en niet als afzonderlijke kennis, vaardigheden en houdingsaspecten op een checklist worden afgevinkt. Ten tweede zijn complexe toetstaken belangrijk. Dit zijn toetstaken zonder vaststaande oplossing waarin zich onverwachte situaties kunnen voordoen. Integratie wordt het meest gestimuleerd wanneer studenten actief op zoek moeten naar oplossingen en de daarvoor benodigde kennis en vaardigheden (Baartman & De Bruijn, 2011). De toetsing is dus een nieuwe uitdagende situatie passend bij de beoordelingscriteria. Een belangrijke voorwaarde hiervoor is wel dat studenten in het onderwijs voldoende hebben geoefend in het omgaan met complexe taken waarin ook integratie van kennis, vaardigheden en houdingen wordt aangesproken en waar bij voorkeur ook feedback op wordt gegeven. Dat wil zeggen, er moet 'constructive alignment' zijn (Biggs, 1996) tussen onderwijstaken en toetstaken.

## 19.6.2 Authentiek toetsen: competentietoetsing voor professioneel gedrag?

Competentietoetsing beoordeelt professioneel handelen in een beroepsrelevante situatie. Het is noodzakelijk dat de toetsing zo veel mogelijk overeenkomt met situaties uit de beroepspraktijk: authentiek toetsen (Gulikers, Bastiaens & Kirschner, 2004; Gulikers, 2005). Voor het vormgeven van een authentieke toets is een kritische beroepssituatie het startpunt. De relevante vraag is:

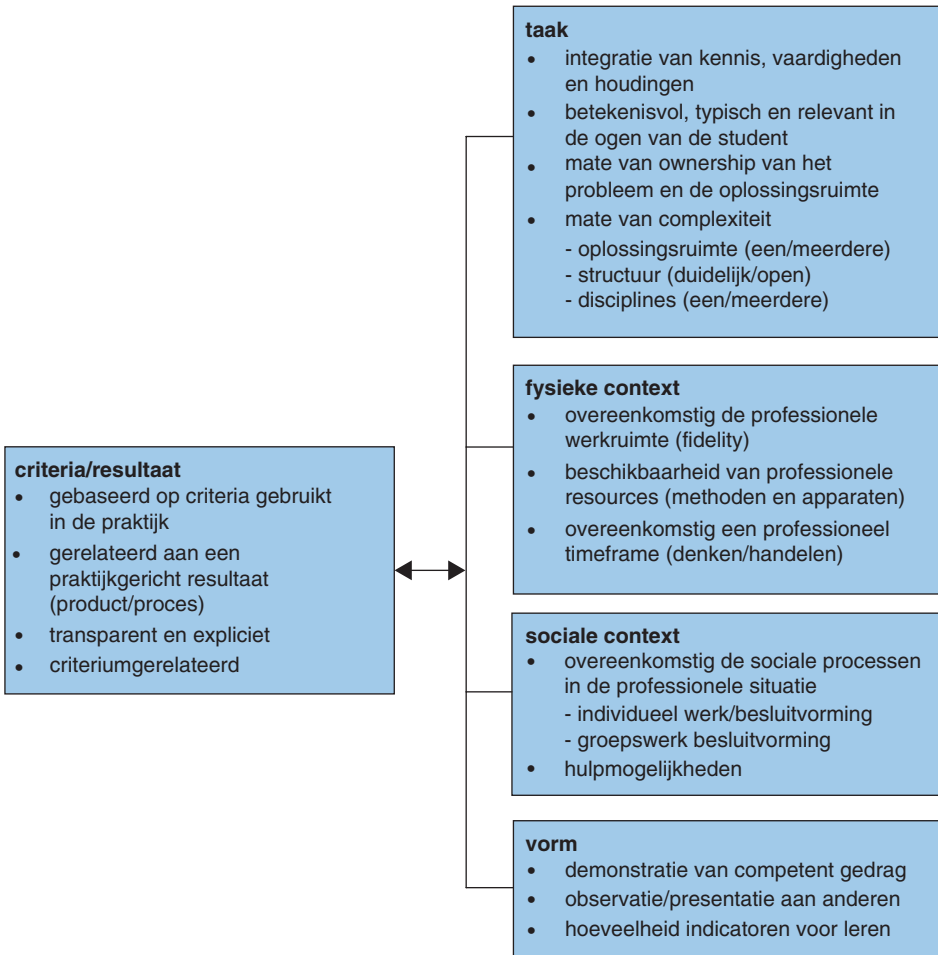
➤ **‘Met welke kritische én representatieve beroepssituaties krijgt het merendeel van onze studenten in de (toekomstige) beroepspraktijk te maken?’**

Let op, van eerstejaars studenten kan niet worden verwacht dat zij taken van een afgestudeerde (startend) beroepsbeoefenaar kunnen uitvoeren. Daarom moeten kritische beroepssituaties worden geïdentificeerd die passen bij de fase van een opleiding. Denk hierbij bijvoorbeeld aan situaties die studenten tijdens een stage moeten uitvoeren of professionele taken die eerstejaars al wel uitvoeren. Zo zal een eerstejaars student ‘Landschapsarchitectuur’ nog geen ontwerp voor een stadspark hoeven te kunnen maken, maar moet deze wel een goed onderbouwd ontwerp voor een speelterreintje in Amsterdam Zuid kunnen maken. Zo’n geïdentificeerde beroepssituatie op het niveau van de student heet *criteriumsituatie* en is het startpunt voor het ontwerpen van een authentieke competentietoets. Denken vanuit beroepsrelevante situaties op het niveau van de student, impliceert ook dat authentiek toetsen mogelijk is gedurende een hele opleiding en niet alleen aan het einde.

Voor het ontwerpen van authentieke toets biedt het 5-dimensionale model (5DM) voor authentieke toetsen handvatten (■ figuur 19.3). Belangrijk uitgangspunt achter dit model is dat iedere toets meer of minder authentiek kan worden gemaakt. Er zijn namelijk vijf toetskenmerken (dimensies) die toetsontwerpers meer of minder op die criteriumsituatie kunnen laten lijken. Hierbij gaat het om:

- Taak: de toetsopdracht(en) die de student moet uitvoeren tijdens de toets.
- Omgeving: de fysieke omgeving waarin de student de toetsopdracht moet uitvoeren.
- Sociale context: de interactie(on)mogelijkheden in de toetsituatie.
- Vorm: de manier van toetsen; de toetsmethode, onafhankelijk van de inhoud.
- Resultaat/Criteria: de output die uiteindelijk wordt beoordeeld en de punten waarop deze prestatie beoordeeld wordt (wat wil ik zien/horen?).

Vanuit de criteriumsituatie worden eerst authentieke beoordelingscriteria en resultaten benoemd: ‘Waar draait het om in de beroepspraktijk?’. Van daaruit kan een docent met behulp van dit model een betekenisvolle en voldoende complexe toetstaak beschrijven die past bij de fase van de opleiding waarin de student zich bevindt. Daarnaast kan een docent benoemen waar de toets moet plaatsvinden, welke mogelijkheden de context moet bieden (bijvoorbeeld welk materiaal moet aanwezig zijn?) en een passende toetsvorm kiezen (bijvoorbeeld observatie van gedrag; presentatie voor opdrachtgevers).



■ **Figuur 19.3** Het vijf-dimensionale model voor authentiek toetsen.

Deze vijf dimensies laten ook zien dat een authentieke toets méér is dan simpelweg toetsen in de beroepspraktijk. Een toets in de beroepspraktijk (fysieke context) waarbij de student verpleegkunde alleen patiëntstatussen mag kopiëren (toetstaak) is verre van authentiek, terwijl een goed opgezette simulatie in een schoolsetting zeer authentiek kan zijn. Onderzoek heeft zelfs laten zien dat authentieke criteria en een authentieke taak de belangrijkste kenmerken zijn van goed authentiek toetsen (Gulikers, 2006). Dit biedt ruime mogelijkheden voor authentieke competentietoetsing in het hoger onderwijs waarin studenten minder in de daadwerkelijke beroepspraktijk (bijvoorbeeld via stages) komen.

### 19.6.3 Meer dan gedragstoetsing: toetsen van onderliggende processen

---

Bewust competent zijn betekent dat een student niet toevallig goed gedrag vertoont in een specifieke toetssituatie, maar bewust kiest voor het vertonen van bepaald gedrag passend binnen een beroepssituatie. Dit impliceert dat een competentietoets uit meerdere onderdelen, of combinaties van toetsvormen, bestaat. Het mede beoordelen van onderliggende processen leidend tot bepaald gedrag vraagt aandacht op twee punten. Ten eerste, het beoordelen van onderliggende processen moet *niet* in isolatie gebeuren, maar moet direct gekoppeld zijn aan voorafgaande of daaropvolgende gedragstoetsing. Dus, een ‘losse’ reflectietoets waarbij niet op het vertoonde gedrag wordt gelet, is geen goede competentietoetsing. Een gesprek waarin wordt beoordeeld of de student in een door de beoordelaars geobserveerd performance assessment bewust bekwaam heeft gehandeld, kan wel een goede competentietoetsing zijn. Ten tweede, beoordelingscriteria moeten helder maken welke onderliggende processen worden getoetst en hoe deze worden beoordeeld. Dit moet rechtstreeks voortvloeien uit de vraag: ‘Wat wil je als docent van een student horen/lezen?’. Wanneer dit niet expliciet is, wordt een criteriumgericht interview al snel een overhoring van kennis of wordt een reflectieverslag een verhaal dat voornamelijk wordt beoordeeld op een criterium als leesbaarheid. Een hulpmiddel dat docenten kunnen gebruiken om de beoordelingscriteria voor onderliggende processen helder te krijgen is het stellen van de vraag: ‘Wanneer is een student gezakt voor deze toets?’. Als docenten die vraag eigenlijk niet kunnen beantwoorden, of alleen maar kunnen beantwoorden met: ‘Als hij niets zegt/niets opschrijft’, dan hebben de docenten niet scherp *wat ze* eigenlijk willen beoordelen. Om scherp te krijgen wat docenten willen horen/lezen, kunnen zij de volgende drie vragen stellen:

- Hoe is een taak uitgevoerd?
- Waarom is deze taak op deze manier uitgevoerd?
- Met welke reden/met welk doel is deze taak op deze manier uitgevoerd?

Ten derde, het is belangrijk dat het niet alleen voor beoordelaars, maar vooral ook voor studenten bekend is waarop ze worden beoordeeld. Dit geldt zowel ten aanzien van hun gedrag als de onderliggende processen. Dit wordt ook wel transparantie van toetsing genoemd en is een cruciaal kwaliteitscriterium bij accreditaties van opleidingen.

### 19.6.4 Toetsing van ontwikkelvermogen

---

Het doel van competentietoetsing is niet alleen het toetsen van een eindniveau en eindproduct, maar ook om relevante leerprocessen en ontwikkeling van competenties te stimuleren. Formatieve toetsing moet daarom een noodzakelijk onderdeel van competentietoetsing zijn. Ook hierbij is een cruciaal aandachtspunt: een formatieve toets (zie ook ► hoofdstuk 8) heeft pas zin als deze direct is gekoppeld – ‘aligned’ (Biggs, 1996) – aan de



uiteindelijke summatieve toetsing; beide moeten gericht zijn op dezelfde beoordelingscriteria en inzicht geven in: ‘Wat je als docent van de student wilt zien of horen?’

Bij formatieve toetsing wordt de prestatie afgezet tegen de beoordelingscriteria. Dit geeft informatie over: ‘Waar sta ik nu?’ (Wiliam, 2011). Vervolgens moeten concrete acties worden benoemd om de discrepantie tussen waar de student naartoe wil/moet (in de summatieve beoordeling) en waar de student nu staat te dichten (Wiliam, 2011). Hier ligt een belangrijke rol weggelegd voor reflectie en feedback (Hattie & Timperley, 2007). Inherent hieraan is dat een student in een volgend toetsmoment de gelegenheid moet krijgen om te laten zien dat concrete stappen zijn gezet en effect hebben gehad. Met andere woorden, competentietoetsing vraagt om doelbewust samenhangende formatieve en summatieve toetsing, waarbij meerdere toetsmomenten elkaar opvolgen en waarbij zowel proces als eindresultaat onderdeel van de toetsing zijn.

### 19.6.5 Rubrics: handig hulpmiddel bij competentietoetsing

---

Een handig hulpmiddel bij competentietoetsing zijn rubrics (zie ► hoofdstuk 17). Rubrics slaan namelijk veel vliegen in één klap: ze ‘dwingen’ om generieke of vage beroepsrelevante uitkomsten/competenties concreet en transparant te maken en om de beheersing hiervan op verschillende niveaus te beschrijven. Hiermee helpen rubrics bij het objectiever beoordelen van gedrag en onderliggende processen, bieden ze mogelijkheden om het huidige ontwikkelniveau in kaart te brengen én bieden ze mogelijkheden om verdere ontwikkelmogelijkheden te benoemen en daadwerkelijke ontwikkeling te duiden en te waarderen. Rubrics zijn zeer geschikt voor samenhangende formatieve en summatieve competentietoetsing, het observeren van gedrag door verschillende beoordelaars en in verschillende beroepssituaties, zelf-beoordeling, peer-beoordeling en het geven van concrete feedback en feed forward (Hattie & Timperley, 2007).

## 19.7 Ten slotte

---

■ Tabel 19.1 vat dit hoofdstuk samen en maakt inzichtelijk (1) wat een docent wil toetsen in competentietoetsing (2) hoe de toets er dan uit moet zien (3) wat hierbij valkuilen zijn.

■ **Tabel 19.1** Samenvatting van kaders voor competentietoetsing en valkuilen.

Wat toetst competentietoetsing?	Hoe moet een competentietoets er dan uit zien?	Wat is hierbij een valkuil voor docenten?
<p>integratie van kennis, vaardigheden en houdingen in adequaat professioneel gedrag</p>	<ul style="list-style-type: none"> <li>– beoordelen concreet gedrag (tonen/ doenlaag van Miller)</li> <li>– een competentietoets is méér dan een vaardigheidstoets</li> <li>– beoordelingscriteria die integratie toetsen</li> <li>– complexe toetstaken met onverwachte en onbekende beroepsdilemma's die integratie uitdagen</li> <li>– authentiek: zo veel mogelijk overeenkomstig met situaties uit de beroepspraktijk (zie ■ figuur 19.3)</li> <li>– kritische beroepssituatie op het niveau van de student (passend bij de fase van de opleiding) als startpunt voor de toets</li> </ul>	<ul style="list-style-type: none"> <li>– geen gedragstoets, bijvoorbeeld gedrag toetsen via (alleen) reflectieverslagen</li> <li>– beoordelingscriteria die geen integratie noodzakelijk maken</li> <li>– weinig uitdagende beoordelings-situatie waarbij studenten kunnen vertrouwen op routine ('de toets is de slagroom op de taart')</li> <li>– toetsen in de beroepspraktijk zonder authentieke taak en criteria</li> <li>– kritische beroepssituaties die niet passen bij het niveau/de fase van de opleiding van de student</li> <li>– geen alignment tussen onderwijs en toetsing</li> </ul>
<p>bewust competent: onderliggende processen leidend tot professioneel gedrag</p>	<ul style="list-style-type: none"> <li>– combinaties van toetsvormen: gedragtoets gecombineerd met een toets die inzicht geeft in onderliggende processen van 'knowing how/ why/who I am'</li> <li>– vooraf expliciet maken in beoordelingscriteria wat je als docent concreet wilt horen/lezen t.a.v. 'knowing how/why/who I am'</li> <li>– beoordelingscriteria vooraf transparant maken aan alle betrokkenen, inclusief studenten</li> </ul>	<ul style="list-style-type: none"> <li>– competenties toetsen met één toets op één moment</li> <li>– onderliggende processen los van een gedragsobservatie toetsen</li> <li>– als beoordelaar niet goed weten wat je als docent eigenlijk wilt horen/lezen t.b.v. 'knowing how/ why/who I am'</li> <li>– beoordelingscriteria niet vooraf bekend maken bij alle betrokkenen ('geheimhouden' voor studenten)</li> </ul>
<p>ontwikkelvermogen</p>	<ul style="list-style-type: none"> <li>– doelbewuste samenhang tussen formatieve en summatieve toets ('alignment')</li> <li>– longitudinaal toetsen: kansen bieden om ontwikkeling te laten zien + 'iets' te kunnen doen met feedback</li> </ul>	<ul style="list-style-type: none"> <li>– formatief toetsen als doel op zich/ geen helder verband tussen formatieve en summatieve toetsing</li> <li>– geen nieuw toetsmoment aanbieden waarin student kan aantonen te hebben geleerd</li> </ul>
<p>aanvullende tips</p>	<ul style="list-style-type: none"> <li>– hoe de toets eruit moet zien wordt bepaald door <i>wat</i> je als docent wilt beoordelen</li> <li>– kiezen van een geschikte toetsvorm passend bij <i>datgene</i> wat je als docent wilt beoordelen (fitness for purpose)</li> <li>– gebruikmaken van rubrics</li> </ul>	<ul style="list-style-type: none"> <li>– inzetten van een 'nieuwe' toetsvorm als doel op zich</li> </ul>

# Toetsen van werkpleklers

*Marcel van der Klink en Jo Boon*

## Inleiding

Hoe kun je leren op de werkplek toetsen?

Wat is het verschil tussen een stage en werkpleklers?

Welke eisen stelt het toetsen van werkpleklers aan de werkplek?

Hoe verhouden begeleiding en beoordeling zich bij het toetsen op de werkplek?

Welke beoordelingsinstrumenten zijn geschikt?

In hoeverre sluit werkplektoetsing aan bij de professionele ontwikkeling van de student?

## 20.1 Vormen van leren op de werkplek – 230

20.1.1 Stages – 230

20.1.2 Duale opleiding – 230

20.1.3 Leerafdelingen – 231

20.1.4 Communities – 231

## 20.2 Dilemma's en problemen in werkpleklers – 231

20.2.1 Opleiding besteedt leren uit maar blijft eindverantwoordelijk – 232

20.2.2 Garanderen van kwaliteit en werkplek – 232

20.2.3 Begeleiding van studenten – 232

20.2.4 Werkpleklers stelt eisen aan de opleiding – 233

20.2.5 Het plan als onderlegger voor activiteiten en beoordeling – 233

20.2.6 Werkpleklers is dynamisch – 234

## 20.3 Instrumenten voor het beoordelen van werkpleklers – 234

20.3.1 Portfolio – 234

20.3.2 360 graden feedback – 235

20.3.3 POP – 236

20.3.4 Ontwikkelingsgesprekken – 236

20.3.5 Performance assessment – 237

20.3.6 Stageverslag – 237

## 20.4 Ten slotte – 238

In het hoger onderwijs is werkplekleren, het leren in de beroepspraktijk, een belangrijk onderdeel van het curriculum (Streumer, 2010). De afgelopen jaren zijn er naast stages andere vormen van werkplekleren ontstaan die niet als een stage zijn te bestempelen, vandaar dat hier de term werkplekleren wordt geprefereerd. In dit hoofdstuk wordt ingegaan op de vraag wat dan de vormen zijn waarin het werkplekleren zich in het hoger onderwijs manifesteert. In het hoger onderwijs lijkt een voorkeur te bestaan voor het toetsen van werkplekleren door middel van een stageverslag. Hoewel zo'n verslag waardevol kan zijn, bestaan er nog tal van andere instrumenten die relevant en soms zelfs beter geschikt zijn om inzicht te verkrijgen in het leerproces en de leerresultaten van de studenten.

## 20.1 Vormen van leren op de werkplek

---

### 20.1.1 Stages

---

De stage is de bekendste vorm waarin het leren op de werkplek in het hoger onderwijs plaatsvindt. Daarbij kan een onderscheid worden gemaakt in een tweetal varianten.

De *beroepsoriënterende stage* is een stage die doorgaans in het eerste jaar van de opleiding is voorzien. Het doel van deze stage is dat studenten door de praktijkervaring een beter beeld krijgen van wat hun toekomstige beroep inhoudt, meer zicht krijgen op de competenties waarover de beroepsbeoefenaren dienen te beschikken en daardoor een betere afweging kunnen maken of het beroep (en de opleiding) past bij hun ambities. Vanwege het oriënterende karakter van deze stage is er niet echt sprake van meewerken, maar eerder van meelopen en observeren.

De *beroepsopleidende stage* is een stage die doorgaans in de tweede helft van de opleiding plaatsvindt, waarbij studenten voor langere tijd in een werksituatie een deel van hun opleiding volgen. Deze stage kan verschillende vormen aannemen. Zo zijn er stages waarbij studenten meewerken met de collega's en naarmate zij meer ervaring opdoen de werkzaamheden steeds zelfstandiger verrichten. Ook zijn er stages waarbij de studenten een project uitvoeren, hun stagewerkzaamheden zijn dan gekoppeld aan de projecttaken. Tot slot zijn er beroepsopleidende stages waarbij er sprake is van een integratie met het afstudeerwerk van de student. Dan is er doorgaans eerst sprake van een beperkte stagetijd gericht op oriënteren en beperkt meewerken, waarna de student start met het afstudeerproject.

### 20.1.2 Duale opleiding

---

Een andere vorm van werkplekleren is de duale opleiding, waarbij de student een belangrijk deel van de opleiding op de werkplek volgt, en de onderwijsinstelling garandeert dat het gedeelte van de opleiding dat op de werkplek plaatsvindt, voldoet aan de eisen die daar door de opleiding aan worden gesteld. Duale opleidingen vinden vooral plaats in sectoren waar sprake is van volwassen deelnemers zodat het voor hen eenvoudiger wordt om werk en opleiding te combineren. Kenmerkend is dat er altijd sprake is van een arbeidsrelatie

tussen de student en de werkgever. Deze vorm van opleiden heeft nogal wat discussie opgeleverd (Bronneman-Helmerts, 2006), vooral gericht op de vraag of duale opleidingen wel het niveau hebben dat van een opleiding in het hoger onderwijs mag worden verwacht (zie Van Wijngaarden, 2008). In het universitaire onderwijs waren er meer bedenkingen dan in het hoger beroepsonderwijs, maar inmiddels is vooral in het hoger beroepsonderwijs de duale opleiding een algemeen geaccepteerd fenomeen.

### 20.1.3 Leerafdelingen

---

Een tamelijk nieuwe vorm van werkpleklers is de leerafdeling. Kenmerkend voor leerafdelingen is dat een student niet individueel maar samen met andere studenten op een afdeling aan het werk gaat, waarbij zij als team alle werkzaamheden uitvoeren die normaliter tot de taken van die afdeling behoren. Er zijn vormen waarbij de zittende werknemers vooral een coachende rol hebben en de studenten het werk uitvoeren, ook zijn er vormen waarbij de studenten met de zittende werknemers meewerken. Voorbeelden hiervan zijn te vinden in de gezondheidszorg en in het onderwijs, waar studenten van lerarenopleidingen door middel van het concept 'Opleiden in de School' voor een omvangrijk deel van hun opleidingstijd (zo'n 40% van de opleidingstijd) op de werkplek worden opgeleid (zie Lappia, 2010).

### 20.1.4 Communities

---

Een variant op de leerafdeling is de community waarin studenten samen met werknemers, en vaak ook docenten, werken en leren. De idee is dat niet alleen de studenten leren maar dat ook de andere betrokkenen leren van de projecten die gemeenschappelijk worden uitgevoerd. Daarbij gaat het om projecten die een innovatief karakter hebben en die moeten leiden tot kennis en producten die voor de organisatie betekenisvol zijn. Voorbeelden hiervan zijn te vinden in de chemie en bouwopleidingen. Ook zijn er vormen van Opleiden in de School die als community zijn te kenschetsen (Hennissen, 2011).

De besproken vormen staan samengevat in [▣ tabel 20.1](#), waarbij dient te worden aangekend dat het niet de bedoeling is om alle vormen van werkpleklers in het hoger onderwijs te presenteren maar wel om de verscheidenheid onder de aandacht brengen.

## 20.2 Dilemma's en problemen in werkpleklers

---

In deze paragraaf passeren problemen en dilemma's de revue die nagenoeg inherent zijn aan het vormgeven van werkpleklers. Onderzoek waaruit blijkt dat werkpleklers niet van problemen is gespeend, laat zien dat het bijvoorbeeld belangrijk is om eisen aan werkplek en begeleiding te stellen (zie bijvoorbeeld Reenalda (2009) over de discutabele kwaliteit van duale opleidingen). Vooral de ontwikkeling van meer complexe (metacognitieve) vaardigheden is doorgaans lastig door middel van werkpleklers te realiseren (zie Nijhof, 2006).

■ Tabel 20.1 Vormen van werkplekleren

Vorm	Plaats in curriculum	Functie
beroepsoriënterende stage	start opleiding	vaststellen van de match tussen student, toekomstig beroep en opleiding
beroepsopleidende stage	tweede helft van de opleiding	variabel: uitvoeren van voor het beroep kenmerkende werkzaamheden, of uitvoeren van een project dat voor beroepspraktijk relevant is
duale opleiding	vindt plaats gedurende gehele opleiding	combineren van werk en opleiding waarbij het werk, of een deel van de werkzaamheden, deel uitmaakt van de opleiding
leerafdelingen	tweede helft van de opleiding (in lerarenopleiding tijdens gehele opleiding)	studenten werken als groep op een afdeling en leren alle werkzaamheden uit te voeren die kenmerkend voor dat beroep zijn
communities	tweede helft van de opleiding (in lerarenopleiding tijdens gehele opleiding)	studenten werken samen met werknemers (en docenten) en leren gezamenlijk door participatie in innovatieve projecten

### 20.2.1 Opleiding besteedt leren uit maar blijft eindverantwoordelijk

Op de eerste plaats dient een opleiding zich te realiseren dat werkplekleren deel uitmaakt van het curriculum en dat impliceert dat de opleiding ook de eindverantwoordelijkheid draagt voor de kwaliteit van het werkplekleren, inclusief de toetsing. Om dat te kunnen garanderen is van belang om na te denken over de doelen die feitelijk met het werkplekleren worden beoogd, en wat dat betekent voor de eisen die aan het bedrijf, de werkplek en de begeleider(s) worden gesteld.

### 20.2.2 Garanderen van kwaliteit en werkplek

Het is aanbevelenswaardig om een instrument te hanteren waarmee vanuit opleidingsperspectief de kwaliteit van een werkplek kan worden bepaald, alvorens een besluit te nemen om een student te plaatsen. Vooral het niveau van de werkzaamheden, de mate waarin deze aansluiten bij de inhoud van de opleiding, en het leerklimaat (tolerantie voor experimenteren en fouten maken) zijn aspecten die niet altijd de aandacht krijgen die nodig is in de besluitvorming hierover (Streumer & Van der Klink, 2004).

### 20.2.3 Begeleiding van studenten

De begeleiding van studenten tijdens het werkplekleren zorgt voor de nodige hoofdbreken. Welke eisen moeten aan die begeleiding worden gesteld? Niet iedere beroepsbeoefenaar is per definitie geschikt om als begeleider van studenten te fungeren. Naast vakinhoudelijke kennis is ook van belang dat de begeleider het bedrijf goed kent en beschikt over de gewenste didactisch-pedagogische kwaliteiten. Stel bijvoorbeeld dat een opleiding

een visie heeft waarin staat omschreven dat studenten dienen te worden uitgedaagd tot exploreren, dan zijn begeleiders die studenten meteen antwoord geven op hun vragen niet erg effectief, en is de opleiding er meer bij gebaat dat ze onderzoekend gedrag bij de studenten stimuleren. Zeker als van begeleiders uit het bedrijf ook wordt verwacht dat ze een rol spelen in de beoordeling van studenten, is het van belang hen te faciliteren om dat op adequate wijze te doen. Alleen een lijstje met beoordelingscriteria aan de begeleiders verstrekken is volstrekt onvoldoende. De beoordelingsprocedure van het werkpleklers moet overeenstemmen met de eisen die de examencommissie van de opleiding hieraan stelt.

#### 20.2.4 Werkpleklers stelt eisen aan de opleiding

---

Verder is het raadzaam de begeleiding die de docenten van de opleiding verzorgen, te onderzoeken. Het is heel verleidelijk om de problematiek van de begeleiding vooral te situeren in het bedrijf waar de student verblijft, maar ook het begeleiden van werkpleklers stelt eisen aan de eigen opleiding. Begeleiding van werkpleklers vraagt om docenten met actuele kennis en een brede interesse voor het vakgebied, flexibiliteit, goede communicatieve vaardigheden en het vermogen om proactief problemen te signaleren en bij te dragen aan oplossingen hiervoor.

Naast de kwaliteit van de begeleiding is ook de frequentie van de begeleiding een punt van zorg. Het gebrek aan communicatie tussen student, opleidingsdocent en begeleider vanuit het bedrijf is een van de hardnekkigste problemen in het werkpleklers. Het is toch eigenlijk vreemd dat een van de belangrijkste opleidingsonderdelen er zo ongunstig van af komt als het gaat om begeleiding: *out of sight, out of mind* lijkt het heersende credo te zijn.

#### 20.2.5 Het plan als onderlegger voor activiteiten en beoordeling

---

Waar het in veel werkpleklers aan ontbreekt, is instrumentatie van het werkpleklers in het licht van de beoogde opleidingsdoelen, bijvoorbeeld informatiebronnen voor studenten en begeleiders (van de opleiding en van het bedrijf) die hen helpen om helderheid en overeenstemming te realiseren over datgene wat gaat worden geleerd, en over de condities waaronder dit dient te gebeuren. In studiegidsen staat vaak wel omschreven wat er met het werkpleklers wordt beoogd, maar doorgaans is dat in dermate globale termen geformuleerd dat dit niet direct kan worden vertaald naar het werkpleklers van student x in bedrijf y. Additionele instrumenten zijn noodzakelijk om verwachtingen over doelen, invulling traject en wijze van beoordeling te expliciteren en op elkaar af te stemmen.

Er is behoefte aan een consistentie tussen doelen, activiteiten op de werkplek, begeleiding en beoordeling: dit moet op één lijn worden gebracht. Om dat te realiseren, is het noodzakelijk een plan op te stellen waarin wordt omschreven welke werkzaamheden worden verricht en tot welke competentieontwikkeling dat idealiter moet leiden. Parallel hieraan moet een begeleidings- en beoordelingsplan worden geformuleerd die garanties bieden dat deze competentieontwikkeling kan plaatsvinden en waarin wordt afgesproken hoe per competentie deze ontwikkeling wordt vastgesteld en beoordeeld, en wie welke rol hierin gaat vervullen. De inhoud van dat plan kan per student verschillen, al naar gelang

de werkcontext, de aard van de werkzaamheden en de beoogde opleidingsdoelen. Dat betekent ook dat per student moet worden besloten welke mix van beoordelingsinstrumenten relevant is.

### 20.2.6 Werkplekleren is dynamisch

---

Tot slot een dilemma dat inherent is aan werkplekleren: het kent een eigen dynamiek. Elementen van die ‘eigen dynamiek’ zijn dat zowel beoogde intentionele alsook toevallige, incidentele leerervaringen door studenten worden opgedaan en die laatste zijn zeer waarschijnlijk de betekenisvolste leermomenten; de beoordeling ervan vereist feedback en reflectie van verschillende betrokkenen (Guile & Griffiths, 2001). Uiteraard behoren tot de elementen van de eigen dynamiek ook de manier waarop bedrijven zelf tegen leren en professionalisering aankijken. De rol van het werkplekleren wordt daarin als steeds prominenter gezien. In toenemende mate is op dit vlak een ontwikkeling gaande van het sterk managergestuurde leren, waarbij de doelen en dus ook de te behalen resultaten door het management werden bepaald, naar medewerkergestuurd leren (Van der Klink, Boon & Schlusmans, 2012) en dat heeft ook repercussies voor wat bedrijven van studenten verwachten. De dynamiek van de werkplek heeft uiteraard gevolgen voor wat de leermogelijkheden zijn en in het verlengde hiervan voor de beoordeling van het leren. Bij het toetsen van competenties op de werkplek is de authenticiteit groot, maar is er nauwelijks sprake van voorgestructureerde opdrachten (zie ► hoofdstukken 16 en 19). Dat houdt in dat taken op verschillende manieren kunnen worden uitgevoerd, dat voor problemen in de werksituatie er doorgaans verschillende oplossingen bestaan en dat het beoordelen van de kwaliteit van de uitvoering van taken afhankelijk is van de perceptie, expertise en voorkeuren van de betrokkenen.

## 20.3 Instrumenten voor het beoordelen van werkplekleren

---

Voor het beoordelen van werkplekleren kunnen verschillende instrumenten worden gebruikt. Het hanteren van een mix van instrumenten verhoogt de betrouwbaarheid en validiteit. In [■ tabel 20.2](#) worden de belangrijkste kenmerken van de instrumenten samengevat.

### 20.3.1 Portfolio

---

Een (onderwijs) portfolio is een verzameling van onderwijsproducten (papers, beoordelingen, reflecties) waarmee een student kan aantonen de competenties van een opleiding te beheersen. Een portfolio (dat vaak een elektronische component heeft) is een manier om onderwijsproducten te ordenen en om de ontwikkeling in het verwerven van competenties aan te tonen. Over de vorm van het portfolio, de mate van gestructureerdheid en de mate waarin het een instrument is van de student zelf, moet in een vroeg stadium duidelijkheid zijn voor student en opleiding. Het blijkt belangrijk vanuit de opleiding een globaal format voor het portfolio aan te geven; dat voorkomt ongestructureerd verzamelen



■ **Tabel 20.2** Instrumenten voor het beoordelen van werkpleklers

Instrument	Doel	Aandachtspunten
portfolio	vastleggen van de ontwikkeling ten behoeve van tussentijds monitoring alsook voor eindbeoordeling	structuur (gevaar vergaarbak) en het vergt discipline van de student
360 graden feedback	van meerdere kanten belichten van iemands functioneren	kan bedreigend overkomen, training in feedback is gewenst
POP	stimuleert om na te denken over gewenste leersetting en leerdoelen	stimuleren studenten om hiermee aan de slag te gaan is doorgaans nodig
ontwikkelingsgesprekken	vaststellen van de gerealiseerde ontwikkeling en formuleren van aandachtspunten voor de toekomst	gesprek is formatief van aard: voorkom dat het een beoordelend karakter krijgt
performance assessment	beoordelen van uitvoering en resultaat van een taak op de werkplek onder reële condities	invloed van context op taakuitvoering; stelt hoge eisen aan beoordelaars; meerdere beoordelaars zijn nodig
stageverslag	variabel, urenverantwoording, reflectie op uitgevoerde werk en competentieontwikkeling	expliciet doel en opzet en de rol ervan in totale beoordeling van werkpleklers

van informatie en onbetrouwbaarheid. Voor studenten en docenten moet het duidelijk zijn wat het ontwikkelingsdeel en wat het beoordelingsdeel van het portfolio is.

Als studenten niet overtuigd zijn van het nut ervan dan wordt het bijhouden en bijwerken als een lastige en tijdrovende taak gezien. Het is daarom zinvol om studenten te overtuigen van de waarde van een portfolio voor ze er aan beginnen. De waarde ervan is dat een student zowel binnen als buiten de opleiding kan aantonen competenties te hebben verworven; het portfolio zorgt ervoor dat er geen ervaringen en bewijzen verloren gaan en door de manier waarop de competentiebewijzen zijn geordend kan het portfolio bijdragen tot reflectie hierop en tot nadenken over toekomstige leerdoelen en leeractiviteiten (Kicken, 2008). In een portfolio kunnen uiteraard ook de resultaten worden opgenomen van de instrumenten die hierna zijn beschreven.

### 20.3.2 360 graden feedback

De werkwijze van de 360 graden feedback komt uit de wereld van human resource management, waar het wordt gebruikt om iemands functioneren letterlijk van alle kanten te belichten (Jellema, 2006). Ook in het onderwijs en vooral in het beoordelen van werkpleklers wordt het op die manier gebruikt. Dat houdt in dat aan alle betrokkenen bij een taak of traject wordt gevraagd hun beoordeling daarover te geven, vaak aan de hand van een hiervoor ontworpen vragenlijstje met meerdere antwoordcategorieën. Hierbij kunnen de praktijkbegeleider, docenten, medestudenten betrokken zijn. Het voordeel van deze methode is dat het redelijk makkelijk te organiseren is, mits duidelijk is welke competenties door middel van dit instrument in kaart worden gebracht. Ook is een voordeel dat verschillende percepties op iemands functioneren aan bod komen. Een nadeel is dat

deze methode door de student als bedreigend of belastend kan worden ervaren. Dat geldt zowel voor de student in de rol van feedbackontvanger als in de rol van feedbackgever. Dat betekent dat een goede voorbereiding nodig is waarin duidelijkheid wordt verschaft over de diverse rollen waarin iemand wordt geplaatst, dat er een voorbereidende training wordt geboden over professioneel feedback geven en ontvangen en dat er voldoende aandacht wordt besteed aan de interpretatie van de via dit instrument verkregen informatie.

### 20.3.3 POP

---

In een Persoonlijk Ontwikkelingsplan (POP) kunnen studenten aangeven welke competenties of leerdoelen ze in de studie of meer specifiek op de werkplek, willen realiseren, hoe men dat wil doen en hoe dat kan worden gemonitord. Na verloop van tijd blijkt het reflecteren op een POP, nagaan wat er is gerealiseerd, wat er is veranderd en mogelijk een nieuwe oriëntatie en planning te zoeken, een zeer nuttige en leerrijke ervaring. Ook hier geldt dat in een vroeg stadium van de opleiding een goede introductie van het nut en de waarde van een POP belangrijk is om studenten te motiveren een POP op te stellen.

Het is aan de opleiding om de structuur of het format van een POP vast te leggen, of in elk geval aan te geven welke vragen erin aan bod moeten komen. Het gaat daarbij over vragen als: wat is mijn ambitie, wat wil ik realiseren, welk soort werk vind ik leuk en wat zeker niet, hoe kijk ik tegen deze werkplek aan, wat verwacht ik ervan, en hoe kan ik dat gaan realiseren (Bolhuis & Simons, 2011). Een POP helpt om studenten verantwoordelijkheid te laten dragen voor hun eigen leertraject en de keuzes die hierin kunnen worden gemaakt. Het zelf kunnen (mee)sturen van het eigen leren is een bekwaamheid die van steeds meer professionals in hun werk wordt verlangd.

### 20.3.4 Ontwikkelingsgesprekken

---

Ontwikkelingsgesprekken worden vooral gebruikt in het bedrijfsleven. Ook dit is een instrument dat uit de wereld van human resource management komt. Resultaat- en ontwikkelingsgesprekken zoals ze ook vaak worden genoemd, zijn bedoeld om afspraken te bespreken en te maken over toekomstig functioneren en om wederzijdse feedback te geven. Men praat over concreet gedrag, over verwachte en gerealiseerde resultaten, over mogelijkheden die zich voordoen, maar ook over problemen en hindernissen. De gesprekspartners kunnen, als ze dat wenselijk achten, zogenaamde informanten vragen om hun feedback voorafgaande aan het gesprek te formuleren.

Ontwikkelingsgesprekken worden op deze manier gebruikt om het leren op de werkplek van studenten te monitoren. De gesprekspartners zijn in eerste instantie de werkplekbegeleider, al dan niet bijgestaan door de docent, en de student. In het gesprek komt ook hier de wederzijdse feedback aan de orde, het gesprek biedt de mogelijkheid om de student zicht te bieden op de gerealiseerde competentieontwikkeling. Uiteraard is hier eveneens het inschakelen van informanten mogelijk. Voordeel van het ontwikkelingsgesprek is dat het inzicht geeft in de mate waarin een student, als actieve gesprekspartner, erin slaagt zich als (aankomend) professional op te stellen. Een nadeel is dat het ontwikkelingsgesprek

door de student kan worden ervaren als een beoordelingsgesprek. Daarom is het van belang om voorafgaande aan het gesprek het formatieve aspect ervan duidelijk te maken en ook het wederzijdse karakter ervan te benadrukken. Tevens moet worden benadrukt dat zowel het intentionele leren alsook het toevallige, incidentele leren belangrijk is voor iemands competentieontwikkeling. Vaak moeten studenten (en werknemers) worden gewezen op datgene wat, toevalligerwijs, is geleerd voordat men dat herkent als leren en het dus ook in een reflectie- of ontwikkelingsgesprek kan verwoorden.

### 20.3.5 Performance assessment

---

Bij een performance assessment gaat het om de beoordeling van de uitvoering van een authentieke taak in de werksetting zelf, waarbij zowel het handelen als de resultaten ervan object van beoordeling kunnen zijn (zie ► hoofdstuk 16). Voor studenten is een performance assessment een zeer leerzame ervaring omdat zij dan kunnen demonstreren hoe goed ze een bepaalde taak onder realistische condities kunnen uitvoeren. Zoals eerder gesteld is werkplekleren dynamisch en dat heeft ook repercussies voor de inzet van performance assessments. Het vereist van beoordelaars dat zij in staat zijn om door middel van observatie de taakuitvoering en -resultaten in een specifieke context te beoordelen. Doorgaans vergt dit een vertaalslag van de standaarden en de criteria (die algemeen zijn) zodat ze passen bij de desbetreffende taak, context en student. Naast de beoordelaars ligt het voor de hand ook de student hierin een actieve rol te laten vervullen. In aanvulling op de observatie van de performance zelf is het gebruikelijk in een gesprek met de student de argumentatie voor de gevolgde taakuitvoering vast te stellen. Een performance assessment kan zowel formatief (ontwikkelingsgericht) als summatief (proeve van bekwaamheid) worden ingezet. Ze vinden doorgaans plaats in een spanningsveld waarin men recht wil doen aan studenten en hun keuze van de taak en visie op de taakuitvoering. Tegelijkertijd dient de invloed van de werksetting op de uitvoering en resultaten van de taak te worden meegewogen. Zeker bij standaarden en criteria die abstract zijn, is er voor beoordelaars (te) veel ruimte om (impliciet) de eigen opvattingen te verweven in de beoordeling (Elshout-Mohr, Oostdam & Dietze, 2001) en daarom is de inzet van meerdere beoordelaars wenselijk. Dat is soms lastig te organiseren maar het gebruik van nieuwe technologieën kan drempelverlagend werken, bijvoorbeeld door aan studenten te vragen hun taakuitvoering te reduceren tot een gecondenseerde videoclip en deze te zenden naar de beoordelaars.

### 20.3.6 Stageverslag

---

Het is gebruikelijk dat aan studenten wordt gevraagd een stageverslag te schrijven, maar wat nu precies wordt verwacht, is niet altijd duidelijk. Wat is het doel ervan en hoe kan de student onder het schrijven monitoren of het verslag voldoet aan de gestelde criteria? Vanuit de opleiding moet helder worden geformuleerd welke functie, inhoud, volgorde, schrijfstijl en dergelijke wordt verlangd, en op welke tijdstip(pen) de begeleidende docent een versie wil zien (tussentijds of alleen bij afronding). Met name over de functie van

het verslag valt veel te zeggen. Wordt er een overzicht verwacht van de bestede uren? Of fungeert het vooral als een procesverslag waarop de student wordt geacht te reflecteren? Wordt het beoordeeld? En zo ja, door wie en met welke criteria? Is het verslag als zodanig het enige materiaal waarop de beoordeling wordt gebaseerd of vormt het de input voor een mondeling beoordelingsgesprek? Het lijkt zo vanzelfsprekend om een verslag van studenten te vragen maar er zitten veel haken en ogen aan. In sommige sectoren moet er rekening mee worden gehouden dat studenten niet alles mogen verwoorden in verband met bedrijfsgevoelige informatie (dat geldt dan ook voor hun afstudeerproject). Dat is natuurlijk een nadeel bij meerdere beoordelingsinstrumenten waarin schriftelijke materialen worden geproduceerd, maar bij verslagen komt het doorgaans zeer pregnant op tafel. Ten slotte, een verslag wordt doorgaans aan het einde van de periode van werkplekleren geschreven. Het geeft dan wel inzicht in het doorlopen traject en wat daarvan is geleerd, maar gedane zaken nemen geen keer: bijsturen in het werkplekleren van de student is dus geen optie. Het verslag als enige beoordelingsinstrument is daarom niet de adequaatste keus als de opleiding het werkplekleren serieus neemt.

## 20.4 Ten slotte

---

Dit hoofdstuk beoogt aan te zetten tot reflectie op het vraagstuk van het vormgeven van werkplekleren en de beoordeling van dit leren in het bijzonder. Het is evident dat werkplekleren een belangrijke component is van de meeste opleidingen in het hoger onderwijs, ook al zitten de studenten tijdens die periode niet op het opleidingsinstituut. Dat neemt niet weg dat de opleiding toch de eindverantwoordelijkheid draagt voor hun leren, inclusief de beoordeling ervan.

Er is betoogd dat werkplekleren eisen stelt aan de werkplek, aan de begeleiding op de werkplek en aan de begeleiding vanuit de opleiding. Vanwege de dynamiek in de werksituatie is het bijna onmogelijk om voor alle studenten dezelfde werk- en leeractiviteiten te definiëren. Veel eerder ligt het voor de hand om samen met het bedrijf en de student te zoeken naar datgene wat noodzakelijk, wenselijk en haalbaar is en om per student te zoeken naar de optimale combinatie en dit vast te leggen in een plan, inclusief een begeleidings- en beoordelingscomponent.

Vanuit de verantwoordelijkheid die een opleiding draagt voor het werkplekleren is het raadzaam na te denken welke beoordelingsinstrumenten aanbevelenswaardig zijn. Er wordt doorgaans gebruikgemaakt van een stageverslag, maar in dit hoofdstuk is benadrukt dat alleen een stageverslag doorgaans onvoldoende is als beoordelingsinstrument. Een combinatie van instrumenten (methodemix) ligt voor de hand om zowel bij aanvang, tijdens en bij afronding de competentieontwikkeling valide en betrouwbaar te beoordelen.

Bij de vraag hoe de beoordeling in te richten, is het van belang na te gaan wat gebruikelijk is in de desbetreffende professie waarvoor de opleiding kwalificeert. Het beoordelen dient in het perspectief van de heersende werkcultuur van de professie te worden gesitueerd (Sluijsmans, 2013). In veel professies wordt van de beroepsbeoefenaren eigen initiatief en verantwoordelijkheid verwacht voor de eigen carrière en leerloopbaan en dat vraagt van hen dat ze in staat zijn hun eigen professionele ontwikkeling te beoordelen

(Boud, 2010). Daarom is het raadzaam om tijdens de opleiding, en zeker gedurende het werkplekleren, studenten beoordelingsvaardigheden aan te leren en dat impliceert dat zij een actieve rol dienen te vervullen, zowel waar het gaat om de keuze van de instrumenten, alsook waar het betreft het verzamelen en interpreteren van de bewijzen voor hun competentieontwikkeling, en het plannen van de toekomstige leeractiviteiten (Fastre e.a., 2012).

# Toetsen van onderzoekvaardigheden

*Quinta Kools en Ilona Mathijsen*

## Inleiding

Wat is onderzoekvaardigheid?

Welke kaders bepalen de kwaliteit van onderzoek?

Wat is de praktische relevantie voor het werkveld?

Welke instrumenten kunnen worden gebruikt?

Wat is de rol van rubrics?

In hoeverre speelt feedback een belangrijke rol?

## 21.1 Wat is onderzoekvaardigheid? – 242

### 21.2 Waarop beoordelen? – 243

21.2.1 Kaders voor beoordeling – 244

21.2.2 Praktische relevantie en betrokkenheid werkveld – 245

### 21.3 Instrumenten voor beoordeling – 246

21.3.1 Criteriumgericht interview – 246

21.3.2 Observatie – 246

21.3.3 Documentanalyse – 247

### 21.4 Kwaliteit van beoordeling – 247

### 21.5 Formatieve toetsing: de rol van feedback – 248

21.5.1 Typen feedback en kwaliteit – 249


21.5.2 Als begeleider beoordelen? – 249

Zowel in het hbo als het wo is het belangrijk dat studenten beschikken over onderzoekvaardigheid. Aangezien dit voor hbo-studenten nog een relatief nieuwe vaardigheid is, wordt in dit hoofdstuk in de eerste twee paragrafen aandacht besteed aan de specifieke aspecten van onderzoekvaardigheid in het hbo. De overige paragrafen zijn van toepassing voor hbo en wo.

## 21.1 Wat is onderzoekvaardigheid?

Onderzoek is een relatief nieuw onderdeel in het curriculum van het hbo en voor veel docenten is het begeleiden en beoordelen van praktijkonderzoek dan ook een nieuwe taak. Voor studenten is het doen van onderzoek geen doel op zich; zij worden opgeleid voor een beroep (fysiotherapeut, leraar, journalist, verpleegkundige en dergelijke) en praktijkonderzoek is voor deze beroepen een ondersteunende activiteit. Alle studenten in het hbo worden geacht aan het eind van hun opleiding onderzoekvaardig te zijn. Voor bijvoorbeeld laboranten of marktonderzoekers is het doen van onderzoek onderdeel van het dagelijks werk. Maar ook voor andere beroepen is onderzoekvaardigheid gewenst. Immers, in elk beroep is behoefte aan professionals die met een kritisch onderzoekende houding naar hun beroepspraktijk kijken en die bijvoorbeeld beslissingen baseren op systematisch verzamelde en geanalyseerde gegevens. Het onderzoek dat hbo-afgestudeerden doen is te typeren als praktijkonderzoek, gericht op het onderzoeken van een vraagstuk of probleem uit de beroepspraktijk. Er kunnen twee doelen voor onderzoek worden onderscheiden, namelijk (1) onderzoek gericht op het verbeteren van eigen handelen als beroepsbeoefenaar en/of (2) onderzoek met een externe probleemstelling waarin een vraagstuk uit de (beroeps)praktijk wordt onderzocht (Verschuren, 2010). Het doen van praktijkonderzoek is te zien als een leerstrategie waarbij de beroepsbeoefenaar op systematische wijze nieuwe kennis en inzichten ontwikkelt om de praktijk te verbeteren (Bolhuis, 2012). Onderzoekvaardigheid is in toenemende mate van belang omdat continue kennisontwikkeling in alle beroepen noodzakelijk is.

Er is een onderscheid te maken in drie niveaus van onderzoekvaardigheid: 1) bekwaam in het gebruik van onderzoeksresultaten; 2) bekwaam in het geven van opdracht tot onderzoek en 3) bekwaam in het zelf uitvoeren van onderzoek (Verschuren, 2011). Dit hoofdstuk gaat in het bijzonder in op onderzoekvaardigheid op niveau 3; beheersing van dit niveau omvat namelijk ook de niveaus 1 en 2.

Globaal gezegd kent ieder onderzoek – onafhankelijk van de discipline – dezelfde stappen, de zogenaamde onderzoekscyclus (zie onder andere: Verhoeven, 2011; Van der Donk & Van Lanen, 2009). Voor elk van de stappen zijn achterliggende kennis en/of vaardigheden vereist. In  tabel 21.1 worden de stappen en voorbeelden van bijbehorende kennis/vaardigheden weergegeven.

Studenten die onderzoekvaardig zijn, worden geacht over de kennis en vaardigheden te beschikken om alle stappen van de onderzoekscyclus in onderlinge samenhang te kunnen uitvoeren. Zij moeten hierbij tevens in staat zijn om handelingen en keuzes te verantwoorden.

■ **Tabel 21.1** Overzicht van stappen uit de onderzoekscyclus met bijbehorende kennis en vaardigheden

Stap uit de onderzoekscyclus	Benodigde kennis/vaardigheden
1. probleem signaleren in communicatie met belanghebbenden (stakeholders, opdrachtgever)	<ul style="list-style-type: none"> <li>– hoofd- en bijzaken kunnen scheiden</li> <li>– helicopterview</li> <li>– vooronderstellingen kritisch benaderen</li> <li>– belangen afwegen</li> </ul>
2. opstellen onderzoeksvragen (en hypothese)	<ul style="list-style-type: none"> <li>– focus kunnen aanbrengen</li> </ul>
3. bepalen methode voor gegevensverzameling: empirische gegevens en gegevens uit literatuur	<ul style="list-style-type: none"> <li>– kennis van geschikte onderzoeksmethoden</li> <li>– kennis van verschillende manieren van gegevensverzameling</li> <li>– literatuur kunnen zoeken</li> <li>– kennis over kwaliteit en geschiktheid literatuur</li> <li>– analytisch kunnen denken in het verbinden van consequenties van inzichten uit literatuur aan de onderzoeksvragen en aan methode van empirische gegevensverzameling</li> </ul>
4. uitvoeren gegevensverzameling	<ul style="list-style-type: none"> <li>– vaardig in gegevensverzameling zoals verrichten van metingen, uitvoeren van interviews, opstellen en afnemen van enquêtes, doen van observaties en/of uitvoeren van documentstudie</li> </ul>
5. analyseren gegevens	<ul style="list-style-type: none"> <li>– vaardig in kwantitatieve en kwalitatieve analysemethoden passend bij de verzamelde gegevens</li> <li>– analytisch kunnen denken in het verwerken van inzichten uit literatuur ook in relatie tot resultaten uit empirische gegevensverzameling</li> </ul>
6. conclusie en antwoord onderzoeksvragen	<ul style="list-style-type: none"> <li>– analytisch kunnen denken</li> </ul>
7. rapporteren, doen van aanbevelingen, demonstratie, tonen van product	<ul style="list-style-type: none"> <li>– schriftelijk rapporteren</li> <li>– valoriseren van kennis in communicatie met belanghebbenden</li> </ul>

## 21.2 Waarop beoordelen?

Bij het toetsen van onderzoekvaardigheid staan het *proces* en het *product* van het onderzoek centraal. Bij het proces gaat het om de uitvoering van het onderzoek (hoe is het gedaan, zijn de stappen goed doordacht en op de juiste manier uitgevoerd?) en bij het product gaat het om de uitkomsten of de opbrengst van het onderzoek (wat is de uitkomst, het antwoord op de onderzoeksvraag). Bij de beoordeling zal de nadruk sterker op het proces liggen omdat de uitvoering van het proces cruciaal is voor het tot stand komen van het product: een slecht proces kan nooit leiden tot een goed product. Een voorbeeld: twee tweedejaars studenten aan een laboratoriumopleiding zijn slordig met het uitvoeren van metingen (proces), waardoor hun waarnemingen niet kloppen en zij tot verkeerde conclusies (product) komen. Doordat er fouten zijn gemaakt in de uitvoering van het proces is de onderzoeksuitkomst van mindere kwaliteit.



Meestal wordt van studenten verwacht dat ze in een eindrapportage (onderzoeksverslag, scriptie, meesterstuk of soms mondelinge presentatie) beschrijven of vertellen hoe het onderzoek is uitgevoerd (proces) en wat de uitkomsten (product) zijn. Vanwege die samenkomst van proces en product lijkt de eindrapportage de geëigende informatie waarop de beoordeling kan worden gebaseerd. Anderzijds betreft zo'n rapportage slechts een enkel onderzoek en is het de vraag of de student hiermee onderzoekvaardigheid in bredere zin (dus ook op andere momenten of over andere onderwerpen) aantoont. Ondanks deze kanttekening is een onderzoekrapportage een geschikte toetsvorm die inzicht biedt in de onderzoekvaardigheid van de student, zeker als rekening wordt gehouden met de hoeveelheid geboden hulp (Oost, 2004). De rapportage is dan niet alleen een manier om het eindresultaat (product) te presenteren, maar is tegelijkertijd een verantwoording van het handelen als onderzoeker (proces). Het schrijven van een eindrapportage doet tevens een beroep op de schriftelijke uitdrukkingsvaardigheid van de student, iets dat vaak als een apart criterium in de beoordeling wordt meegenomen.

### 21.2.1 Kaders voor beoordeling

Er zijn verschillende kaders die direct of indirect over de kwaliteit van praktijkgericht onderzoek gaan, bijvoorbeeld:

- internationaal geformuleerde niveauaanduidingen voor niveau: bachelor – master, zoals de Dublin-descriptoren;
- nationale kaders als de gedragscode praktijkgericht onderzoek in het hbo (Andriessen, Onstenk, Delnooz, Smeijsters & Peij, 2010) of het Hobéonkader (Stapert, 2012);
- beoordelingsinstrumenten voor studenten aan een specifieke opleiding of aan een cluster van opleidingen.

De Dublin-descriptoren bestaan uit een lijst met criteria (kennis & inzicht; toepassing kennis & inzicht; oordeelsvorming; communicatie; leervaardigheid), met per criterium een globale uitwerking van het eindniveau van een bachelor dan wel master. Deze descriptoren zijn niet per se geformuleerd in relatie tot onderzoek. Toch bieden ze houvast bij bijvoorbeeld de verantwoording van opleidingspecifieke beoordelingsinstrumenten voor onderzoek, zeker wanneer het onderzoek is bedoeld als afronding van de opleiding. Het Dublin-kader geeft wel aanknopingspunten voor verschil tussen bachelor en master, maar vermeldt nauwelijks iets over kwaliteit: wanneer is een onderzoek goed?

Het beoordelingskader van Hobéon (Stapert, 2012) is wel specifiek gericht op onderzoek en richt zich vooral op criteria voor een 'eindwerkstuk'. Er staan criteria beschreven voor de opzet en uitvoering van het onderzoek en voor de (schriftelijke) rapportage. Het Hobéon-kader wordt gebruikt bij accreditaties en bevat ook aanduidingen voor een oordeel over de kwaliteit van het onderzoek. Voor het opstellen van opleidingspecifieke beoordelingskaders is het aan te raden om de eisen die de accreditatie stelt erbij te betrekken om te voorkomen dat eisen van de opleiding afwijken van eisen in de accreditatie.

Ook de gedragscode praktijkgericht onderzoek in het hbo is normatief (Andriessen e.a., 2010). Deze bestaat uit vijf gedragsregels, te weten: ‘Onderzoekers aan het hbo ...:

1. ... dienen het professionele en maatschappelijke belang
2. ... zijn respectvol
3. ... zijn zorgvuldig
4. ... zijn integer
5. ....verantwoorden hun keuzes en gedrag’

Gedragsregel 1 heeft vooral betrekking op de keuze van het onderwerp en op de bruikbaarheid van de uitkomst (product), de andere regels gaan vooral over het proces van onderzoek doen. De rode draad die door alle regels heen loopt, is dat de uitvoerders van het onderzoek op een transparante wijze te werk gaan: zij handelen weloverwogen, verantwoord keuzes en zijn daarop aanspreekbaar. De gedragscode koppelt de vijf regels aan de stappen uit de onderzoekscyclus en per stap staat beschreven wat hiervan de betekenis is voor het handelen. De gedragscode leent zich daarmee voor toepassing in beoordelingsinstrumenten op opleidingsniveau.

De gedragscode, het Hobéon-kader en de beoordelingskaders op opleidingsniveau zijn te typeren als een analytische benadering waarbij verschillende criteria voor onderzoekvaardigheid worden beoordeeld op niveau of kwaliteit. De Dublin-descriptoren zijn te typeren als een meer holistische benadering, waarbij een globaal denk- en handelingsniveau wordt beschreven. Voor het opstellen van eigen opleidingskaders is het aan te raden om een combinatie te maken van deze kaders.

### 21.2.2 Praktische relevantie en betrokkenheid werkveld

---

Een specifiek kenmerk van onderzoek aan het hbo is de praktijkrelevantie. Volgens de gedragscode (Andriessen e.a., 2010) dient de onderzoeker aan het hbo het professionele en maatschappelijke belang. Dit heeft vooral betrekking op de keuze van het onderwerp van onderzoek en op de bruikbaarheid van het onderzoek voor de (beroeps)praktijk. Vanwege dit kenmerk zou hbo-onderzoek daarom moeten worden beoordeeld op de relevantie van het onderzoek voor de beroepspraktijk. Praktijkonderzoek is ‘goed’ wanneer dat onder andere bijdraagt aan (a) de oplossing van een probleem, waarbij (b) alle stakeholders betrokken zijn en dat (c) diepgaande kennis en inzichten oplevert waarmee (d) de praktijk zich blijvend kan verbeteren (Anderson & Herr, 1999). Het gaat dus naast de bruikbaarheid van de uitkomsten voor het werkveld (a, c, d), ook over de wijze waarop belanghebbenden bij het onderzoek worden betrokken (b). Bij dit laatste kan bijvoorbeeld worden gelet op de communicatieve vaardigheden van de student in relatie tot de stakeholders. Voor het beoordelen van deze aspecten van praktijkrelevantie is het raadzaam om (een vertegenwoordiger van) het werkveld te vragen een kritische review te geven. De criteria van Anderson en Herr (1999) kunnen daarbij worden gebruikt als kader.

## 21.3 Instrumenten voor beoordeling

---

Er zijn grofweg drie instrumenten om gegevens te verzamelen waarmee de docent kan vaststellen of de student onderzoekvaardig is, namelijk criteriumgerichte interviews, observaties en documentanalyse.

### 21.3.1 Criteriumgericht interview

---

Een criteriumgericht interview is een interview waarbij beoordelaars vragen stellen over relevant bevonden beoordelingscriteria. Er kan worden doorgevraagd naar het proces van het onderzoek en naar het product. Vragen over het proces zijn bijvoorbeeld: ‘Waarom heb je op die manier gegevens verzameld?’, ‘Vertel eens hoe je te werk bent gegaan bij de analyse van gegevens?’ Vragen over het product zijn bijvoorbeeld: ‘Wat wilde je met het product bereiken?’, ‘Hoe heb je ervoor gezorgd dat het product voldoet aan...?’ Ook zou de student een casus met een onbekende onderzoeksvraag of onbekende onderzoeksgegevens kunnen krijgen voorgelegd met de vraag: ‘Stel, je zou de opdracht krijgen om... hoe ga je dat aanpakken...?’ Door dergelijke vragen te stellen, wordt getoetst of de student over de benodigde kennis en vaardigheden beschikt om een praktijkprobleem te onderzoeken. De antwoorden op de vragen vormen – samen met het oordeel over de kwaliteit van het gerealiseerde product of van het doorlopen proces – een beeld van de onderzoekvaardigheid van de student. Een criteriumgericht interview staat veelal dan ook niet op zichzelf, maar gaat vaak gepaard met een observatie en/of documentanalyse. In dat geval dient het interview om meer informatie te verzamelen dan waartoe de observatie of de documentanalyse gelegenheid biedt.

### 21.3.2 Observatie

---

Observaties hebben betrekking op het product ‘in actie’, op de ontwikkelaar van het product ‘in actie’ of op het onderzoeksproces. Een voorbeeld van het product ‘in actie’ is een demonstratie door de student van de werking van het ontwikkelde product. Wanneer het product ook gedrag van de student impliceert (bijvoorbeeld een dans), dan kan de kwaliteit van het product worden vastgesteld door de student in actie te observeren. Wanneer de demonstratie mislukt, hoeft dat overigens niet per se te betekenen dat de student niet over de benodigde kennis en vaardigheden beschikt. Van observatie van het onderzoeksproces is bijvoorbeeld sprake wanneer de student de opdracht krijgt om tijdens een assessment ter plekke literatuur over een bepaald onderwerp te zoeken en daarmee de vaardigheid in het zoeken van literatuur demonstreert. Of een student die een geluidsopname laat horen over een gevoerd interview en de rationale bij de aanpak van dit interview tijdens een criteriumgericht interview nader toelicht. Observaties zijn met name geschikt wanneer andere verklaringen voor de kwaliteit van demonstraties, anders dan benodigde kennis en vaardigheden, kunnen worden uitgesloten. Daarbij is het de vraag of uit getoond gedrag conclusies over kennis kunnen worden geformuleerd. Observatie zal dan vaak ook niet het

enige instrument zijn waarop de beoordeling is gebaseerd. Een criteriumgericht interview zal er veelal op volgen, of aan voorafgaan.

### 21.3.3 Documentanalyse

---

Een documentanalyse heeft betrekking op verschillende typen documenten: op documenten over het product waarin de student het product beschrijft en verantwoordt, de werking van het product uitlegt of achtergronden verwoordt (bijvoorbeeld handleidingen); of op documenten waarin de student het onderzoeksproces toelicht en verantwoordt, bijvoorbeeld een onderzoeksverslag met een paragraaf of hoofdstuk over de methode van onderzoek. Een analyse van een document met een verantwoording van zowel het product als het proces is het geschiktst voor het toetsen van de onderzoekvaardigheid. Zo'n eindrapportage is voor de student een manier om zowel de uitkomsten (product) alsook de aanpak (het proces) in onderlinge relatie te beschrijven. Tevens verantwoordt de student hiermee het onderzoekend handelen. In het eerder in dit hoofdstuk beschreven voorbeeld van de twee studenten van de laboratoriumopleiding schrijft student A een onderzoeksrapportage met de verkeerde conclusies zonder verdere kanttekeningen over het proces. Student B laat in zijn onderzoeksrapportage zien dat hij zich bewust is van slordigheden en geeft aan dat hij twijfelt over de juistheid van de conclusies. Beide studenten hebben dezelfde slechte uitkomsten opgeleverd, maar student B laat zien dat hij zich bewust is van kritische momenten in het proces. Op de beoordeling van het proces zal student B hoger scoren.

Wanneer een rapportage te vaag is of te zeer beperkt tot een eenmalig onderzoek, kan ook nog een criteriumgericht interview en/of een observatie plaatsvinden om een oordeel te kunnen vellen over de onderzoekvaardigheid.

## 21.4 Kwaliteit van beoordeling

---

Ongeacht of de beoordeling plaatsvindt via een interview, observatie en/of studie van documenten, de beoordeling van een onderzoek wint aan kwaliteit wanneer deze transparant, valide en betrouwbaar is (Stokking, Van der Schaaf, Jaspers & Erkens, 2004; Joosten-ten Brinke, 2011).

Transparantie wil zeggen dat studenten van tevoren weten wat zij kunnen doen om de kans op een positief oordeel te vergroten. En dan gaat het niet alleen om datgene waarop wordt gelet (de criteria), maar ook welke normen daarbij worden gehanteerd (wanneer een oordeel over een criterium positief of negatief uitvalt). Rubrics zijn daartoe een geschikt middel (zie ► hoofdstuk 17). Voordeel van rubrics is dat deze ook kunnen worden gebruikt in verschillende stadia in de ontwikkeling van de student en/of in de realisatie van het onderzoek. Beoordelaars kunnen hun tussentijdse feedback formuleren aan de hand van de rubric: 'je zit nu wat betreft criterium... op beginnersniveau, probeer eens ... om op gevorderd niveau te geraken.'

Validiteit wil zeggen dat men datgene beoordeelt wat men wil beoordelen. Dat betekent dat niet alleen het geschikte instrument moet worden gekozen, maar ook dat de beschreven criteria (in aard, omvang, nadruk enzovoort) representatief zijn voor datgene wat moet worden beoordeeld. Als een opleiding of docent het belangrijk vindt dat de student vooral vaardigheid toont in het verzamelen en verwerken van empirische gegevens, dan dient dit bij de beoordeling de nadruk te krijgen. Als daarentegen de relevantie voor de beroepspraktijk de nadruk heeft, dan is het zaak om bijvoorbeeld samen met het werkveld criteria te formuleren. In het geval van ‘high-inference’-criteria, criteria die op hoog niveau zijn (bijvoorbeeld ‘kritisch onderzoekende houding’), is het belangrijk om te definiëren waaruit zo’n criterium bestaat, waar het uit blijkt. Dit komt ook de communicatie richting studenten ten goede, die dan beter begrijpen wat precies van hen verwacht wordt (zie transparantie).

Bij betrouwbaarheid gaat het niet alleen om de betrouwbaarheid van de beoordeling door één beoordelaar (*intra*beoordelaarsbetrouwbaarheid), maar ook om de betrouwbaarheid van eenzelfde student door meerdere beoordelaars (*inter*beoordelaarsbetrouwbaarheid). In het eerste geval wordt de betrouwbaarheid verhoogd als de beoordelaar zich niet laat leiden door andere informatie dan waarop moet worden beoordeeld, door toevallige omstandigheden op het moment van beoordeling, en dergelijke. Wat betreft de interbeoordelaarsbetrouwbaarheid – dat wil zeggen de mate waarin docenten overeenkomen in hun oordeel – is het wenselijk dat docenten gezamenlijk criteria met normen ontwikkelen aan de hand van concrete onderzoeksproducten van studenten, en bestaande criteria met normen (blijven) aanpassen en toepassen op concrete onderzoeksproducten ten behoeve van consensus nu of in de nabije toekomst. Ook moet rekening worden gehouden met ‘beoordelaarseffecten’ (zie de bijlage achterin dit boek).

## 21.5 Formatieve toetsing: de rol van feedback

Onderzoekvaardigheid kan op verschillende momenten worden beoordeeld: van tussentijdse beoordelingen (formatief, gericht op feedback) tot eindbeoordelingen (summatief) (Joosten-ten Brinke, 2011). De eerder beschreven instrumenten zijn zowel formatief als summatief in te zetten. Het is belangrijk om de instrumenten met de studenten te bespreken zodat ze weten wat het verwachte eindniveau is en op welke aspecten zij worden beoordeeld. Beoordelingen gedurende het proces zijn eigenlijk te typeren als ‘begeleiding’, het gaat om formatieve beoordelingen waarmee de student feedback krijgt om het onderzoeksproces te verbeteren. De feedback die docenten geven kan enorm variëren (Arts, Jaspers & Joosten-ten Brinke, 2013). Om ervoor te zorgen dat de studenten leren van de feedback en niet blindelings de verbeter suggesties overnemen, moet de feedback aan bepaalde voorwaarden voldoen. Zo moet de feedback verwijzen naar de criteria of doelen en aangeven in welk opzicht de student hier nog niet aan voldoet.

### 21.5.1 Typen feedback en kwaliteit

---

Er kunnen vier typen feedback worden onderscheiden (Hattie & Timperly, 2007):

- taak: deze vorm betreft uitspraken over of het werk wel of niet goed is, wel of niet relevant is, en dergelijke. Het gaat dus om feedback over het onderzoeksproduct: de resultaten of uitkomsten van onderzoek;
- proces: dit komt tot uiting in uitspraken over hoe de student te werk is gegaan, in het geval van onderzoek over hoe de student het onderzoek heeft aangepakt;
- zelfregulatie: dit is feedback bedoeld om de student zichzelf te laten evalueren;
- persoon: dit is feedback gericht op kenmerken van degene die de feedback krijgt.

In relatie tot het voorbeeld van de twee laboranten zouden deze typen feedback respectievelijk kunnen luiden: ‘interessante resultaten’ (feedback op taak), ‘jullie hebben het slordig aangepakt’ (feedback op proces), ‘waarom hebben jullie dat instrument gebruikt, en niet bijvoorbeeld...?’ (feedback on self-regulation), ‘jullie zijn onnauwkeurig’ (feedback op de persoon). Feedback over het proces en feedback om zelfregulatie te bevorderen zijn het krachtigst. Feedback op het product is geëigend behulpzaam als studenten daarmee hun handelen kunnen verbeteren. Dat betekent dat niet alleen moet worden aangegeven wat er goed/minder goed is, maar ook waarom. Feedback op de persoon ten slotte, wordt door Hattie & Timperly (2007) als niet-effectief beschouwd.

### 21.5.2 Als begeleider beoordelen?

---

Sommige begeleiders van studentonderzoek vragen zich weleens af of hun oordeel over het eindproduct wel voldoende valide is, gezien de grote hoeveelheid feedback die zij tijdens de begeleiding hebben verstrekt: beoordeel ik nu de student of mezelf? Bij het begeleiden is een grote mate van betrokkenheid gewenst, terwijl bij het beoordelen juist kritische distantie noodzakelijk is (Oost, 2004). Hoe is te voorkomen dat begeleiders ervaren dat zij zichzelf beoordelen in plaats van de student?

Ten eerste kan de begeleider de begeleidende rol geleidelijk afbouwen. Oost (2004) adviseert meer ondersteuning en betrokkenheid bij de start van een project en steeds minder bij de realisatie van het eindproduct. In aanvulling hierop geldt het advies een rubric te gebruiken waarmee de begeleider tussentijds feedback kan geven op een tussenproduct (‘je zit nu op dit niveau, denk ook eens aan ... of probeer eens... om op een volgend niveau te komen’) en waarmee de begeleider ook het eindproduct beoordeelt. Ten tweede bestaat de mogelijkheid om de rol van de begeleidende docent en de ‘zelfverantwoordelijkheid van de student’ mee te wegen als een criterium in het eindoordeel. Ten derde is de kwaliteit van de verstrekte feedback een belangrijke factor. Als de begeleider bijvoorbeeld het schrijven van stukken tekst overneemt, is dat geen goede vorm van begeleiding. Het overnemen van het proces van de student heeft weinig nut voor het leren van de student. Als de docent daarentegen commentaar geeft en concrete suggesties doet ter verbetering, dan is het aan de student om dit commentaar en die suggesties te verwerken. In een uiterst geval kan de begeleider letterlijk ‘voorschrijven’ (bijvoorbeeld voorbeeldzinnen formuleren), in de

verwachting dat de student het type formulering bijvoorbeeld oppakt in de rest van het verslag.

Ten slotte is het noodzakelijk om een tweede of soms zelfs derde beoordeling door een niet-betrokken collega (intern of extern) te laten uitvoeren ten gunste van de betrouwbaarheid van het eindoordeel.

# Toetsen van academische vaardigheden

*Albert Pilot*

## Inleiding

Wat zijn academische vaardigheden?

Welke academische vaardigheden worden onderscheiden in het hbo en wo?

Met welke toetsvormen zijn academische vaardigheden te toetsen?

Welke toets past het best bij het beoordelen van welke academische vaardigheden?

Langs welke stappen kan de keuze voor een bepaalde toets het beste plaatsvinden?

Welke informatie is nodig om die stappen te doorlopen?

**22.1 Kenmerkende competenties voor het hoger onderwijs – 252**

**22.2 Professionele vorming – 253**

**22.3 Academische vorming – 254**

**22.4 Werkwijze om geschikte toetsvormen te selecteren – 255**

**22.5 Drie voorbeelden van een toetsvorm – 257**

22.5.1 De arbeidsproef (hands-on-toets) – 257

22.5.2 Het assessmentcenter (simulatietoets) – 258

22.5.3 De Script Concordance Test – 259

**22.6 Competenties en toetsvormen – 261**

**22.7 Ten slotte – 261**



Beoordelen is moeilijk, en dat geldt in hoge mate als het gaat om academische vorming, beroepscompetenties en complexe vaardigheden. Gezien de inhoud van het onderwijs met dit type doelstellingen is het voor docenten moeilijk geschikte toetsvormen te selecteren. De vraagstelling van dit hoofdstuk is dan ook: welke stappen moeten docenten zetten om een toets te selecteren die optimaal aansluit bij dergelijke doelen van de cursus, en welke informatie hebben zij daarvoor nodig?

## 22.1 Kenmerkende competenties voor het hoger onderwijs

De Wetenschappelijke Raad voor het Regeringsbeleid (WRR, 1995; p. 122) en recenter bevestigd door de Commissie Toekomstbestendig Hoger Onderwijs Stelsel (de commissie Veerman (2010)) maakt een onderscheid tussen de kwalificatie-eisen aan afgestudeerden van het hoger beroepsonderwijs enerzijds en het wetenschappelijk onderwijs anderzijds en stelt voor elke groep drie doelen.

Voor het *hbo* gaat het in het WRR-rapport om:

- Het bereiken van een professioneel denk- en werkniveau door persoonlijke, sociale en algemene vorming.
- Het verwerven van theoretische kennis met betrekking tot een bepaald vakgebied.
- Leren deze kennis toe te passen in een specifiek beroep.

Voor het *wo* gaat het om:

- Het aankweken van een academisch denk- en werkniveau met het oog op een actieve deelname aan de samenleving.
- De voorbereiding op specifieke beroepen (professies) waarvoor een academisch denk- en werkniveau vereist is.
- De voorbereiding op specifiek wetenschappelijke beroepen, zoals docent en onderzoeker.

Deze indeling is weliswaar zeer globaal, maar belangrijk, zeker in het licht van de discussies over het onderscheid tussen *hbo* en *wo*, waar de Commissie Veerman een verdere uitwerking aan gaf naar interdisciplinariteit en een brede algemene vorming. Als aanvulling op het voorgaande past nog een enkele opmerking over het onderscheid tussen *professioneel* en *academisch* (Nedermeijer & Pilot, 2000). De WRR duidt het onderscheid aan met enkele voorbeelden uit verschillende vakgebieden (WRR, 1995, p. 133):

» Een academisch gevormd ingenieur hoort routines op hun merites te kunnen beoordelen en, als het te pas komt, zelf nieuwe routines te kunnen ontwerpen. Een ingenieur uit het hoger beroepsonderwijs dient met gebruikmaking van dergelijke routines, complexe technische of bedrijfskundige problemen op te lossen. (...) In het *hbo* liggen de problemen aan de praktische kant, dat wil zeggen problemen die zich voordoen bij het ontwerpen van producten, het oplossen van bedrijfskundige kwesties. In het *wo* moet de kunde worden ontwikkeld grensverleggend onderzoek te doen, resulterend in nieuwe kennis. «

In het hbo gaat het daarbij om praktijkgericht onderzoek. (Zie ook het themanummer van het Tijdschrift voor Hoger Onderwijs, juli 2013, en ► hoofdstuk 21). Hbo'ers zullen steeds meer competenties moeten leren die inspelen op de ontwikkeling naar een kennismaatschappij. Zij zullen vaker zelfstandig informatie verzamelen en beoordelen op kwaliteit. Zij moeten leren reflecteren op hun ervaringen. De problemen waarmee zij te maken krijgen worden steeds complexer. Het voorgaande is slechts ter toelichting bedoeld op het genoemde onderscheid; een absolute grens is zeker niet te trekken gezien de breedte van het hoger onderwijs en de grote verschillen tussen vakgebieden en professies. Van primair belang is in dit verband dat duidelijk moet worden gemaakt wat de doelen en beoordelingscriteria van een specifieke opleiding zijn teneinde een consistent opleidingsprogramma te ontwikkelen. Deze doelen en beoordelingen dienen het niveau te weerspiegelen dat de studenten volgens de opleidingen – en vaak ook het latere beroepenveld – moeten bereiken (zie voor de bepaling van het 'niveau van een opleiding': Rexwinkel, Haenen & Pilot, 2012).

## 22.2 Professionele vorming

De Vereniging Hogescholen heeft ten behoeve van de visitatiecommissies een algemene karakterisering gemaakt van het niveau van hoger beroepsonderwijs en dat in tien competenties vastgelegd (hbo-raad/Vereniging Hogescholen, 1999; 2011):

- *Brede professionalisering*: kan zelfstandig taken uitvoeren als beginnend beroepsbeoefenaar in een organisatie.
- *Multidisciplinair en geïntegreerd*: heeft kennis, inzichten, houdingen en vaardigheden geïntegreerd vanuit het perspectief van beroepsmatig handelen.
- *Past uitkomsten van wetenschappelijk onderzoek toe* bij vraagstukken waarmee hij in de beroepsuitoefening wordt geconfronteerd.
- *Transfer en brede inzetbaarheid*: past kennis, inzichten en vaardigheden toe in uiteenlopende beroepssituaties.
- *Creativiteit en complexiteit in handelen*: weet om te gaan met vraagstukken waarvan het probleem op voorhand niet duidelijk is omschreven en waarop standaardprocedures niet van toepassing zijn.
- *Probleemgericht werken*: definieert en analyseert complexe probleemsituaties zelfstandig.
- *Methodisch en reflectief denken en handelen*: stelt realistische doelen, pakt werkzaamheden planmatig aan en reflecteert op het beroepsmatig handelen.
- *Sociaal-communicatieve bekwaamheid*: communiceert en werkt samen met anderen in een arbeidsorganisatie.
- *Basiskwalificering managementfuncties*: kan eenvoudige leidinggevende en managementtaken uitvoeren.
- *Besef van maatschappelijke verantwoordelijkheid*: heeft de juiste beroepshouding en betreft daarin relevante ethische aspecten.

In verschillende opleidingen is dit verder specifiek uitgewerkt.

■ **Tabel 22.1** Profiel van de academicus m/v voor de 21e eeuw.

Wetenschapsgebonden competenties	
De academicus	<ul style="list-style-type: none"> <li>– is thuis in het theoretische en methodologische kennisdomein van een of meer wetenschapsdisciplines</li> <li>– heeft basiservaring met de werkwijze van onderzoekers (dus met het wetenschappelijk verantwoord opzetten, uitvoeren en rapporteren van een onderzoeksproject)</li> <li>– is in staat wetenschappelijke en vakpublicaties op te sporen in de eigen (wetenschappelijke of maatschappelijke) beroepspraktijk en is in staat deze publicaties in dat licht kritisch te bestuderen en op hun waarde te schatten</li> <li>– is in staat systematisch en creatief te werken aan het verhelderen en oplossen van concrete problemen die centraal staan in de eigen (wetenschappelijke of maatschappelijke) beroepspraktijk en is in dat verband in staat relevante theoretische en methodologische kennis te gebruiken en toe te passen</li> </ul>
Algemene competenties	
De academicus	<ul style="list-style-type: none"> <li>– is in staat bevindingen op heldere wijze schriftelijk te rapporteren en mondeling te presenteren, afgestemd op een gegeven forum; kan zich zowel mondeling als schriftelijk goed uitdrukken in de Nederlandse taal</li> <li>– beschikt over de denk- en redeneervaardigheden die nodig zijn voor adequate wetenschapsbeoefening en -toepassing</li> <li>– is in staat projecten planmatig op te zetten en uit te voeren; kan werken in teamverband en heeft de communicatieve en sociale vaardigheden die nodig zijn voor het functioneren in arbeidsorganisaties, voor het samenwerken met anderen en voor de omgang met opdrachtgevers, cliënten en tegenspelers</li> <li>– beheerst moderne vreemde talen (actief en passief)</li> <li>– beheerst praktische vaardigheden, bijvoorbeeld computer- en laboratoriumvaardigheden</li> </ul>
Speciale competenties uit een bedrijfstak	
De academicus	<ul style="list-style-type: none"> <li>– heeft de startbekwaamheden voor nadere opleiding en ontwikkeling tot zelfstandige wetenschapsbeoefening in een bepaald vakgebied of tot professionele wetenschapstoepassing (postdoctoraal, postacademisch of in-service)</li> <li>– heeft wetenschappelijk onderwijs genoten van voldoende diepgang en niveau</li> </ul>
Holleman e.a., 1999	

### 22.3 Academische vorming

Een concretere en gedetailleerdere indeling van profielen van academici en academische vorming is gemaakt door Holleman e.a. (1999). Zij onderscheiden een aantal profielen voor academici op basis van een analyse van de literatuur, visitatierapporten, interviews met docenten en studenten. Het profiel 'academicus (m/v) voor de 21e eeuw' is een poging tot synthese. Het profiel is opgenomen in ■ tabel 22.1.

Andere profielen die Holleman e.a. (1999) uitwerken, zijn:

- vier traditionele profielen die nog steeds hun weerklank vinden in het huidige debat over academische vorming: de breed gevormde academicus, de wetenschappelijk gevormde academicus, de disciplinair gevormde academicus en de intellectueel met maatschappelijk verantwoordelijkheidsbesef;
- twee profielen die in het afgelopen decennium naar voren zijn gebracht: de wetenschappelijke vakman en de competente academicus met brede oriëntatie.

In elk van deze profielen worden andere accenten gelegd door specifieke doelen te beschrijven met behulp van zogenaamde taxonomieën, waarvoor tegenwoordig vaak de Structure of Observed Learning Outcomes (SOLO-taxonomie) wordt gebruikt (Biggs & Collis, 1982; Biggs, 1999; Rexwinkel e.a., 2012). Bij de breed gevormde academicus ligt bijvoorbeeld een zwaar accent op algemene beroepsbekwaamheden en attitudes ('is competent in debatteren, disputeren en argumenteren') en algemene persoonlijke bekwaamheden ('is geletterd, cultureel gevormd, deelgenoot van de Europese beschaving').

Essentieel in deze benadering is dat het idee van die ene, alomvattende typering van de 'academicus' en de 'academische vorming' wordt verlaten (Nedermeijer & Pilot, 2000). Uit de analyse volgt een differentiatie met een veelheid aan kenmerken die worden geordend in een beperkt aantal profielen.

## 22.4 Werkwijze om geschikte toetsvormen te selecteren

De werkwijze voor het kiezen van een toetsvorm bestaat uit een aantal stappen. De volgorde van de stappen is niet dwingend. Het is eerder een cyclus waarin men regelmatig de resultaten van vorige stappen moet herzien of aanvullen. Voor het ontwikkelen van de hier beschreven werkwijze (eerder beschreven in Nedermeijer & Pilot, 2000; Elsen e.a., 2001) is vooral gebruikgemaakt van de argumentatie voor de aanpak van het toetsen van vaardigheden van Stiggins (1992), Moerkerke & Dochy (2000), Straetmans (1995) en Van der Vleuten (1995). Voor de lijst van mogelijke toetsvormen is gebruikgemaakt van de studie van Wolters (1998) naar mogelijk bruikbare toetsvormen.

### ■ Stap 1: Welke competenties moeten worden getoetst?

Stap 1 in het selectieproces van een toetsvorm is dat een docent vaststelt welke competenties er in een bepaald onderwijsgedeelte centraal staan. Vanzelfsprekend moet deze stap al worden gezet voor aanvang van het onderwijs. De volgende competenties en deelcompetenties vormen het uitgangspunt:

- *Beroepsspecifieke competenties:*
  - zelfstandig taken kunnen uitvoeren als beginnend beroepsbeoefenaar in een organisatie
  - multidisciplinair kunnen werken
  - competenties in uiteenlopende beroepssituaties kunnen toepassen
- *Wetenschappelijk werken en denken:*
  - systematisch kunnen denken en werken
  - (ingewikkelde) problemen kunnen analyseren en beschrijven
  - zich kunnen inwerken in wetenschappelijke discussies
  - eigen en andermans werk kritisch en creatief kunnen benaderen en bespreken
- *Lezen:*
  - kritisch en doelgericht kunnen lezen
  - teksten kunnen interpreteren, analyseren en samenvatten
  - uitkomsten van wetenschappelijk onderzoek kunnen interpreteren, en de waarde hiervan kunnen beoordelen

- *Onderzoeken en probleemgericht werken:*
  - een probleemstelling kunnen formuleren
  - een eenvoudig onderzoek binnen het eigen vakgebied kunnen plannen
  - een eenvoudig onderzoek binnen het eigen vakgebied kunnen uitvoeren
  - informatie kunnen verzamelen, selecteren en structureren
  - een oplossing of advies kunnen formuleren en beargumenteren
  - verkregen kennis/inzicht op nieuwe problemen kunnen toepassen
- *Schriftelijk kunnen formuleren:*
  - kunnen formuleren
  - kunnen structureren
- *Mondeling kunnen formuleren:*
  - kunnen formuleren
  - kunnen structureren
  - kunnen presenteren volgens de eisen die het vakgebied stelt
- *Sociale en managementcompetenties:*
  - kunnen discussiëren
  - kunnen samenwerken in teams of projecten
  - eenvoudige leidinggevende en managementtaken kunnen uitvoeren

#### ■ **Stap 2: Wat is de belangrijkste toetsfunctie?**

Een belangrijke keuze die moet worden gemaakt, betreft het vaststellen van de toetsfunctie: waarom wordt er getoetst? Toetsen hebben ten minste drie verschillende functies (zie ook ► hoofdstuk 1): selectie, diagnose en sturen en activeren. Vanzelfsprekend kan een bepaalde toets meer toetsfuncties tegelijk vervullen. Waar het om gaat is dat een docent van tevoren vaststelt welke functie(s) de toets moet vervullen.

#### ■ **Stap 3: Aan welke kwaliteitsindicatoren moet een toets minimaal voldoen?**

Of een bepaalde toets voldoet, hangt af van de kwaliteit die de toets op een aantal aspecten kan bereiken. Er zijn diverse soorten kwaliteitsindicatoren (zie ► hoofdstuk 2), waarvan inhoudsvaliditeit, betrouwbaarheid, bruikbaarheid, uitvoerbaarheid en transparantie de belangrijkste zijn. Afhankelijk van de functie die een toets heeft, hebben bepaalde kwaliteitsindicatoren voor toetsen een hogere prioriteit. Bij selectie bijvoorbeeld, en in mindere mate bij diagnose, is het belangrijk het aantal foute beslissingen te beperken. Toetsen meten vrijwel altijd gebrekkig, maar het spreekt voor zichzelf dat in een situatie waarin van de toets veel afhangt, de toets zo inhoudsvalide en betrouwbaar mogelijk moet zijn. Anderzijds kunnen toetsen studenten activeren tot effectief studiegedrag. Het vaststellen van de kwaliteitsindicatoren waaraan een toets minimaal moet voldoen, gezien een bepaalde toetsfunctie en gezien de onderwijssetting, bepaalt in stap 3 de selectie van een toetsvorm.

#### ■ **Stap 4: Welke toetsvormen zijn er en wat zijn hun gebruiksmogelijkheden?**

Om tot de keuze van een toetsvorm te komen, zullen docenten zich moeten verdiepen in de mogelijkheden die er zijn en in de gebruiksmogelijkheden van de verschillende toetsvormen. Dit is om twee redenen belangrijk. Allereerst om docenten zich te laten realiseren dat de toetsinstrumenten die ze al jaren kennen en gebruiken (bijvoorbeeld een

schriftelijke of een mondelinge toets) niet alleen voor het meten van kennis inzetbaar zijn, maar ook voor het meten van verschillende competenties. Bovendien maken docenten zo kennis met nieuwere toetsvormen die zich voornamelijk richten op het meten van competenties. De volgende toetsvormen worden onderscheiden:

- de schriftelijke vraag-antwoordtoets
- de mondelinge toets
- het werkstuk
- het referaat
- toetsen die de beroepspraktijk als uitgangspunt nemen:
  - *hands-on*-toetsen (de arbeidsproef en de 360° feedbackmethode)
  - simulatietoetsen (het stationsexamen en het assessmentcenter)
  - *hands-off*-toetsen (casustoetsen en potlood-en-papiertoetsen)
- toetsvormen waarin verslag wordt gedaan van leeractiviteiten en leerresultaten:
  - het inspanningscontract
  - het leerverslag
  - het logboek
  - de portfolioanalyse
- *peer-, self- en co-assessment*.

Het valt buiten de doelstelling van dit hoofdstuk de genoemde toetsvormen uitgebreid te bespreken; zie daarvoor Wolters (1998), Nedermeijer & Pilot (2000), Milius e.a., (2001), Ruijter e.a. (1995) en verschillende hoofdstukken in dit boek. Wel worden hierna enkele voorbeelden beschreven van toetsvormen: de arbeidsproef, het assessmentcenter en de Script Concordance Test.

## 22.5 Drie voorbeelden van een toetsvorm

### 22.5.1 De arbeidsproef (*hands-on-toets*)

De arbeidsproef, ook wel *worksample test* genoemd (Straetmans, 1998) bestaat uit de opdracht om een aantal taken uit te voeren in een echte arbeidssituatie. Beoordelaars observeren de studenten. De taken moeten representatief zijn voor wat de toets beoogt te meten. De student moet ook op de hoogte zijn van de toetsituatie. Stimuli lokken een bepaalde deelvaardigheid uit. Het is vooraf bekend welke deelvaardigheden zullen worden uitgelokt, waardoor de beoordelaars zich op de beoordelingstaak kunnen voorbereiden. Voorbeelden van dit type toetsen zijn het geven van een les aan het eind van de opleiding tot leraar en het uitvoeren van diagnostisch onderzoek of behandeling van een echte patiënt in de geneeskunde of tandheelkunde.

*Kwaliteitsindicatoren:* de betrouwbaarheid van de toets is zwak doordat de mogelijkheden om de toets te standaardiseren beperkt zijn. Ook de uitvoerbaarheid is zwak omdat er steeds reële werksituaties nodig zijn en omdat de toets veel tijd in beslag neemt.

*Toetsfuncties:* door de grote authenticiteit is deze toets sterk activerend. Dit is dan ook het grote voordeel van de toets. Vanwege de geringe betrouwbaarheid is de toets

niet geschikt voor selectiedoeleinden. De betrouwbaarheid wordt echter verhoogd door gebruik te maken van heldere beoordelingscriteria, getrainde beoordelaars en meerdere metingen. De toets kan in ieder geval diagnostisch worden ingezet, mits de risico's op de werkvloer beperkt blijven.

*Competenties:* met de arbeidsproef kunnen beroepsmatige competenties worden getoetst. De competenties dienen in elk geval expliciet te worden gemaakt. Zo kunnen bij bepaalde arbeidsproeven mondelinge en schriftelijke communicatie, sociale competentie, lezen of onderzoeken worden getoetst. Ook wetenschappelijk werken en denken kunnen worden getoetst bij ingewikkelder opdrachten.

### 22.5.2 Het assessmentcenter (simulatietoets)

Een assessmentcenter bestaat uit een serie gestandaardiseerde opdrachten die kandidaten moeten uitvoeren. De opdrachten vragen om een reactie van de student in termen van waarneembaar gedrag. Bijvoorbeeld: deelnemen aan een groepsdiscussie, reageren op een gespeelde conflictsituatie in het werk, het afhandelen van een stapel post, waarbij van elk stuk moet worden bepaald welke actie hierop moet volgen. Het gevraagde gedrag sluit nauw aan bij de functie of werksituatie waarvoor wordt opgeleid. Het gedrag wordt geobserveerd door getrainde beoordelaars die het gedrag zo objectief mogelijk beschrijven, classificeren en scoren volgens gedetailleerde, vooraf bepaalde richtlijnen. De kandidaten krijgen vervolgens een inhoudelijk verslag waarin hun scores worden weergegeven, alsmede de basis waarop men tot dit oordeel is gekomen. Het verslag wordt met elke kandidaat afzonderlijk besproken. Indien het gaat om een beoordeling in het kader van een keuze voor een opleiding of de keuze voor een bepaalde functie, wordt advies gegeven. Als het gaat om een beoordeling waarbij selectie het doel is, wordt het oordeel toegelicht.

Naast beoordeling door een getrainde beoordelaar, is het ook mogelijk de student zelf of collega-studenten een rol te geven in het beoordelingsproces, de zogenaamde self- of peer-assessment. Het voordeel van self-assessment is dat het de motivatie en betrokkenheid van de student kan verhogen (Harlen & Crick, 2003). Het nadeel is echter dat de studenten zichzelf over het algemeen milder beoordelen dan anderen. Hetzelfde nadeel geldt voor peer-assessment.

*Kwaliteitsindicatoren:* assessments kunnen een redelijke betrouwbaarheid halen, omdat er een hoge mate van standaardisatie mogelijk is. De validiteit van de toets hangt sterk af van de mate waarin men erin slaagt de opdrachten representatief te maken voor de vaardigheden of competenties die men beoogt te meten. Aangezien het bereiken van voldoende standaardisatie en het trainen van de beoordelaars arbeidsintensief is, is de bruikbaarheid van deze toetsvorm beperkt.

*Toetsfuncties:* een assessment kan activerend zijn, omdat de studenten bezig zijn met het demonstreren van hun bekwaamheid in een aantal beroepscompetenties. Assessments worden echter afgelegd in een wat kunstmatige situatie, wat wellicht enige afbreuk aan het activerende karakter doet. Als de opdrachten de competenties toetsen die in de doelen worden omschreven, is een assessment op de juiste manier sturend. Assessments met externe beoordelaars zijn geschikt voor zowel selectieve als diagnostische doeleinden.

Self- en peer-assessments zijn alleen geschikt voor diagnostische doeleinden, of als onderwijsvorm, als het erom gaat zichzelf of elkaar te leren beoordelen.

*Competenties:* Het assessmentcenter is bij uitstek geschikt voor het beoordelen van beroepsmatige competenties. Verder kunnen competenties als onderzoeken en probleemgericht werken, schriftelijke en mondelinge communicatie en sociale competenties worden beoordeeld, mits deze geëxpliciteerd worden. Wetenschappelijk werken en denken, evenals lezen, kunnen eveneens worden beoordeeld, mits de opdracht daarvoor voldoende aanknopingspunten biedt en mits geëxpliciteerd.

### 22.5.3 De Script Concordance Test<sup>1</sup>

Het oordelen en beslissen in situaties van relatieve onzekerheid, op basis van slechts beperkte informatie of onder tijdsdruk, is voor velen – zeker in hogere functies en beroepen – onderdeel van hun professionele praktijk. In een Script Concordance Test (SCT) worden problemen en situaties beschreven in de vorm van een casus met bijbehorende vragen. De vragen hebben de vorm van te toetsen hypothesen of suggesties voor een verdere aanpak, én aanvullende casusinformatie. Deelnemers wordt gevraagd te beoordelen in hoeverre de gesuggereerde hypothese of aanpak bevestigd c.q. verworpen moet worden op basis van de totale casusinformatie in het vignet en de aanvulling. Een casusvignet bevat de eerste informatie over de casus voor zover die relevant is voor de vragen en vergelijkbaar met een 'script'. Het geheel aan informatie geeft geen totaalbeeld van de casus, maar artificiële beperkingen in de beschikbare informatie c.q. onrealistische onzekerheden wordt zorgvuldig vermeden. In principe wordt die informatie gegeven die in de praktijk ook over het (klinische) probleem bekend is op het moment waarop de overweging of beslissing zich voordoet. De test is ontwikkeld en onderzocht in het medisch domein om het redeneren en beslissen in realistische klinische situaties te kunnen meten (Charlin, Roy, Brailovsky, Goulet & Van der Vleuten, 2000; Charlin, Tardiff & Boshuizen, 2000; Charlin & Van der Vleuten, 2004).

Kenmerkend voor de ontwikkeling van een SCT is dat: (a) deze uitgaat van authentieke omstandigheden en problemen en omstandigheden waarbij – net als in de klinische praktijk – meerdere opties mogelijk zijn, (b) gevraagd wordt naar een schatting van waarschijnlijkheden en onderbouwing daarvan op basis van herkenning c.q. het beredeneren van samenhangende patronen, en waarbij (c) enige verscheidenheid in het klinisch probleemoplossen van experts verdisconteerd wordt in de 'standaard' die gehanteerd wordt voor beoordeling van de door deelnemers gekozen aanpak.

#### ■ **Formatief gebruik van deze toets**

Om studenten adequaat voor te bereiden op die professionele praktijk wordt toenemend in opleidingen aandacht besteed aan authentieke vraagstukken en situaties, dat wil zeggen die problemen en omstandigheden die in hoge mate realistisch zijn voor de professionele praktijk en dezelfde (mentale) activiteiten en processen oproepen. Voor het monitoren

<sup>1</sup> Deze paragraaf is gebaseerd op een studie van Ramaekers, Kremer, Pilot, Van Beukelen & Van Keulen (2009).



■ Tabel 22.2 Competenties en geschikte toetsvormen.

	Wetenschap- pelijk werken en den- ken	Lezen	Onder- zoeken en pro- bleem- gericht werken	Schrifte- lijke com- municatie	Mon- delinge commu- nicatie	Sociale en manage- mentcom- petenties	Beroeps- matige compe- tenties
schrifte- lijke toets	++	++	+	+/-	-	-	+/-
mondelin- ge toets	++	++	+	-	+/-	+ (discus- siëren)	+/-
werkstuk	++	+	++	++	-	+/- (samen- werken)	+ (schrij- ven)
referaat	++	+	++	-	++	(+ discus- siëren)	+ (spre- ken)
hands-on- toets	+/-	+/-	+ (toe- passen, oplossen)	+/-	+	++ (360° feedback)	++
simulatie- toets	+/-	+/-	++ (toe- passen, oplossen)	+/- (as- sessment- center)	+ (assess- mentcen- ter)	++ (assess- mentcenter)	++
hands-off- toets	+ (casus; syst. werken)	+/-	++ (toe- passen, oplossen)	+/-	-	+ (leiding- geven)	+
inspan- nings- contract	+/-	+/-	+/-	+/-	+/-	+/-	+/-
leerverslag	+/-	+/-	+/-	+/-	+/-	+/-	+/-
logboek	+/-	+/-	+/-	+/-	+/-	+/-	+/-
portfoli- analyse	+/-	+/-	+/-	+/-	+/-	+/-	+/-
self-, co- en peer-as- sessment	+/-	+/-	+/-	+/-	+/-	+/-	+/-

++ Deze toets is bij uitstek geschikt voor het meten van deze competentie.

+ Deze toets is geschikt voor het meten van deze competentie.

+/- In principe kan deze toets deze competentie meten, maar dit gebeurt bij voorkeur in combinatie met een andere toets. De toets richt zich vooral op andere competenties.

- Deze toets is niet geschikt voor het toetsen van deze competentie.

van de ontwikkeling die studenten daarin doormaken, alsook voor onderzoek naar de effectiviteit van dit onderwijs is het van bijzonder belang te beschikken over een instrumentarium waarmee het probleemoplossend vermogen van studenten in realistische omstandigheden in kaart wordt gebracht. Onzekerheden en ambiguïteit die deel uit kunnen maken van authentieke situaties zijn, vanuit perspectief van validiteit, belangrijke ingrediënten van de test (Van der Vleuten & Schuwirth, 2005).

## 22.6 Competenties en toetsvormen

---

Ter wille van het overzicht is een matrix ontwikkeld (zie [tabel 22.2](#)) waarin per competentie is aangegeven welke toetsvorm geschikt kan worden gemaakt om deze competentie te meten. Als indicator voor geschiktheid is de validiteit genomen, dus de mate waarin een toets kan meten wat men ermee beoogt te meten. De matrix ondersteunt de vierde stap in het toetskeuzeproces.

## 22.7 Ten slotte

---

Uit bovenstaande voorbeelden blijkt dat er problemen zijn en blijven bij het kiezen van een toetsvorm. Het eerste probleem betreft het dilemma van de keuze voor het optimum, want elke toets kent naast voordelen ook nadelen. Er moet meestal een afweging worden gemaakt van voor- en nadelen, of docenten moeten kiezen voor het gecombineerd gebruik van verschillende toetsvormen. Wanneer docenten de tijd nemen om de beschreven stappen systematisch te doorlopen en daarbij weloverwogen keuzes te maken, zal dit leiden tot de keuze voor een redelijk geschikte toetsvorm. Een ander probleem heeft te maken met de selectieve functie van toetsen en de betrouwbaarheid die daarvoor wenselijk is. Veel docenten geven een hoge prioriteit aan de selectiefunctie. Om te selecteren moet een toets echter een hoge betrouwbaarheid bezitten, en hoge betrouwbaarheid is bij veel toetsvormen moeilijk of niet te bereiken, zeker niet bij toetsvormen die niet kunnen worden gestandaardiseerd. De docent moet zich daarom expliciet afvragen of de selectiefunctie in de betreffende cursus wel zo'n hoge prioriteit moet hebben. Een derde probleem is dat authentieke toetsen aantrekkelijk lijken in verband met het activerende karakter, maar dat de uitvoerbaarheid en betrouwbaarheid van dergelijke toetsen nogal eens een knelpunt is. De voor de hand liggende oplossing om dan een combinatie van toetsvormen (methodemix) te gebruiken, brengt weer andere problemen met zich mee: de beschikbare tijd (hoe meer toetsing, des te meer (onderwijs)tijd dat van docent en studenten vergt) en de weging van de resultaten (hoe de toetsuitslagen te wegen ten opzichte van elkaar).

In de bovenstaande procedure is getracht een zo goed mogelijke oplossingsrichting aan te geven voor een lastig probleem: het beoordelen van academische vorming, beroepscompetenties en het kunnen aanpakken van complexe opdrachten. Het is eerder een verzameling van heuristische aanwijzingen dan een verzameling gegarandeerd succesvolle voorschriften. De beschreven procedure leidt echter zeker tot een bewuste en systematische keuze voor een geschikte toetsvorm die bovendien kan worden verantwoord.

# Toetsen van schriftelijke werkstukken

*Sjoerd Romme*

## Inleiding

Wat is het belang van beoordelingscriteria in het toetsen van schriftelijke werkstukken?

Wat zijn goede criteria?

Wie dient werkstukken te beoordelen?

Is de beoordelingsprocedure afhankelijk van het soort werkstuk?

Wanneer en hoe moeten externe deskundigen deelnemen in de beoordeling?

Wat zijn de voor- en nadelen van het betrekken van medestudenten in het beoordelen?

Wat is een afstudeerkring?

**23.1      Onbetrouwbaarheid en terugkoppelingseffecten – 264**

**23.2      Beoordelingscriteria: wat wordt beoordeeld? – 265**

**23.3      Wie beoordeelt? – 268**

**23.4      Hoe wordt getoetst? De afstudeerkring als voorbeeld – 270**

In het hoger onderwijs wordt frequent gebruikgemaakt van schriftelijke werkstukken om bepaalde onderwijsdoelen te realiseren. In het algemeen zijn er twee typen werkstukken te onderscheiden: relatief kleine werkstukken die verslag leggen van de uitwerking van een specifieke opdracht, zoals de analyse van een case of het verslag van een practicum; en grotere werkstukken die naar aanleiding van een specifieke project-, stage- of afstudeeropdracht worden geschreven. Vooral bij afstudeerwerkstukken nemen behalve de docent-begeleider ook andere beoordelaars – zoals collega-docenten, externe deskundigen en de opdrachtgever – deel aan de beoordeling. Daarnaast bestaat de mogelijkheid om medestudenten te betrekken in het beoordelen van schriftelijke werkstukken.

Dit hoofdstuk behandelt de toetsing van schriftelijke werkstukken, zoals afstudeerscripties, stageverslagen en andere werkstukken die in het kader van veel opleidingen dienen te worden gemaakt. Een viertal aspecten van dit toetsingsvraagstuk komt aan bod: waarom is een goede aanpak van de beoordeling van werkstukken noodzakelijk, welke criteria spelen een rol bij het beoordelen, wie beoordeelt: docent, externe deskundige, opdrachtgever en/of medestudent, en hoe wordt beoordeeld in geval van meerdere beoordelaars?

Dit hoofdstuk gaat primair over het beoordelen van individuele werkstukken. Niet dat het beoordelen van werkstukken met meerdere auteurs wezenlijk anders is, maar aan het beoordelen van groepsproducten is een apart hoofdstuk gewijd (► hoofdstuk 24).

## 23.1 Onbetrouwbaarheid en terugkoppelingseffecten

Onderzoek naar de beoordelingspraktijk van schriftelijke werkstukken laat zien dat er zich regelmatig problemen voordoen, enerzijds in de sfeer van betrouwbaarheid, anderzijds met de effecten die de beoordeling heeft op het leergedrag van studenten.

### ■ Betrouwbaarheid

Veel scriptiebeoordelingen stijgen in de praktijk nauwelijks uit boven een 'dubbelsteenresultaat'. Indien bijvoorbeeld collega-docenten een groot aantal goedgekeurde doctoraalscripties anoniem opnieuw beoordelen, krijgt bijna de helft een onvoldoende (Geerligns, 1993). Het onderzoek is wel gedateerd, maar een recenter onderzoek is niet voorhanden. Juist op dit gebied is de laatste tijd wel wat veranderd. Visitatiecommissies in het hoger onderwijs beoordelen ook de afstudeerscripties. Voor zover bekend komen die niet in die mate tot een ander oordeel. Echter dit is niet systematisch onderzocht.

Een belangrijke bron van onbetrouwbaarheid in scriptiebeoordeling is de subjectiviteit van de beoordelaar. Deze zogenaamde beoordelaarseffecten kunnen onder meer optreden in de vorm van persoonlijke beoordelingstendenties (verschillen in gehanteerde normen) en halo-effecten (uitstraling van in het oog springende kenmerken naar andere kenmerken van de scriptie). In de bijlage staat een overzicht van verschillende beoordelaareffecten.

### ■ Terugkoppelingseffecten

Een ander probleem dat zich in de praktijk veel voordoet, is het feit dat een bepaalde manier van terugkoppeling door docenten op werkstukken ongewenst gedrag van studenten uitlokt. Als op ingeleverd werk van studenten niet of nauwelijks feedback wordt gegeven,

werkt dit demotiverend en zullen studenten zich een volgende keer minder inspannen. Dit gaat dan ten koste van kritische (zelf)reflectie, diepgaand begrip en zelfstandigheid (Boud, 1995; Sambell e.a., 1997). Daarnaast is een terugkoppeling die aan de hand van een checklist wordt gegeven, vaak gedetailleerd. Op zichzelf is dat niet zo erg, maar studenten zullen bij het schrijven van de volgende werkstukken het voldoen aan de checklist opvatten als garantie voor een voldoende. Dat is echter lang niet een geval. Een werkstuk is meer dan een optelsom van de te vermijden fouten.

Geen van beide problemen is geheel op te lossen. Enerzijds is de subjectiviteit van de beoordelaar tot op zekere hoogte onvermijdelijk en anderzijds hebben de meeste docenten niet of nauwelijks invloed op eerdere beoordelingen die hun studenten hebben ondergaan. Een goed ontwerp van de beoordelingsmethode volgens een aantal keuzen op het terrein van 'wat', 'wie' en 'hoe' kan genoemde nadelige effecten echter in hoge mate beperken.

## 23.2 Beoordelingscriteria: wat wordt beoordeeld?

---

In het algemeen gaat het bij schriftelijke werkstukken in het hoger onderwijs om het ontwikkelen, gebruiken en toetsen van hogere cognitieve vaardigheden, zoals diepgaand begrijpen, samenhang zien en kritisch reflecteren. Gestandaardiseerde beoordelingsinstrumenten voor het beoordelen van afstudeerwerkstukken en dergelijke zijn zeer beperkt voorhanden. Voor zover wel beschikbaar en toegepast, benadrukken dergelijke instrumenten het belang van lagere cognitieve prestaties – zoals vormaspecten van het werkstuk – ten koste van hogere cognitieve prestaties. Toch is het zeer belangrijk om op basis van de leerdoelen beoordelingscriteria te formuleren die de student duidelijk maken waarop het werkstuk wordt beoordeeld. Dergelijke richtlijnen dienen enerzijds voldoende houvast te bieden voor zowel studenten als docenten en anderzijds hogere cognitieve prestaties te benadrukken.

### ■ Globale beoordeling

Een beoordeling is globaal indien deze plaatsvindt op grond van de totale indruk die het werkstuk op de beoordelaar maakt, en er geen gebruik wordt gemaakt van nader uitgewerkte criteria (Geerligts, 1993). Belangrijk voordeel van globale beoordeling is dat veel ruimte wordt gelaten voor het waarderen van hogere cognitieve prestaties, zoals de samenhang en diepgravendheid van het betoog. Het grote nadeel is dat snel vertekening en vervuiling van de beoordeling optreedt omdat in belangrijke mate beoordelaareffecten kunnen optreden. Specifieke criteria dienen voorop te staan indien het belangrijk is dat de gegeven feedback door de student kan worden begrepen, geaccepteerd en verwerkt. Het is dan wel een voorwaarde dat de feedback is gericht op criteria. Overigens worden vaak zeer algemeen geformuleerde criteria gehanteerd – zoals 'praktijkrelevantie' of 'samenhang van betoog' – die wel in meer specifieke en concrete bewoordingen zijn te omschrijven.

### ■ Analytische beoordeling

Van een analytische beoordeling is sprake indien specifieke en meetbare criteria worden gehanteerd. 'Specifiek' betekent in dit verband dat het betreffende criterium duidelijk

anders is dan andere criteria. Het criterium ‘opbouw van betoog’ is niet duidelijk genoeg te onderscheiden van het criterium ‘relatie tussen bevindingen en conclusies’. ‘Meetbaar’ impliceert dat toepassing van het criterium in een eenduidige conclusie resulteert. Het criterium ‘duidelijke probleemstelling’ laat in het midden wanneer een werkstuk nu wel of niet aan dit vereiste voldoet. Een beter meetbaar criterium is bijvoorbeeld:

De centrale vraag of probleemstelling in dit werkstuk is eenvoudig en begrijpelijk geformuleerd, bijvoorbeeld analoog aan een van de volgende algemene vraagstellingen:

- In hoeverre kan methode of benadering A bijdragen aan het oplossen van vraagstuk B?
- Hoe kan verschijnsel C het beste worden verklaard: met behulp van theorie D of theorie E?

Het belangrijkste voordeel van analytische beoordeling is dat de schrijver van het werkstuk gestructureerde feedback krijgt. Een mogelijk groot nadeel is dat een te nauwkeurige en smalle formulering van criteria kan leiden tot ongewenst gedrag van de student (de eerdergenoemde terugkoppelingseffecten).

#### ■ Beoordelingsschema

■ Tabel 23.1 geeft een voorbeeld van een beoordelingsschema dat een beperkt analytisch karakter heeft en ruimte laat voor een aantal globaal omschreven criteria. De verschillende deelbeoordelingen dienen te worden samengevoegd in een totaaloordeel. Dit beoordelingsformulier is ontwikkeld in onderlinge samenwerking met een aantal studenten in een zogenaamde afstudeerkring (waarover later in dit hoofdstuk meer) in het kader van een universitaire opleiding Bedrijfswetenschappen. In deze opleiding wordt veel aandacht besteed aan de ontwikkeling van schrijfvaardigheden in een eerdere fase van de opleiding. Een en ander betekent dat het formulier ook voor zelfbeoordeling van (conceptversies van) een werkstuk kan worden gehanteerd. Een dergelijk formulier kan in het kader van de beoogde leerdoelen van de betreffende opleiding worden ontwikkeld. Naar keuze kunnen bepaalde criteria, zoals ten aanzien van vormeisen, nader worden gespecificeerd. Andere voorbeelden van beoordelingsschema's zijn te vinden in Jankowicz (2005). Zie ook ► hoofdstuk 17 over rubrics.

Het omzetten van een beoordeling op basis van nader omschreven criteria in een eindcijfer is geen lineair, eenduidig proces. Vooral indien sprake is van verschillende beoordelaars, is het van belang om vooraf vast te stellen hoe de beoordelingen worden omgezet in een eindcijfer, bijvoorbeeld: ‘Voor het cijfer 7 dient op alle aspecten van een voldoende of goede prestatie sprake te zijn.’ Bovendien is het wenselijk om beperkingen op te leggen aan compensatoir beoordelen, bijvoorbeeld: ‘Indien de vorm van het werkstuk voldoende of goed is, is de inhoud (zie criterium 1 t/m 6 in ■ tabel 23.1) geheel bepalend voor het cijfer. Een onvoldoende vorm betekent automatisch ook een onvoldoende eindcijfer, ongeacht de kwaliteit van het werkstuk gezien vanuit de andere criteria.’

■ **Tabel 23.1** Voorbeeld van een beoordelingsformulier voor schriftelijke werkstukken.

Criteria	Omschrijving	Beoordeling o - v - g - zg	Toelichting
1 Vraagstelling	Wordt de centrale vraag op een duidelijke manier ingeleid, of komt deze als het ware uit de lucht vallen? Is de centrale vraag of probleemstelling in dit werkstuk eenvoudig en begrijpelijk geformuleerd, bijvoorbeeld analoog aan een van de volgende algemene vraagstellingen: Hoe kan methode A bijdragen aan het oplossen van vraagstuk B? Kan verschijnsel C het beste worden verklaard met behulp van theorie D of theorie E?		
2 Methode	Beschrijft de auteur de gebruikte onderzoeksmethode(n) op een duidelijke manier (beschrijving); wordt nader toegelicht waarom gekozen is voor deze in plaats van andere methoden (rechtvaardiging)?		
3 Presentatie van bevindingen	Worden de gegevens – in kwalitatieve of kwantitatieve vorm – overzichtelijk volgens een nader aangegeven ordeningsprincipe (bijvoorbeeld chronologisch, naar auteurs, of naar items) gepresenteerd? Wordt op verantwoorde wijze gebruikgemaakt van tabellen, figuren, diagrammen of schema's?		
4 Analyse van bevindingen	Worden de gegevens niet alleen goed beschreven maar ook nader geanalyseerd met behulp van kwalitatieve en/of kwantitatieve analysetechnieken die passen in de gekozen onderzoeksmethode(n)?		
5 Conclusies	Volgen de conclusies logisch uit de (analyse van de) bevindingen? Geven de conclusies een duidelijk antwoord op de centrale vraag?		
6a Relevantie	Voor werkstuk zonder expliciete opdrachtgever: zijn de vraagstelling, bevindingen en conclusies interessant, dat wil zeggen van een nader omschreven belang voor bepaalde personen, groepen of organisaties?		
6b Aanbevelingen	Voor werkstuk gemaakt voor opdrachtgever: volgen de aanbevelingen logisch uit de conclusies? Hoe evalueert de opdrachtgever de resultaten en aanbevelingen die in dit werkstuk worden gedaan? Beschrijft de auteur in de aanbevelingen specifieke acties en initiatieven, of worden deze in het midden gelaten?		
7 Vorm	Heeft het werkstuk een omvang die past bij de inhoud? Zijn er veel of weinig typografische fouten? Is de tekst prettig leesbaar? Wordt op verantwoorde wijze van appendices gebruikgemaakt, en op correcte wijze naar literatuur verwezen? Is de alinea-indeling consequent? Enzovoort.		

Tabel 23.1 Vervolg

Criteria	Omschrijving	Beoordeling o - v - g - zg	Toelichting
Eindcijfer	Richtlijn: 5 of lager, indien één of meer aspecten onvoldoende is/zijn 6 indien alle aspecten voldoende zijn 7 indien minimaal de helft van de aspecten goed en de overige aspecten voldoende zijn 8 indien alle aspecten goed zijn (of tegenover elke voldoende een zeer goed staat) 9 indien meeste aspecten zeer goed zijn en de overige goed zijn 10 indien alle aspecten zeer goed zijn		

### 23.3 Wie beoordeelt?

#### ■ Begeleider-beoordelaar

De begeleider heeft over het algemeen de bevoegdheid en deskundigheid om schriftelijke werkstukken te beoordelen. De dubbelrol van begeleider en beoordelaar impliceert het gevaar dat de begeleider tot op zekere hoogte zichzelf beoordeelt omdat in de frequente interactie tussen begeleider en student gezamenlijke beslissingen ten aanzien van het schriftelijk werkstuk worden genomen, waardoor de begeleider als het ware mede-eigenaar is van het eindproduct en de wijze waarop het tot stand is gekomen. Hiermee hangt het gevaar samen dat het *proces* in plaats van het *product* wordt beoordeeld. Naarmate er meer sprake is van zelf- en procesbeoordeling, zal het ook moeilijker zijn om bijvoorbeeld een negatief oordeel uit te spreken.

#### ■ Externe beoordelaar

Externe beoordelaars spelen een belangrijke rol in het corrigeren van mogelijke vertekeningen in de beoordeling als gevolg van genoemde dubbelrol. Het is dus zaak om vooral bij uitgebreide werkstukken, zoals stageverslagen en scripties, externe beoordelaars in de toetsing te betrekken (zie ook ► hoofdstuk 21 over het toetsen van onderzoeksvaardigheid). Externe beoordelaars zijn 'extern' in die zin dat zij niet bij de totstandkoming van het werkstuk worden betrokken maar wel over relevante deskundigheid beschikken. In veel onderwijsinstellingen vervult een collega-docent deze rol. Een externe beoordelaar kan ook in een geheel andere rol dan die van de docent bij het werkstuk betrokken zijn, zoals een externe opdrachtgever en/of begeleider vanuit een stageorganisatie. In geval van opdrachtgevers kunnen deze de beoordeling bovendien verdiepen en aanvullen vanuit hun specifieke gezichtspunt en belang. Belangrijk is dat externe beoordelaars op de hoogte zijn van de criteria en hebben geleerd hoe zij met deze criteria moeten omgaan. Bij het inzetten van beoordelaars van buiten de onderwijsinstelling kan zich een juridisch vraagstuk voordoen. Alleen personen die zijn benoemd als examinator mogen studenten beoordelen. En deze mogelijkheid is als zodanig meestal niet expliciet in het examenreglement



opgenomen (maar ook niet is uitgesloten). In dat geval kan de begeleider de externe beoordelaar in ieder geval altijd zelf in het beoordelingsproces betrekken, om mogelijke vertekeningen in de eigen beoordeling te kunnen corrigeren.

#### ■ Medestudenten

Deelname in de beoordeling door medestudenten is vooral interessant om de uitwisseling van kennis en ervaring tussen studenten te bevorderen. Daarnaast dragen medestudenten ook in hoge mate bij aan transparantie en legitimiteit van de beoordeling: indien immers de medestudent de beoordeling begrijpt en accepteert, is de kans groot dat dit ook geldt voor de auteur van het werkstuk. Echter, ook hier doet het probleem zich voor dat studenten niet bevoegd zijn tot het afnemen van examens. De oplossing is dat de beoordeling van studenten over het werkstuk in formele zin wordt uitgesproken door de examenbevoegde begeleider-beoordelaar. Ook hier is het van belang dat de student als beoordelaar zich heeft geschoold op het gebied van beoordelen.

#### ■ Analogie met de professionele praktijk bij de totstandkoming en beoordeling van product

De positie van de externe beoordelaar en medestudenten in de beoordeling kan naar analogie van bedrijfsprocessen worden begrepen. Hierin is het oordeel van de klant of opdrachtgever uiteindelijk doorslaggevend; dit oordeel heeft geheel betrekking op het product zelf, zonder veel kennis van of betrokkenheid bij het productieproces en de producent. Docent-begeleiders vervullen een sturende rol waarin zij zich nadrukkelijk ook richten op zowel de persoon van de auteur als het totstandkomings- en begeleidingsproces. Medestudenten spelen mogelijk een rol als beoordelaar door als collega's van de auteur, al dan niet tussentijdse, feedback te geven (te vergelijken met interne kwaliteitscontrole).

De analogie met beoordeling van producten in het bedrijfsleven impliceert dat de compleetste en betrouwbaarste beoordeling plaatsvindt door verschillende beoordelaars die vanuit verschillende perspectieven en belangen het werkstuk toetsen. Er vindt dan confrontatie en afstemming tussen verschillende beoordelingen plaats, waardoor een mogelijke vertekening van de afzonderlijke beoordelingen zichtbaar wordt gemaakt en veel verschillende aspecten van het werkstuk aan bod kunnen komen. Bij het beoordelen van werkstukken van een zeer beperkte omvang wegen de voordelen van het inschakelen van verschillende beoordelaars natuurlijk niet op tegen de extra kosten.

De formatieve functie van studenten als beoordelaars is vooral van belang voor de kwaliteit van het beoordelingsproces. Het niet betrekken van medestudenten in de beoordeling vermindert de kans op mogelijke onderlinge leereffecten en kan ook leiden tot geringere transparantie van het beoordelingsproces en de uiteindelijke beoordeling. Het beoordelingsproces wordt erg kwetsbaar indien men de externe beoordelaar weglaat, terwijl medestudenten wel deelnemen in de beoordeling. Onder studenten kan bijvoorbeeld de neiging ontstaan elkaar te ontzien en het commentaar op en de beoordeling van anderen werkstuk af te zwakken. Een extreme vorm van dit gedrag is het maken van afspraken over de beoordeling ('Ik beoordeel jouw werkstuk met een 9, en jij doet hetzelfde met mijn werkstuk'). Elk scenario zonder externe beoordelaars is derhalve af te raden.

### ■ Integrale benadering

De effectiefste aanpak van de toetsing van werkstukken met een grote omvang is die waarin docent, externe beoordelaar en medestudent(en) deelnemen. Deze integrale benadering is alleen mogelijk door een duidelijke procedure te hanteren waarin is vastgelegd welke rollen en bevoegdheden de verschillende deelnemers hebben en hoe de verschillende beoordelingen op elkaar worden afgestemd. Een dergelijke integrale aanpak loont zich natuurlijk vooral bij het beoordelen van omvangrijke werkstukken. Onderzoek naar dergelijke integrale benaderingen laat zien dat deze het leergedrag van studenten en de kwaliteit van de beoordeling bevorderen (Sambell e.a., 1997; Dochy e.a., 1999). Een integrale benadering leidt niet per definitie ook tot grotere transparantie, zeker niet wanneer het gesprek tussen de verschillende personen geen inhoudelijk maar een politiek karakter heeft ('Jij zit op een zeven, ik op een zes. Zullen we er een zesenhalf van maken?'). Als gevolg hiervan krijgt de student een eindcijfer voor het werkstuk zonder precies te weten waarop dit is gebaseerd.

## 23.4 Hoe wordt getoetst? De afstudeerkring als voorbeeld

---

Indien een groter aantal (dat wil zeggen drie of meer) beoordelaars werkstukken van grotere omvang beoordeelt, dan is de vraag hoe dit dient te worden aangepakt. In het algemeen dient in dat geval sprake te zijn van een duidelijke procedure waarin bevoegdheden, beslisregel en besluitvormingsproces zijn vastgelegd. De vraag naar een transparante procedure kan op vele manieren worden beantwoord. Vooropstaat dat behalve een duidelijke omschrijving van de criteria en de deelnemende personen ook een nadere afbakening en beschrijving van de procedure nodig is. De beoordelingsprocedure in zogenaamde afstudeerkringen kan hier als referentiepunt dienen (Romme & Nijhuis, 2000).

### ■ Afstudeerkring

Een afstudeerkring is een instrument voor het begeleiden en beoordelen van afstudeerwerkstukken dat vooral wordt gebruikt in situaties waarin een groter aantal studenten een werkstuk schrijft over een gemeenschappelijk thema. Een afstudeerkring bestaat uit een of meer docent-begeleiders en vijf tot vijftien studenten in de afstudeerfase, die gezamenlijk de verantwoordelijkheid voor de begeleiding en beoordeling van de werkstukken van de deelnemende studenten dragen. Als zodanig vormt een afstudeerkring een permanente begeleidingsgroep waarin studenten in principe op elk moment kunnen instromen, en na hun afstuderen weer vertrekken. Het doel van deze manier van werken is zowel het vergroten van de kwaliteit en doelmatigheid van de begeleidingswerkzaamheden als het vergroten van de kwaliteit van de werkstukken. Voordeel voor de docent is grotere doelmatigheid door schaalvoordelen: minder voorbereiding doordat het leeswerk tijdens het begeleidingsproces over verschillende deelnemers wordt verdeeld, en minder gesprekstijd omdat de docent met verschillende studenten gelijktijdig een bijeenkomst heeft. De afstudeerkring stimuleert daarnaast samenwerkend leren en versterkt de feedbackfunctie van toetsing.

Tijdens de bijeenkomsten worden voorstellen voor werkstukken en tussentijdse conceptversies van hoofdstukken en dergelijke besproken. In feite vindt dus een doorlopend proces van formatieve toetsing van (delen van) de werkstukken-in-voorbereiding plaats, waarbij de feedback zowel voor de student die schrijft als voor andere studenten leerzaam kan zijn. De deelnemende student schakelt in vrijwel elke bijeenkomst minimaal één keer tussen de rol van schrijver en die van medebegeleider van de andere studenten. De eindversie van het werkstuk wordt beoordeeld in een bijeenkomst van de kring, waarin ook een externe beoordelaar deelneemt. Zowel collega-docenten als deskundigen van buiten de onderwijsinstelling (zoals de eventuele opdrachtgever) kunnen de rol van externe beoordelaar vervullen.

### Spelregels en beoordelingsprocedure in afstudeerkring

Hier volgt een weergave van belangrijke spelregels met betrekking tot beoordeling, zoals in een reglement vastgelegd. Voor een nadere uitwerking van spelregels en beoordelingsprocedure in afstudeerkringen (inclusief de examenreglementaire aspecten), zie Romme & Nijhuis (2000). De regels onder a t/m d betreffen algemene afspraken, en vanaf e wordt de beoordelingsprocedure weergegeven.

#### Algemeen

- a. Het doel van de afstudeerkring is het begeleiden van afstudeerprojecten op het gebied van (domeinomschrijving), op zodanige wijze dat dit leidt tot afstudeerwerkstukken waarmee de betrokken studenten hun studie kunnen afronden.
- b. De docent-begeleider(s) bepaalt (bepalen) het domein van de afstudeerkring.
- c. De kring neemt beleidsbeslissingen op basis van het geïnformeerd *consent-beginsel* (*informed consent*). Deze beslissingen betreffen onder meer de inhoud, vorm en frequentie van de bijeenkomsten, gespreksleiderschap, secretarisfunctie en beoordeling van werkstukken.
- d. Volgens het '*informed consent*'-beginsel staat bij de besluitvorming de argumentatie voor het nemen van een beslissing centraal. Dat wil zeggen dat een voorstel is aangenomen als geen van de aanwezige kringleden daartegen een beargumenteerd bezwaar heeft.

#### Beoordeling

- e. De criteria voor de beoordeling van werkstukken zijn in een afzonderlijk document vastgelegd.
- f. In geval van eindbeoordeling van een afstudeerwerkstuk neemt minimaal één externe beoordelaar deel aan (het betreffende deel van) de bijeenkomst. Deze externe beoordelaar(s) is/zijn niet betrokken geweest bij de totstandkoming of begeleiding van het werkstuk. In voorkomende gevallen levert/leveren de externe beoordelaar(s) een schriftelijke beoordeling op basis van de vastgestelde criteria.
- g. In geval van eindbeoordeling van een afstudeerwerkstuk, neemt minimaal één bevoegde docent-begeleider deel aan (het betreffende deel van) de bijeenkomst.
- h. In de eindbeoordeling van een afstudeerwerkstuk nemen alleen de aanwezige studentleden van de kring deel die minimaal twee eerdere bijeenkomsten hebben bijgewoond.

- i. De auteur van het afstudeerwerkstuk neemt niet deel in de besluitvorming over het eindoordeel van het betreffende werkstuk (die in zijn/haar aanwezigheid plaatsvindt).
- j. Indien de afstudeerkring niet tot een beslissing komt, bepalen de externe beoordelaar(s) en docent-begeleider(s) in onderling overleg het eindcijfer.

De volgende procedure vormt een handvat voor de gespreksleider bij het beoordelen van scripties en stageverslagen op basis van het 'informed consent'-principe (zie c en d):

1. De auteur van het te beoordelen werkstuk (hierna: scriptant) houdt een korte inleiding – van maximaal vijftien minuten – waarin het eigen werkstuk wordt geëvalueerd en waarbij kan worden ingegaan op de belangrijkste conclusies en leerpunten.
2. Ieder aanwezig kringlid, behalve de scriptant, krijgt een stembriefje en schrijft daarop de eigen naam en het voorgestelde cijfer. Alle deelnemers leveren hun stembriefje bij de gespreksleider in. De gespreksleider vraagt vervolgens in een ronde iedere deelnemer om argumenten voor het voorgestelde cijfer te geven (zie beoordelingsformulier).
3. De gespreksleider geeft de scriptant de gelegenheid om te reageren (zonder discussie) en, indien gewenst, om het werkstuk uit de beoordelingsprocedure terug te trekken.
4. Iedere deelnemer krijgt in een ronde de gelegenheid zijn/haar cijfervoorstel te wijzigen naar aanleiding van de gehoorde argumenten.
5. De gespreksleider doet een voorstel voor een cijfer en vraagt alle deelnemers (behalve de scriptant) in een ronde om consent voor dit voorstel. Heeft niemand een overwegend bezwaar, dan is het voorstel aangenomen.
6. Deelnemers die geen consent geven, moeten hun bezwaar na laatstgenoemde ronde toelichten. Vervolgens kan een open discussie ontstaan waarin de gegeven argumenten worden getest op hun geldigheid en gewicht.
7. De gespreksleider vraagt aan de personen die in stap 6 geen consent gaven, of hun bezwaar overwegend is. Is dit niet het geval, dan wordt het voorstel alsnog aangenomen. Is minstens één bezwaar wel overwegend, dan vraagt de gespreksleider om een gewijzigd cijfervoorstel van de deelnemers die eerder in stap 2 of 4 (indien gewijzigd) het cijfer hadden voorgesteld waartegen nu een overwegend bezwaar bestaat. Ook andere deelnemers kunnen eventueel hun voorstel wijzigen.
8. Op basis van de nieuwe informatie kan de gespreksleider met een nieuw voorstel komen. Dit proces herhaalt zich vanaf stap 6 tot er een voorstel is waartegen geen van de deelnemers een overwegend bezwaar heeft.

# Toetsen van groepswerk

*Anneke Bax, Jacob Perrenet en Henk van Berkel*

## Inleiding

Waarom is samenwerken belangrijk?

Waarom kiezen voor groepsproducten?

Hoe groepsproducten beoordelen?

Kunnen groepsoordelen ook een oordeel zijn over individuele bijdragen?

Hoe is 'meeliftgedrag' te ontdekken?

Hoe zijn een hoge validiteit en betrouwbaarheid te bereiken?

- 24.1 Overwegingen voor een groepsproduct – 274
- 24.2 Beoordelen van groepsproducten – 275
- 24.3 Opzetten van een beoordelingssysteem – 276
- 24.4 Methoden voor beoordeling: het betrekken van medestudenten – 280
- 24.5 Uitwerking van enkele beoordelingsmethoden – 281
- 24.6 Kwaliteitseisen aan de beoordeling – 283

Er zijn weinig beroepen waarin afgestudeerden solitair, zonder collega's, werken. Veelal zal er sprake zijn van samenwerking of teamwerk, waarbij afgestudeerden zich samen met collega's buigen over een bepaald probleem, afspraken maken, vergaderen, en soms het voortouw (moeten) nemen. Hierbij is het belangrijk dat zij elkaar feedback geven en het groepsproces goed laten verlopen om tot een gewenst resultaat te komen.

Dus samenwerken en het gezamenlijk maken van een product is belangrijk. Onderwijs moet studenten op deze activiteiten voorbereiden. Een onderwijsvorm die zich hierop richt is het actief participeren in een groep en het maken van een groepsproduct. De onderwijsgroep fungeert dan als een practicum waarin studenten de voor samenwerking benodigde vaardigheden oefenen en toepassen, en waarin zij leren in een groep te werken. Daarom is het alleszins redelijk dat de opleiding toetst in hoeverre studenten dit soort vaardigheden na verloop van tijd machtig zijn.

Het beoordelen van samenwerking en groepswerk is nogal complex en moeilijk objectief uit te voeren, zeker als niet alleen het product maar ook het groepsproces en de samenwerking onder de beoordeling vallen. Meestal krijgen alle groepsleden hetzelfde eindcijfer, ook als er binnen de groep grote verschillen bestaan in wat ieder heeft gedaan en geleerd. Als het evident is dat bepaalde studenten meer kennis van zaken hebben dan de rest van de groep en een grotere bijdrage hebben geleverd aan het functioneren van de groep, zou een docent hun een hoger cijfer moeten kunnen toekennen; of juist een lager cijfer aan iemand die voornamelijk heeft 'meegelift' met de groep (Bax, 2004). Mogelijkheden voor een gedifferentieerde waardering wanneer er binnen de groep grote verschillen bestaan in geleverde bijdragen, zullen daarom aanwezig moeten zijn. Opleidingen die groepswerk belangrijk vinden, zullen daarom een beoordelingssysteem moeten gaan ontwerpen dat hen in staat stelt de verschillende facetten die kleven aan het beoordelen van groepswerk op een correcte manier toe te passen.

## 24.1 Overwegingen voor een groepsproduct

Vrijwel iedere opleiding in het hoger onderwijs kent studieonderdelen waarbij de afsluiting bestaat uit het maken van een groepsproduct. Hierbij gaat het erom dat studenten in een team kennis en inzicht weten te integreren in een gezamenlijk eindproduct. Dit komt voor bij specifieke onderwijsvormen, zoals projectonderwijs, maar ook bij afzonderlijke vakken, practica, en dergelijke. Twee of meer studenten maken een groepsproduct dat kan bestaan uit een schriftelijk werkstuk, een project- of practicumverslag, een ontwerp of een uitwerking van – meestal wat complexere – opgaven. Of een groepsproduct de aangewezen vorm is, hangt samen met de doelstellingen. Groepsproducten hebben betrekking op complexere vaardigheden en minder op het verwerven van algemene basiskennis en basisvaardigheden. Het gezamenlijk maken van een eindproduct is in veel beroepsgroepen een essentieel onderdeel van het werk. Daarnaast somt Davies (2009) als voordelen op dat het werken in groepen diep en actief leren bevordert, kennis structureert en vaak een positieve invloed heeft op motivatie. Bovendien, en zeker niet minder belangrijk, geldt dat samenwerken in een groep vrijwel altijd leidt tot iets beters,

tot betere producten of tot meer kennis. Het spreekwoord ‘twee weten meer dan één’ is in deze situatie zeer toepasselijk. Bij het samenwerken worden kennis en vaardigheden overgedragen en gedeeld.

Indien doelstellingen een van de volgende zinsneden bevatten, ligt een groepsproduct voor de hand:

- het gezamenlijk kunnen werken aan een opdracht (taakuitvoering)
- het projectmatig kunnen werken en functioneren in een groep (procesgericht)
- het als groep kunnen opzetten van een verslag (schriftelijk rapporteren)
- het kunnen verdedigen van een groepsproduct (presentatie)

Hoewel de keuze voor groepsproducten op de eerste plaats moet zijn ingegeven door de doelen, kunnen er ook praktische argumenten zijn, bijvoorbeeld om redenen van efficiëntie en kostenbesparing: kostbare apparatuur voor practica, kosten voor proefpersonen, vereiste aantallen proefdieren, nakijkwerk en/of aantal beoordelaars.

Als is gekozen voor de afsluiting van een studieonderdeel met een groepsproduct, moeten de opdrachten of taken zich hiervoor lenen. Ook moet vooraf duidelijk zijn geformuleerd waarom dit beoordelingsinstrument wordt gehanteerd en hoe de beoordeling zal plaatsvinden. Dit is van invloed op de werkwijze en het studeergedrag van de groep studenten. Voor een docent is het van belang bij de beoordeling duidelijkheid te verkrijgen over het aandeel van iedere student bij het tot stand komen van het eindproduct. Voor studenten is het van belang om te worden ‘afgerekend’ op wat zij daadwerkelijk hebben gepresteerd. Wanneer de groep bijvoorbeeld niet individueel maar *als groep* wordt beoordeeld op een gezamenlijk eindproduct, is het zeer frustrerend wanneer bepaalde groepsleden hun afspraken niet nakomen of als één student al het werk doet en anderen slechts ‘meeliften’. In ► paragraaf 24.4 staan enkele manieren om meeliftgedrag te ontdekken. Om het te voorkómen heeft Davies (2009) een praktische handleiding geschreven waarin hij ook uiteenzettingen geeft over de aard en omvang van de samenstellingen van groepen, over allerlei problemen die zich in groepen kunnen voordoen en over de soorten opdrachten. Daarnaast noemt hij nog dat het samenwerken in groepen ook een nadeel kan hebben voor vooral de betere studenten. Die voelen zich juist door het (lagere) niveau van de medestudenten benadeeld omdat zij de kwaliteit van het groepsproduct omlaag trekken.

## 24.2 Beoordelen van groepsproducten

In dit hoofdstuk wordt uitgegaan van een groepsproduct waar een groep studenten gedurende minimaal enkele weken aan heeft gewerkt. Als een opleiding het van belang vindt dat studenten in een groep leren om gezamenlijk te werken aan een eindproduct, zal dit leerproces deel moeten gaan uitmaken van de beoordeling. Het beoordelen van groepsproducten is daardoor nogal complex. Maar dat is niet de enige reden voor de complexiteit. Het maken van een groepsproduct vereist vaak hogere cognitieve vaardigheden, die doorgaans moeilijk objectief zijn te meten. Bovendien kunnen er verschillende vakgebie-

den bij betrokken zijn die niet alleen een plaats moeten krijgen in het groepsproduct, maar ook in de beoordeling. Daar komt nog dat bij groepsproducten niet op voorhand duidelijk is welk aandeel de afzonderlijke groepsleden hebben gehad.

Er zijn vier belangrijke aandachtsgebieden bij het beoordelen van groepsproducten: het beoordelingssysteem, het betrekken van studenten bij het geven van oordelen, het beoordelen van de groep versus individuele studenten en de kwaliteit van het oordeel. Op alle genoemde gebieden zijn richtlijnen en methoden ontwikkeld die docenten in het hoger onderwijs ten dienste staan.

### 24.3 Opzetten van een beoordelingssysteem

---

Een beoordelingssysteem van groepswerk geeft antwoorden op vragen als:

- wat is het te beoordelen product en/of proces?
- wie beoordeelt?
- wie wordt beoordeeld?
- hoe vindt de beoordeling plaats?
- wanneer beoordelen?
- gevolgen van beoordeling?

#### ■ Wat is het te beoordelen product en/of proces?

Vooraf moet worden bepaald *wat* precies zal worden beoordeeld. Dat is afhankelijk van de doelen. Er is reeds gesteld dat het bij groepsproducten om complexe opdrachten gaat die alle deel moeten uitmaken van het beoordelingssysteem. Te onderscheiden zijn (zie onder meer Cluitmans, 2000):

- *concreet groepsproduct (onderscheiden naar inhoud en vorm)*: paper, project- of practicumverslag, ontwerp, uitwerking opgave;
- *taakproces*: gehanteerde methode, stappenplan om tot het product te komen, de marge tussen planning en realisatie;
- *groepsproces*: afspraken, taakverdeling, uitvoering taken, vergadertechniek, communicatieve aspecten;
- *houding/attitude*: het kunnen geven van gerichte feedback, het openstaan voor kritiek/stressbestendigheid;
- *presentatie*: mondelinge verdediging in de vorm van een betoog of groepsgesprek aan de hand van vragen.

Wanneer doelstellingen zinsneden bevatten die een of meer van bovenstaande elementen uit het onderwijsproces impliceren, dient het beoordelingssysteem ook die elementen te beslaan. Immers, een toets dient representatief te zijn voor een of meer doelstellingen (zie ► hoofdstuk 1).



### ■ Wie beoordeelt?

De meest voorkomende situatie is dat de begeleidende docent ook beoordelaar is. Deze dubbelrol is echter verre van ideaal omdat een begeleider ook een sturende rol vervult bij het tot stand komen van het eindproduct. In zekere zin beoordeelt de beoordelaar zichzelf. Het splitsen van de rol begeleider-beoordelaar zou een oplossing zijn, maar dit gebeurt te weinig in het hoger onderwijs, vooral vanwege het kostenaspect. Een ander probleem dat zich in deze situatie voordoet, is dat studenten inspelen op de persoon van de beoordelaar. Studenten kennen hun begeleider doorgaans goed, en dus ook diens stokpaardjes. Daardoor bestaat de kans dat studenten hun begeleider-beoordelaar ‘naar de mond gaan praten’. Wanneer de beoordelaar zich strikt houdt aan een beoordelingslijst (zie verder), wordt die kans kleiner.

Over het algemeen geldt: hoe meer beoordelaars, hoe meer dit ten goede komt aan de objectiviteit van het oordeel en daardoor aan de betrouwbaarheid. Maar net als het splitsen van de rol begeleider-beoordelaar is het inzetten van meer dan één beoordelaar eerder uitzondering dan regel.

Er zijn verschillende situaties en manieren om twee beoordelaars in te zetten:

- In een situatie waarin het oordeel van de begeleider-beoordelaar in de buurt van de zak/slaaggrens ligt, kan een tweede beoordelaar worden ingeschakeld. Indien deze tweede beoordelaar tevens begeleider is van een andere, vergelijkbare groep, hoeft hij of zij zich niet extra inhoudelijk te verdiepen. Dat is tijdwinst.
- De beoordelaars kunnen onderling de aspecten waarop wordt beoordeeld, verdelen, zodat de beoordelaars wel het werk van alle studenten zien, maar alleen een deel ervan.
- De beoordelaars kunnen al het werk van een deel van de studenten verdelen, waardoor de beoordelaars wel het hele product zien, maar niet alle studenten.
- De beoordelaars kunnen alles van alle studenten beoordelen en komen tot een score door de scores te middelen tot één eindcijfer voor de hand.

Als aanvulling op de beoordeling door de docent kunnen ook de studenten zelf worden ingezet bij groepsbeoordeling (*co- en peer-assessment*). Diverse methoden zijn betrouwbaar en valide gebleken, zoals: *peer nominations*, *peer ranking* en *peer rating*. In de volgende paragraaf staan hiervan enkele voorbeelden. Uit juridisch oogpunt moet de docent (examinator) de eindverantwoordelijkheid houden en de studentbeoordeling er in adviserende zin bij gebruiken.

Vooral in het kader van projectonderwijs komen groepsproducten vaak tot stand buiten de opleiding, namelijk in een bedrijf of instelling. Wanneer er sprake is van externe opdrachtgevers, is aan te raden hen in te schakelen bij het beoordelen van het groepsproduct. Zij hebben immers zicht gehad op het functioneren van de groep studenten die de opdracht heeft uitgevoerd en hebben bovendien een specifieke invalshoek om het product te beoordelen. Het is niet aan te raden de beoordeling van het groepsproduct volledig in handen te leggen van externen. De opleiding is uiteindelijk verantwoordelijk voor het niveau van de afgestudeerden. Het is wel verstandig dat externen een vorm van

inspraak hebben alvorens docenten een oordeel over het groepsproduct en de samenwerking geven.

Het komt verder ook voor dat studenten het groepsproduct moeten presenteren, aan andere docenten of aan medestudenten. Hierbij is de achterliggende gedachte dat presenteren dikwijls deel uitmaakt van de latere beroepspraktijk. De toehoorders vormen een soort forum dat in staat is in ieder geval de presentatie te beoordelen (zie ► hoofdstuk 10 voor het beoordelen van presentaties). Daarom kunnen docenten ook een dergelijk forum inschakelen om tot een beter afgewogen oordeel te komen.

Dus, de uiteindelijke beoordeling van samenwerkingsvaardigheden en het groepsproduct moet worden gedaan door een examiner/docent, maar deze kan hierbij gebruikmaken van input van:

- de begeleidende docent;
- een tweede docent, bijvoorbeeld iemand die begeleider is bij een andere groep, in gevallen rond de zak/slaaggrens;
- meer docenten, van wie één eventueel ook de begeleider is;
- in het geval een project buiten de onderwijsinstelling plaatsvindt: samen met de opdrachtgever;
- een forum van deskundigen;
- de docent en studenten (*co-assessment*);
- de medestudenten uit de groep (*peer-assessment*).

#### ■ Wie wordt beoordeeld?

Uiteindelijk moet elke student een eigen beoordeling krijgen. Vooraf moet bekend zijn of er aan elk lid van een groep één groepscijfer wordt gegeven of dat er individuele verschillen kunnen zijn. Bij twijfel over de inbreng van afzonderlijke groepsleden kan op advies van de begeleider bijvoorbeeld worden afgeweken van het groepscijfer. Dit moet wel duidelijk aantoonbaar zijn, waarbij een probleem is dat de begeleider slechts een deel van het proces ziet. In combinatie met peer-assessment wordt het beeld echter completer (zie de voorbeelden hierna).

Een andere methode is het eindcijfer te laten bestaan uit het gemiddelde van het cijfer voor het groepsproduct en het cijfer voor de individuele procesbeoordeling.

Daarnaast staat het de docent natuurlijk vrij om in aansluiting op het beoordelen van het groepsproduct studenten mee te laten doen een individuele toets over het onderwerp waarop het groepsproduct betrekking heeft.

#### ■ Hoe vindt de beoordeling plaats?

Studenten moeten vooraf weten wat ze kunnen verwachten. Telt bijvoorbeeld een tussentijdse bespreking mee? Moet ieder groepslid het totale groepsproduct kunnen verdedigen? De wijze van beoordeling van een groepsproduct heeft – net als bij andere toetsvormen – een sterk sturende werking op het studiegedrag, en in dit geval ook op het groepsgedrag van studenten. Het hanteren van beoordelingsvoorschriften of het werken met rubrics (zie ► hoofdstuk 17) die vooraf bekend zijn bij studenten, geeft duidelijkheid. Een dergelijk

voorschrift bevat verschillende beoordelingsmethoden die zich van elkaar onderscheiden in geschiktheid voor het geven van feedback aan studenten:

- *globaal/holistisch*: de beoordeling vindt plaats op basis van een algemene indruk of met een korte aandachtspuntenlijst als ondersteuning;
- *op detailniveau/analytisch*: de beoordeling geschiedt met behulp van gedetailleerde beoordelingsdimensies;
- *kwalitatief*: de docent identificeert sterke en zwakke punten die vervolgens de groep in staat stellen te werken aan hun zwakke punten;
- *kwantitatief*: het (tussentijdse) resultaat wordt in een cijfer uitgedrukt; deze beoordelingen vormen in de regel de basis voor eindbeoordelingen waarbij er beslissingen over studenten worden genomen, bijvoorbeeld slagen of zakken.

Analytische en kwalitatieve beoordelingsmethoden zijn het geschiktst voor het geven van feedback.

#### ■ Wanneer beoordelen?

Doorgaans vindt beoordeling plaats na afloop van, maar ook wel tijdens de onderwijsactiviteiten. Soms is er al een beoordeling voordat studenten gaan deelnemen aan het onderwijs om het beginniveau van studenten vast te stellen. Dit voorkomt een ‘valse start’ van (een van) de groepsleden. Maar een beoordeling vooraf kan ook als selectiemiddel dienen dat studenten opspoort die in onvoldoende mate beschikken over noodzakelijk geachte kennis en vaardigheden. Studenten mogen bijvoorbeeld pas aan een groepsproject deelnemen nadat zij voor vak x en y een voldoende hebben gehaald.

Keuzemogelijkheden voor beoordeling zijn dus:

- *bij de start van het project*, met als doel selectie of vaststelling van het beginniveau;
- *tussentijds*, met als doel controle op de voortgang (formatief) en bijsturing;
- *aan het eind*, met als doel afsluiting van het studieonderdeel (summatief).

#### ■ Gevolgen van beoordeling?

Er zijn diverse consequenties aan het beoordelingsresultaat van groepsopdrachten te verbinden. Bij een flinke onvoldoende is het duidelijk dat de hele groep er niet aan ontkomt het project over te doen of anderszins met extra studie-inspanningen te voldoen aan de verplichtingen. Als het resultaat net onvoldoende is, zou door de groep of door een enkele student op korte termijn een aanvullende taak kunnen worden gemaakt. Een aanvullende taak kan desgewenst ook worden gebruikt om een sterke student in een zwakke groep zich alsnog naar boven toe te kunnen laten onderscheiden – bijvoorbeeld omdat deze student in aanmerking wil komen om ‘cum laude’ af te studeren. Ook is het denkbaar dat een student die binnen de groep duidelijk aantoonbaar onder de maat heeft gepresteerd, het project in een volgende groep moet overdoen.

Mogelijke consequenties van de beoordeling zijn:

- *hele groep geslaagd*: alle studenten krijgen een voldoende;

- *bijna geslaagd, twijfelachtig*: aanvulling vereist door de hele groep of door individuele studenten;
- *niet geslaagd*: de hele groep of enkele leden uit de groep moeten het project overdoen.

Nadat aldus een beoordelingssysteem is opgezet dat in grote lijnen de wijze van beoordeling schetst, de beoordelaars benoemt, het beoordelingsobject aangeeft, alsmede het tijdstip en de consequenties, kan de feitelijke beoordeling beginnen.

## 24.4 Methoden voor beoordeling: het betrekken van medestudenten

Hierna staat een aantal methoden voor beoordeling van samenwerking en van groepsproducten genoemd waarbij ofwel de docent of beoordelaar alleen betrokken is, ofwel de docent samen met studenten. Soms gaat het slechts om een kleine variatie in een methode. Vaak is het van belang om niet alleen het groepsproduct maar ook het proces en individuele studenten erbij te betrekken, bijvoorbeeld wanneer teambuilding tot de doelstellingen behoort, of de mate van reflectie van de studenten op het groepsproces.

### ■ Beoordeling door de docent

Een docent kan op diverse manieren een cijfer of waardering toekennen aan een groepsproduct en eventueel aan de bijdrage van de afzonderlijke groepsleden. Mogelijke manieren zijn:

- Elke student krijgt hetzelfde groeps cijfer dat voor het groepsproduct is gegeven.
- Elke student krijgt als basis het groeps cijfer dat voor het groepsproduct is gegeven, maar hieraan wordt een individuele beoordeling gekoppeld, die voldoende moet zijn. Deze methode gaat ‘meeliftgedrag’ tegen. Het ontdekken van meeliften is mogelijk:
  - met behulp van observaties van de inbreng van studenten door de begeleider tijdens de groepsbijeenkomsten;
  - door individueel gerichte vragen te stellen bij het groeps gesprek of de groeps presentatie;
  - door tijdens de looptijd van het project de individuele bijdrage aan de voortgang te observeren; als bijvoorbeeld gebruik wordt gemaakt van een digitale leeromgeving, kan de begeleider op de computer zien in hoeverre elke student regelmatig tussenproducten heeft aangeleverd.
- Elke student krijgt een individueel cijfer dat tot stand is gekomen uit een combinatie van het groeps cijfer voor het groepsproduct en van een individueel cijfer voor de taken die door groepsleden zijn uitgevoerd. Middeling van deze cijfers ligt voor de hand, al zijn andere methoden, waarbij de docent een bepaald gewicht toekent aan een deeltaak, ook toepasbaar. De docent moet in dit geval wel duidelijk onderscheid maken tussen wie welke taken heeft uitgevoerd, bijvoorbeeld door gebruik te maken van een logboek dat door de groep is bijgehouden.

### ■ Medebeoordeling door de studenten

Het is mogelijk om studenten in te schakelen voor het beoordelen van de individuele bijdragen aan een groepsproduct en/of het beoordelen van samenwerkingsvaardigheid. Die methoden vallen onder een van de volgende noemers: *co-*, *peer-* en *self-assessment*. *Co-assessment* houdt in dat de studenten en de docent(en) gezamenlijk beoordelen, waarbij de docent verantwoordelijk blijft voor het eindoordeel. Bij *peer-* en *self-assessment* worden de groepsleden betrokken bij de beoordeling van de prestaties van elkaar en bij voorkeur ook van zichzelf. Uit onderwijskundig oogpunt is belangrijk dat hiermee tevens de academische vaardigheid wordt getraind van het kunnen beoordelen van collega's en jezelf (zie Van den Berg, Pilot & Admiraal, 2005). Deze vormen van beoordelen zijn mogelijk door bijvoorbeeld van studenten het volgende te vragen:

- Noem een of meer studenten van je groep, inclusief jezelf, die het best hebben gepresteerd (*peer nominations*).
- Zet je groepsleden en jezelf in een rangorde aan de hand van ieders prestatie (*peer ranking*).
- Beoordeel elk groepslid, ook jezelf, op een absolute schaal (*peer rating*).

Er zijn ook andere methoden die het midden houden tussen de twee hiervoor genoemde:

- Het scheiden van het beoordelen van samenwerking/proces en het product, waarbij het product door de docent wordt beoordeeld – eventueel met behulp van studenten uit andere groepen (*co-assessment*) – en het proces door middel van *peer-* en *self-assessment*.
- Het scheiden van groepstaken en individuele taken, waarbij het individuele deel door de docent wordt beoordeeld en de groepstaken door middel van *peer-* en *self-assessment*. Het eindcijfer is dan afhankelijk van de cijfers van de onderdelen. Het is uiteraard mogelijk een verschillend gewicht toe te kennen aan de onderdelen om tot een eindcijfer te komen.
- De docent geeft een cijfer aan de hele groep en vermenigvuldigt dit met het aantal studenten. De studenten verdelen vervolgens onderling de uitkomst van de vermenigvuldiging. De docent kan nadere restricties hiervoor geven. Een voorbeeld: het cijfer voor het product is een '8' en de groep bestaat uit zeven studenten. De vermenigvuldiging levert 56 punten op die de zeven studenten onderling mogen verdelen. Groepsleden die beter of slechter hebben gepresteerd in de ogen van de groep, verkrijgen hierdoor een hogere of lagere waardering. De restrictie van de docent kan luiden: geen enkel lid mag een cijfer krijgen lager dan een '5' en hoger dan een '9'.

## 24.5 Uitwerking van enkele beoordelingsmethoden

In de volgende voorbeelden worden drie beoordelingsmethoden nader uitgewerkt.

### Voorbeeld 1

Met de checklist van Goldfinch (zie Lejk, Wyvill & Farrow, 1996) beoordeelt iedere student zowel zichzelf (*self-assessment*) als de andere groepsleden (*peer-assessment*) op procesaspecten. Het toekennen van punten wordt gerelateerd aan de bijdrage van zichzelf ten opzichte van de andere groepsleden op het betreffende aspect. Punten op de criteria hebben de volgende betekenis:

- 3 : beter dan de meeste andere groepsleden
- 2 : ongeveer hetzelfde als de andere groepsleden
- 1 : niet zo goed als de anderen
- 0 : geen enkele nuttige bijdrage
- 1 : storend voor de groep

In de tabel staan peer- en self-assessment gericht op procesaspecten, conform Goldfinch.

Schrijf de namen van de groepsleden naast jezelf en geef ieder een punt	Jezelf	Student a	Student b	Student c
enthousiasme/participatie	2	2	2	0
met ideeën komen	3	3	2	1
begrijpen wat vereist wordt	2	3	2	2
bijdrage aan functioneren als team	2	3	2	-1
organiseren/controleren van de groep	1	3	1	0
efficiënt taken uitvoeren	2	2	2	0
Totaal aantal punten:	12	16	11	2

Het eindcijfer van het groepsproduct kan in dit voorbeeld voor de afzonderlijke groepsleden als volgt worden bepaald: elk groepslid dat van iedereen gemiddeld twaalf punten krijgt, krijgt het groepsresultaat voor het eindproduct als eindcijfer. Voor studenten die er op bepaalde procesaspecten uitspringen, kan de docent het groepsresultaat naar boven (in het voorbeeld student a) of naar beneden (student c) bijstellen op grond van de oordelen van de andere groepsleden. De docent zou zelfs kunnen overwegen om student c het project over te laten doen in een andere groep, bijvoorbeeld wanneer het kunnen samenwerken in een groep een belangrijke doelstelling is.

### Voorbeeld 2

Als het vooral om taakaspecten gaat, kan gebruik worden gemaakt van de checklist van Conway e.a. (zie Lejk, Wyvill & Farrow, 1996) waarbij studenten alleen hun medegroepsleden beoordelen. Punten op de criteria hebben de volgende betekenis:

- 1 : geen enkele nuttige bijdrage
- 2 : goed geprobeerd, zonder succes
- 3 : gemiddeld
- 4 : bovengemiddeld
- 5 : bijzonder goed

In de tabel staat peer-assessment gericht op taakaspecten en presentatie, conform Conway.

Schrijf de namen van de groepsleden hiernaast en geef ieder een punt	Student a	Student b	Student c	Student d
zoeken naar literatuur	3	1	5	3
analyseren literatuur	3	1	5	3
rapport schrijven	3	2	3	3
groepspresentatie	3	2	4	3
Totaal aantal punten:	12	6	17	12

Bij afwijking van het gemiddelde kan de docent ook in dit voorbeeld besluiten om het eindcijfer van het groepsproduct omhoog of omlaag bij te stellen op grond van de oordelen van de andere groepsleden. Als de verschillen vrij groot zijn, zoals voor student b, kan worden overwogen om deze student een aanvullende taak op te dragen om daarmee alsnog te voldoen aan de doelstellingen.

### Voorbeeld 3

De verdeling van de punten voor een individueel eindcijfer volgens Habeshaw (zie Lejk, Wyvill & Farrow, 1996), gaat als volgt. Iedere studenten uit een groep van drie mag gemiddeld zestig punten verdelen over de leden van de groep, inclusief zichzelf. De toekenning van punten gebeurt door de studenten zelf, in onderling overleg met de medestudenten. Indien de groep van mening is dat iedereen goed heeft gewerkt en dat er weinig onderlinge verschillen waren in de bijdrage aan het groepsproduct, zullen zij ieder hetzelfde aantal punten en dus hetzelfde eindcijfer geven. Bij afwijkingen, zoals in dit voorbeeld, kan de waardering voor ieders bijdrage verschillend zijn.

In de tabel staat peer-assessment gericht op een totaal oordeel, conform Habeshaw.

Voorbeeld van toedeling punten aan elkaar	Student a	Student b	Student c
student a geeft	80	40	60
student b geeft	60	60	60
student c geeft	70	50	60
Totaal aantal punten:	210:3=70	150:3=50	180:3=60

## 24.6 Kwaliteitseisen aan de beoordeling

Ook het beoordelen van samenwerking/groepsproces en groepsproducten moet zo betrouwbaar en valide mogelijk plaatsvinden (zie ook [tabel 24.1](#)). Voor het procesdeel is het hanteren van rubrics, checklists of een logboek een handig hulpmiddel. Een zorgvuldige formulering van de rubrics en checklists en/of het hanteren van de juiste kapstukken in het logboek, verhoogt de representativiteit van het oordeel en daarmee de inhoudsvaliditeit. Voor het beoordelen van het product zijn meer mogelijkheden om tegemoet te komen aan

■ **Tabel 24.1** Checklist voor het beoordelen van groepsprocessen en groepsproducten.

<p><b>Wat wordt beoordeeld?</b></p> <ul style="list-style-type: none"> <li>– concreet groepsproduct (inhoud en vorm)</li> <li>– taakproces</li> <li>– groepsproces</li> <li>– houding/attitude</li> <li>– presentatie</li> </ul>	<p><i>Tips voor een hogere validiteit:</i> Laat studenten een logboek schrijven. Hanteer beoordelingslijsten die representatief zijn voor:</p> <ul style="list-style-type: none"> <li>– de doelstellingen</li> <li>– de afzonderlijke vakgebieden</li> <li>– het beoogde cognitieve niveau</li> </ul> <p><i>Tips voor een hogere betrouwbaarheid:</i></p> <ul style="list-style-type: none"> <li>– schakel bij voorkeur twee beoordelaars in</li> <li>– laat bij twijfel over een voldoende/onvoldoende een tweede beoordelaar het werk nogmaals beoordelen en middel de scores</li> <li>– voorkom een dubbelrol van begeleider en beoordelaar</li> <li>– hanteer een korte beoordelingslijst</li> <li>– maak de procedure inzichtelijk</li> <li>– zorg ervoor dat de criteria vooraf bekend en duidelijk zijn voor studenten</li> </ul>
<p><b>Wie beoordeelt?</b></p> <ul style="list-style-type: none"> <li>– begeleidend docent</li> <li>– tweede docent</li> <li>– meer docenten (onder wie eventueel de begeleider)</li> <li>– opdrachtgever(s)</li> <li>– forum van deskundigen</li> <li>– docent en studenten (<i>co-assessment</i>)</li> <li>– medestudenten (<i>peer-assessment</i>)</li> </ul>	
<p><b>Wie wordt beoordeeld?</b></p> <ul style="list-style-type: none"> <li>– groep</li> <li>– individu binnen de groep</li> <li>– combinatie groeps- en individueel aspect</li> </ul>	
<p><b>Hoe vindt de beoordeling plaats?</b></p> <ul style="list-style-type: none"> <li>– globaal</li> <li>– op detailniveau/analytisch</li> <li>– kwalitatief</li> <li>– kwantitatief</li> </ul>	
<p><b>Wanneer beoordelen?</b></p> <ul style="list-style-type: none"> <li>– bij de start van het project</li> <li>– tussentijds</li> <li>– aan het eind</li> </ul>	
<p><b>Gevolgen van beoordeling</b></p> <ul style="list-style-type: none"> <li>– hele groep geslaagd</li> <li>– bijna geslaagd, twijfelachtig: aanvulling (hele groep of individuele student)</li> <li>– niet geslaagd: overdoen (hele groep of individuele student)</li> </ul>	

de validiteitseis. Om validiteit te bereiken, moet de beoordeling van het groepsproduct representatief zijn voor de doelstellingen. Wanneer daarbij sprake is van informatie uit diverse vakgebieden, zal dit in het groepsproduct tot uiting moeten komen. Wanneer er bij complexe groepsproducten verschillende vakgebieden betrokken zijn, verhoogt het de validiteit wanneer de beoordeling ook onderscheid maakt tussen die verschillende vakgebieden.

Wanneer bij het beoordelen van groepsproducten meer dan één beoordelaar is betrokken, geldt dat de mate van overeenstemming tussen verschillende beoordelaars hoog moet zijn. Als dat het geval is, is het oordeel betrouwbaar. Als er geen eenduidige beoordelingscriteria zijn, verlaagt dit over het algemeen de overeenstemming tussen de beoordelaars. Maar ook wanneer er slechts een enkele beoordelaar is, zijn onduidelijke criteria onwenselijk omdat de beoordelaar zal moeten gaan interpreteren. Dit verhoogt de subjectiviteit en verlaagt dus de objectiviteit.



Ook bij *peer-assessment* moet worden gestreefd naar een betrouwbare beoordeling. Als een groot aantal studenten in groepjes is ingedeeld, is onderlinge vergelijking tussen groepen moeilijker omdat elke groep verschillend is van samenstelling, strengheid, en dergelijke. Daardoor kunnen beoordelareffecten optreden (zie de bijlage achterin het boek). Wanneer verschillende groepsproducten door dezelfde studentbeoordelaars worden beoordeeld of wanneer verschillende studentbeoordelaars telkens hetzelfde deel van het groepsproduct nakijken, gaat de betrouwbaarheid waarschijnlijk omhoog. De objectiviteit is soms discutabel en afhankelijk van de opstelling van de beoordelaars. Het komt regelmatig voor dat groepen die een en hetzelfde soort opgave of verslag hebben gemaakt, van de ene beoordelaar een hoger cijfer krijgen dan van een andere beoordelaar, ondanks een lijst met beoordelingscriteria. Dit is te verhelpen door via een steekproef hetzelfde verslag door twee beoordelaars te laten bekijken. Mochten hier grote verschillen optreden, dan kan alsnog worden besloten de andere producten ook door twee beoordelaars te laten beoordelen en de scores vervolgens te middelen.

Verder kan aan de eis van transparantie worden voldaan door te zorgen voor een duidelijke, eerlijke en inzichtelijke procedure. Door de studenten zelf te betrekken bij de beoordeling – van met name de procesaspecten – worden ze medeverantwoordelijk voor hun eigen leren. Dit vergt wel een duidelijke instructie vooraf en eventueel een training in het geven van feedback. De eindverantwoordelijkheid hoort altijd bij de verantwoordelijke docent te liggen en niet bij de student(en).

# Beoordelaarseffecten

## Inleiding

In diverse hoofdstukken van dit boek komen verschillen tussen beoordelaars ter sprake. In algemene zin zijn verschillen tussen beoordelaars die hetzelfde werk beoordelen niet wenselijk. De verschillen zorgen voor ongewenste effecten op de score van de student. De persoon van de beoordelaar mag immers, zo is de algemene opvatting, bij het geven van scores geen rol spelen. Toch is in de meeste gevallen niet uit te sluiten dat de persoon van de beoordelaar het uiteindelijk oordeel ongewenst deels heeft beïnvloed. Dat zal ertoe leiden dat de beoordeling minder betrouwbaar wordt. Er is veel onderzoek verricht naar deze ongewenste effecten. In zijn standaardwerk, *Methodologie* (1961), benoemt en definieert De Groot deze effecten. Het onderstaande is daaraan deels ontleend.

### ■ Contaminatie-effect

Het contaminatie-effect is het effect dat optreedt bij beoordelen wanneer de vrijheid in de beoordeling, onwillekeurig of willekeurig, wordt gebruikt voor andere doeleinden dan die van een onbevagen, onbevooroordeelde beoordeling. Dit is bijvoorbeeld het geval als beoordelaars lagere scores geven om aan te tonen dat hun vak moeilijk is.

### ■ halo-effect

Het halo-effect treedt op wanneer beoordelaars hun oordeel mede laten beïnvloeden door andere prestaties van de student dan geuit op de te beoordelen prestatie. De beoordelaars hebben in dat geval de neiging om een 'goede' student bij een wat mindere prestatie wat hoger te beoordelen dan de prestatie van de student rechtvaardigt. De prestatie wordt overgewaardeerd.

### ■ Horn-effect

Het horn-effect is het tegenovergestelde van het halo-effect. Het horn-effect treedt op wanneer beoordelaars hun oordeel mede laten beïnvloeden door andere prestaties van de student dan geuit op de te beoordelen prestatie. De beoordelaars hebben in dat geval de neiging om een 'slechte' student bij een wat hogere prestatie wat lager te beoordelen dan de prestatie van de student rechtvaardigt. De prestatie wordt ondergewaardeerd.

### ■ Normverschuiving

De normverschuiving is het effect dat optreedt wanneer een beoordelaar zich aanpast aan de geleverde prestaties van studenten. Beoordelaars kunnen bijvoorbeeld minder streng worden als na een aantal beoordelingen blijkt dat de meeste studenten eenzelfde vraag fout beantwoorden.

### ■ Restriction-of-range

Restriction-of-range wordt veroorzaakt door de vrijheid in de beoordelingstaak die willekeurig of onwillekeurig leidt tot bepaalde verdelingen van de beoordelingen die algemeen menselijke of persoonlijke neigingen uitdrukken. Zo zal de ene beoordelaar wel alle schaalwaarden (scores van 1 tot 10) gebruiken, terwijl een andere beoordelaar steeds waarden zal toekennen in de buurt van het midden.

- **Sequentie-effect**

Het sequentie-effect treedt op wanneer een beoordelaar de beoordeling ten onrechte mede laat leiden door een of meer voorafgaande beoordelingen. Bijvoorbeeld, beoordelaars kijken toetsvragen in een bepaalde volgorde na. Een groot aantal slechte antwoorden dat wordt gevolgd door een correct antwoord, kan er dan toe leiden dat het correcte antwoord onevenredig hoog wordt beoordeeld.

- **Signifisch effect**

Het significans effect treedt op wanneer beoordelaars letten op verschillende aspecten, of als ze de aspecten verschillend wegen in hun oordeel. Dit effect treedt bijvoorbeeld op wanneer beoordelaars schrijfproducten beoordelen. De ene beoordelaar let vooral op de grammaticale (on)juistheden en de andere op structuur of inhoud.

# De redacteuren

#### ■ **Anneke Bax**

Drs. A.E. Bax (1951) studeerde Andragologie aan de Vrije Universiteit te Amsterdam. Vanaf 1979 was zij – telkens als onderwijs(beleids)medewerker – werkzaam bij het Centraal Bureau voor Onderzoek en Wetenschappelijk Onderwijs van de Academische Raad te 's-Gravenhage (tot 1985), de VSNU te Utrecht (tot 1986), de provincie Limburg te Maastricht (tot 1989), de faculteit Technologie Management van de TU Eindhoven (tot 2000) en het Onderwijscentrum van de Vrije Universiteit te Amsterdam (tot 2004). Vervolgens werkte ze bij Hogeschool Zuyd als toetsdeskundige bij het Programma Grensverleggend Onderwijs (tot 2008), als kwaliteitszorgmedewerker bij de faculteit Onderwijs (tot 2011) en als zelfstandige aan diverse tijdelijke projecten binnen Hogeschool Zuyd te Maastricht, Heerlen en Sittard. Zij heeft enkele boeken op het gebied van toetsen en kwaliteitszorg op haar naam staan en is redactielid geweest van *Onderzoek van Onderwijs*, *de Hoger Onderwijs Reeks* en *Tijdschrift voor Hoger Onderwijs*.

#### ■ **Henk van Berkel**

Dr. H.J.M. van Berkel (1948) studeerde Psychologie aan de Universiteit van Amsterdam. In 1984 promoveerde hij op een toetstechnisch onderwerp. Na een aantal jaren werkzaam te zijn geweest bij het Centrum voor Onderzoek van het Wetenschappelijk Onderwijs (COWO) werd hij in 1986 coördinator van de Eerstegraads lerarenopleiding aan de Universiteit Maastricht. Hij is lid en voorzitter geweest van een zevental visitatiecommissies. Hij heeft verschillende boeken op zijn naam staan op het gebied van toetsen, kwaliteitszorg en probleemgestuurd onderwijs. Hij is auteur van tientallen artikelen op deze drie terreinen. Hij is oprichter en (tot 1 juli 2013) hoofdredacteur van *EXAMENS – Tijdschrift voor de toetspraktijk*, oprichter en editor van *Advances in Health Sciences Education* en redactielid van de *Hoger Onderwijs Reeks*. Daarnaast is hij voorzitter van de examencommissie Gezondheidswetenschappen. Zie ook: ► <http://www.personeel.unimaas.nl/h.vanberkel/>.

#### ■ **Desirée Joosten-ten Brinke**

Dr. Desirée Joosten-ten Brinke (1967) studeerde Toegepaste Onderwijskunde (meetmethoden en data-analyse) aan de Universiteit Twente. In 2008 promoveerde zij op Assessment of Prior Learning (EVC). Zij is universitair hoofddocent bij het Centre for Learning Sciences and Technologies (CELSTEC) van de Open universiteit en lector Eigentijds toetsen en beoordelen aan de Fontys lerarenopleiding in Tilburg. Haar onderzoek en onderwijs richten zich op het professionaliseren van docenten ten behoeve van de kwaliteit van toetsen en beoordelen. Onderwerpen als formatief en summatief toetsen, EVC, digitaal toetsen staan daarbij centraal. De context van haar werk betreft primair het hoger onderwijs in verschillende inhoudelijke domeinen. Vanaf 1 juli 2013 is zij hoofdredacteur van *EXAMENS – Tijdschrift voor de toetspraktijk*.

# De auteurs

**Liesbeth Baartman**

Hogeschool Utrecht  
liesbeth.baartman@hu.nl

**Anneke Bax**

baxanneke@hotmail.com

**Niek van Benthum**

Wageningen University  
Stoas Wageningen, Vilentum Hogeschool  
niek.vanbenthum@wur.nl  
n.van.benthum@stoasvilentum.nl

**Henk van Berkel**

Universiteit Maastricht  
h.vanberkel@maastrichtuniversity.nl

**Jo Boon**

Open Universiteit  
jo.boon@ou.nl

**Paul van den Bos**

Vrije Universiteit Amsterdam  
p.j.p.vanden.bos@vu.nl

**Patris van Boxel**

Vrije Universiteit Amsterdam  
p.van.boxel@vu.nl

**Marcel van Brunschot**

Stichting Leerstation Zorg  
m.v.brunschot@leerstationzorg.nl

**Carel Burghout**

Fontys lerarenopleiding Tilburg  
c.burghout@fontys.nl

**Silvester Draaijer**

Vrije Universiteit Amsterdam  
s.draaijer@vu.nl

**Tom Erkens**

Cito  
tom.erkens@cito.nl



**Titus Geerligs**

ctg@home.nl

**Erik de Graaff**

Aalborg Universitet, Denemarken

degraaff@plan.aau.dk

**Judith Gulikers**

Wageningen University

judith.gulikers@wur.nl

**Desirée Joosten-ten Brinke**

Open Universiteit

Fontys lerarenopleiding Tilburg

desiree.joosten-tenbrinke@ou.nl

d.tenbrinke@fontys.nl

**Ruud Klarus**

Hogeschool van Arnhem en Nijmegen

ruud.klarus@han.nl

**Marcel van der Klink**

Open Universiteit

Zuyd Hogeschool

marcel.vanderklink@ou.nl

marcel.vanderklink@zuyd.nl

**Raymond Kloppenburg**

Hogeschool Utrecht

raymond.kloppenburg@hu.nl

**Irma Kokx**

Universiteit Maastricht

i.kokx@maastrichtuniversity.nl

**Quinta Kools**

Fontys lerarenopleiding Tilburg

q.kools@fontys.nl

**Wim Lansu**

Fontys lerarenopleiding Tilburg

w.lansu@fontys.nl

**Arnoud van Leuven**

Vereniging van hogescholen  
leuven@hbo-raad.nl

**Riet Martens**

Universiteit Twente  
h.l.m.martens@utwente.nl

**Ilona Mathijssen**

Fontys lerarenopleiding Tilburg  
i.mathijssen@fontys.nl

**George Moerkerke**

Open Universiteit  
george.moerkerke@ou.nl

**Arno Muijtjens**

Universiteit Maastricht  
a.muijtjens@maastrichtuniversity.nl

**Jacob Perrenet**

Technische Universiteit Eindhoven  
j.c.perrenet@tue.nl

**Albert Pilot**

Universiteit Utrecht  
a.pilot@uu.nl

**Frans Prins**

Universiteit Utrecht  
f.j.prins@uu.nl

**Sjoerd Romme**

Technische Universiteit Eindhoven  
a.g.l.romme@tue.nl

**Henk Schmidt**

Erasmus Universiteit Rotterdam  
schmidt@fsw.eur.nl

**Lambert Schuwirth**

Flinders University, Adelaide, Australië  
lambert.schuwirth@flinders.edu.au

**Dominique Sluijsmans**

Zuyd Hogeschool

dominique.sluijsmans@zuyd.nl

**Gerard Straetmans**

Saxion

Cito

g.j.j.m.straetmans@saxion.nl

**Wynand Wijnen**

Universiteit Maastricht

(overleden op 21-11-2012)

# Lexicon

### Absoluut normeren

Bij het geven van cijfers volgens een absolute methode, houden docenten geen rekening met de score van de overige studenten. De toetscore wordt vergeleken met een norm die vooraf is bepaald. In bijna alle gevallen zal de norm bestaan uit een percentage van de maximaal te behalen score. Bijvoorbeeld, studenten die tachtig procent van de maximale score behalen krijgen het cijfer '8', of: indien studenten veertig van de zestig driekeuzevragen correct beantwoorden krijgen ze een '6'. Veel docenten hanteren deze methode van cijfergeven. Het feit dat een toets onverwacht moeilijk, of gemakkelijk, kan uitvallen, is het belangrijkste argument tegen een absolute wijze van het bepalen van de zak/slaaggrens.

### Afleider

Een afleider is een alternatief in een meerkeuzevraag dat niet correct is.

### Afstudeerkring

Een afstudeerkring is een instrument voor het begeleiden en beoordelen van afstudeerwerkstukken (thesis) dat vooral wordt gebruikt in situaties waarin een aantal studenten een werkstuk schrijft over een gemeenschappelijk thema. Een afstudeerkring bestaat uit een of meer docent-begeleiders en vijf tot vijftien studenten in de afstudeerfase. Als zodanig vormt een afstudeerkring een permanente begeleidingsgroep waarin studenten in principe op elk moment kunnen instromen, en na hun afstuderen weer vertrekken. Het doel van deze manier van werken is het vergroten van zowel de kwaliteit en doelmatigheid van de begeleidingswerkzaamheden als de kwaliteit van de werkstukken. De afstudeerkring stimuleert daarnaast samenwerkend leren.

### Alternatieven

Een gesloten vraag bestaat uit twee delen: de stam en de alternatieven. De stam is het gedeelte van de vraag dat het probleem bevat. De alternatieven zijn de antwoordmogelijkheden waaruit studenten kunnen kiezen. De niet-correcte alternatieven worden afleiders genoemd; het correcte alternatief is de sleutel.

### Antwoordmodel

Zie correctievoorschrift

### Arbeidsproef

Een arbeidsproef bestaat uit de opdracht om een aantal taken uit te voeren in een echte arbeidssituatie. Beoordelaars observeren de studenten. De taken moeten representatief zijn voor wat de toets beoogt te meten. Stimuli lokken een bepaalde deelvaardigheid uit. Het is vooraf bekend welke deelvaardigheden zullen worden uitgelokt, waardoor de beoordelaars zich op de beoordelingstaak kunnen voorbereiden. Voorbeelden van dit type toetsen zijn het geven van een les aan het eind van de opleiding tot leraar en het uitvoeren van diagnostisch onderzoek of van behandeling van een 'echte' patiënt in de geneeskunde of tandheekkunde.

### Assessment center

Een assessment center bestaat uit een serie gestandaardiseerde opdrachten die kandidaten moeten uitvoeren. De opdrachten vragen om een reactie van de student in termen van waarneembaar gedrag. Bijvoorbeeld: deelnemen aan een groepsdiscussie, reageren op een gespeelde conflictsituatie in het werk en het afhandelen van een stapel post. Het gevraagde gedrag sluit nauw aan bij de functie of werksituatie waarvoor wordt opgeleid. Het gedrag wordt geobserveerd door beoordelaars.

### Beantwoordingstijd

De beantwoordingstijd van een toetsvraag is het maximaal aantal seconden dat voor het overgrote deel van de toetsdeelnemers voldoende is om de toetsvraag te kunnen beantwoorden. Deze tijd kan variëren tussen gering voor een gesloten vraag met twee antwoordmogelijkheden tot wel een uur voor het antwoord op een open vraag waar studenten, bijvoorbeeld, een standpunt moeten verdedigen.

### Beoordelaarsbetrouwbaarheid

De beoordelaarsbetrouwbaarheid geeft de mate aan waarin beoordelaars consistent beoordelen. Wanneer beoordelaars hun eigen beoordeling op een ander tijdstip nog een keer overdoen, spreekt men van intrabeoordelaarsbetrouwbaarheid. De overeenstemming tussen verschillende beoordelaars over hetzelfde antwoord, product en dergelijke, wordt aangeduid met interbeoordelaarsbetrouwbaarheid.

### Betrouwbaarheid

De betrouwbaarheid van een toets geeft aan in hoeverre er vertrouwen kan zijn in de toets-score als meting, ongeacht de inhoud van de toets. Meet de toets echt iets of kan er net zo goed worden gedobeld om uitspraken te doen over studenten? Als een weegschaal de ene keer 85 kilo aangeeft en vijf minuten later, met dezelfde persoon erop, 90, is de weegschaal niet betrouwbaar. Zo ook met toetsen. Als na een tweede afname, direct na de eerste, blijkt dat een vergelijkbare toets bij dezelfde deelnemers leidt tot geheel andere toetscores, is de toets niet betrouwbaar. De correlatie tussen de scores op de beide toetsen is een aanduiding voor de betrouwbaarheid van de toets, de zogenoemde test-hertestbetrouwbaarheid. Deze correlatie is een belangrijke indicatie voor de kwaliteit van de toets als geheel. Omdat een hertest in het onderwijs niet voorkomt, zijn alternatieven bedacht om een indicatie van de betrouwbaarheid te verkrijgen met behulp van één toets. De bekendste is coëfficiënt alfa (Cronbach's alpha) die waarden kan aannemen tussen '0' en '1'.

### Betrouwbaarheid versus validiteit

De relatie tussen betrouwbaarheid en validiteit is vergelijkbaar met de relatie tussen een fundering en het huis dat erop staat. Een toeschouwer ziet alleen datgene wat zichtbaar is, de toets en de vragen. De fundering (de betrouwbaarheid) en alles wat daarop van invloed is, zoals materiaal, waterstand, diepte, stevigheid, blijft verborgen. Slechts na een bodemonderzoek (toetsanalyse) zal de kwaliteit van de fundering blijken. Als de ondergrond drijfzand is (de toets is onbetrouwbaar), heeft het geen zin een huis (de toets) te bouwen.

### Blauwdruk

Een blauwdruk is in de simpelste vorm een lijst met onderwerpen die in de toets aan de orde moeten komen. Deze onderwerpen zijn doorgaans uit de doelen af te leiden. Een toets moet representatief zijn ten opzichte van de doelen van het studieonderdeel. De blauwdruk is een hulpmiddel om representativiteit en evenwichtigheid te bereiken. Zie ook: specificatietabel/toetsmatrijs

### Casus

Een casus is een vorm van een toetsvraag die een gevalbeschrijving bevat waarover studenten vragen moeten beantwoorden. In medische opleidingen bestaat een casus veelal uit een patiënt-doktersituatie, bij lerarenopleidingen uit een leerling-docentsituatie.

### Cijfergeven

Cijfergeven is het proces waar de verantwoordelijke een cijfer toekent aan een toetsprestatie. Cijfers symboliseren een bepaalde waardering. In Nederland is waardering op een 10-puntschaal gebruikelijk. In plaats van cijfers worden ook wel letterscores gegeven, zoals onvoldoende (O), voldoende (V) en goed (G).

### Classificatie van doelen

Doelen in het onderwijs laten zich in drie categorieën classificeren: cognitieve doelen (alles wat te maken heeft met verstandelijk functioneren), affectieve doelen (alles wat te maken heeft met emoties en andere gevoelsaspecten, met afkeuren en voorkeuren, maar ook met houdingen (attitudes)) en psychomotorische doelen (alles wat betrekking heeft op motorische handelingen).

### College van Beroep voor de Examens

Studenten kunnen in beroep gaan tegen iedere beslissing van de examencommissie die hen raakt. Aan iedere instelling van het hoger onderwijs functioneert een College van Beroep voor de Examens (CBE of CoBex) voor klachten van studenten. Het CBE behandelt klachten niet inhoudelijk. Het zal onderzoeken of de examencommissie in redelijkheid tot de gewraakte beslissing is gekomen. Is dat oordeel negatief, dan vernietigt het CBE het besluit van de examencommissie en draagt zij de examencommissie op een herzien besluit te nemen, rekening houdend met het gestelde in de uitspraak van het CBE.

### Competentie

Het begrip 'competentie' is een aanduiding voor een verzameling bekwaamheden (kennis, vaardigheid en attitude) die, in een bepaalde verhouding, noodzakelijk is om adequaat te kunnen handelen of oordelen. In vrijwel alle beroepen is competentie een sleutelbegrip. Omdat 'competentie' een mengeling is, is het lastig toetsbaar.

Het toetsen van competenties kan met allerlei toetsvormen plaatsvinden: performance assessment, self-assessment, peer-assessment, reflectieverslagen en/of portfolio's. Een competentietoets dient een directe beoordeling van gedrag te bevatten, zoals een directe ob-

servatie of observatie via video. Bij competentietoetsing moeten studenten in dit gedrag integratie (van kennis, vaardigheden en houdingen) laten zien. Dit maakt competentietoetsing méér dan een vaardigheidstoets.

### Compromismethode

De compromismethode bij geven van cijfers houdt het midden tussen een relatieve en absolute methode. Een docent die deze methode hanteert, stelt bijvoorbeeld dat een student in principe zestig procent van de maximale score dient te behalen. Maar indien dan blijkt dat meer dan negentig procent is geslaagd of meer dan vijftig procent is gezakt, past de docent de grens aan. Het compromis kan ook ontstaan door bijvoorbeeld 60% van de maximaal behaalde score te nemen als cesuur. Dit is de zogenaamde methode Cohen. De compromismethode leidt ertoe dat het aantal gezakten en geslaagden op de verschillende toetsmomenten geen al te grote fluctuaties vertonen.

### Contaminatie-effect

Het contaminatie-effect is het effect dat optreedt bij beoordelen wanneer de vrijheid in de beoordeling, onwillekeurig of willekeurig, wordt gebruikt voor andere doeleinden dan die van een onbevangen, onbevooroordeelde beoordeling. Dit is bijvoorbeeld het geval als beoordelaars lagere scores geven om aan te tonen dat hun vak moeilijk is.

### Correctievoorschrift (antwoordmodel)

Een hulpmiddel bij het beoordelen van open vragen is het hanteren van een antwoordmodel, ook wel aangeduid als correctievoorschrift. Wanneer beoordelaars een correctievoorschrift hanteren, verkleint dat de verschillen tussen beoordelaars. Een correctievoorschrift leidt ertoe dat (extreem) strenge en (extreem) soepele beoordelaars zich matigen omdat ze niet hun eigen intuïtieve oordelen mogen geven. Het hanteren van een dergelijk voorschrift voorkomt ook dat docenten verschillende criteria hanteren bij de beoordeling. Een correctievoorschrift bevat algemene en vakspecifieke regels voor de beoordeling en voorbeelden van juiste, gedeeltelijk juiste en onjuiste antwoorden. Ook bevat het correctievoorschrift instructies hoe de maximaal te behalen score per vraag moet worden verdeeld in deelscores.

### Criteriumgericht interview

Het criteriumgericht interview is een individueel vraaggesprek ten behoeve van een beoordeling. De beoordelaars onderzoeken op een gestructureerde manier door middel van vragen in hoeverre een student voldoet aan bepaalde criteria. Met een dergelijk interview worden de volgende doelen nagestreefd: 1. Zicht krijgen op kennis, vaardigheden en persoonlijke kwaliteiten die de beoordelaars niet zelf hebben waargenomen of kunnen waarnemen. 2. Nagaan of de student bewust competent is: het weten waaróm je zo hebt gehandeld is een belangrijk aspect van competent zijn. Een criteriumgericht interview staat veelal niet op zichzelf, maar gaat vaak gepaard met een observatie en/of documentanalyse. In dat geval dient het interview om meer informatie te verzamelen dan waartoe de observatie of de documentanalyse gelegenheid biedt.



### Creatief toetsen

Toetsen in het onderwijs zijn doorgaans sterk gericht op het kunnen reproduceren van kennis. Deze benadering van toetsen heeft tot gevolg dat studenten het accent leggen op uit het hoofd leren, waardoor het toepassen van deze kennis in het gedrang komt. Zodra toepassing van die kennis wordt verlangd, zullen studenten hun leerinspanning daar op gaan richten, want zoals bekend: de toets stuurt het studeergedrag. Het is gewenst dat studenten in het hoger onderwijs met een grote variëteit aan toetsvormen in aanraking komen, waarin enerzijds een beroep wordt gedaan op kennis en vaardigheden sec en anderzijds op integratie ervan en op het creatief kunnen omgaan met opgedane kennis.

### Digitaal toetsen

Digitaal toetsen is een breed begrip waaronder wordt verstaan het inzetten van computers bij het toetsproces. Dit kan zijn bij het beheer van toetsvragen in een itembank en het afnemen en/of het nakijken van toetsen.

### Doelen van toetsen

In het onderwijs worden verschillende doelen toegekend aan toetsen: selectie, leermiddel, voortgangscontrole en kwalificatie. Selectie vindt plaats voorafgaande aan de start van de opleiding, de toets als leermiddel en voortgangscontrole tijdens het onderwijs en het kwalificatiedoel aan het eind.

### Dublin-descriptoren

Op initiatief van Nederland en Vlaanderen zijn in de jaren 2001-2004 voor het hoger onderwijs in Europa de zogenaamde Dublin-descriptoren opgesteld, genoemd naar de plaats waar in 2004 een overeenkomst over deze eindtermen werd bereikt. De Dublin-descriptoren zijn de eindtermen voor de bachelor- en masterstudies aan universiteiten en hogescholen in Europa. De aanleiding voor het opstellen van deze descriptoren was het streven naar opleidingen van gelijkwaardig niveau in de verschillende landen van Europa. Ook wilde men dat helder zou worden wat de verschillen zijn tussen de kwalificaties of eindtermen voor de twee onderwijstrajecten bachelor en master in het hoger onderwijs. Dit mede met het oog op studenten die een deel van hun studie aan een buitenlandse universiteit willen doen.

De Dublin-descriptoren bestaan uit een lijst met criteria (kennis en inzicht; toepassing kennis en inzicht; oordeelsvorming; communicatie; leervaardigheid), met per criterium een globale uitwerking van het eindniveau van een bachelor dan wel master.

### EVC (Erkenning van Verworven Competenties)

In een kennisintensieve samenleving wordt steeds meer kennis ontwikkeld en benut buiten het geïnstitutionaliseerde onderwijs. Er wordt dan ook steeds meer buiten een formeel opleidingsinstituut/school geleerd. EVC is het erkennen van formeel, informeel en non-formeel leren. Hoe en waar is geleerd is binnen de EVC-gedachte minder belangrijk dan wat is geleerd. Toetsprocedures voor EVC zijn gericht op toetsen van leerresultaten, ongeacht de gevolgde leerweg. Het gaat om leerwegonafhankelijk toetsen.

Vrij recentelijk is een kwaliteitscode EVC ontwikkeld. Nederland heeft ongeveer honderd organisaties die zijn erkend zijn als EVC-aanbieder. In de code staan uitgangspunten aanzien van vijf onderdelen, te weten het doel van EVC, de rechten van de kandidaat, het EVC-onderzoek, de assessoren en begeleiders en de kwaliteitszorg.

### Examencommissie

De examencommissie is het orgaan dat op objectieve en deskundige wijze vaststelt of een student voldoet aan de voorwaarden die de onderwijs- en examenregeling stelt ten aanzien van kennis, inzicht en vaardigheden die nodig zijn voor het verkrijgen van een graad (WHW; 7.12). De examencommissie heeft ook tot taak de kwaliteit van examens en tentamens te borgen. De examencommissie functioneert onafhankelijk van het (instellings/faculteits)bestuur.

### Formatief toetsen

Formatief toetsen betreft het proces van het zoeken, aggregeren en interpreteren van informatie die studenten en docenten gebruiken om te bepalen waar studenten staan in hun leerproces, waar zij naar toe moeten werken en op welke manier. De informatie die formatief toetsen oplevert, kan de student gebruiken om het eigen leerproces te sturen, maar geeft docenten ook informatie om het eigen onderwijs te evalueren.

### Gedragscode praktijkgericht onderzoek

De Gedragscode voor praktijkgericht onderzoek beoogt een bijdrage te leveren aan het bevorderen van de kwaliteit van het onderzoek aan hogescholen. De gedragscode beschrijft gewenst gedrag van een gespecificeerde groep beroepsbeoefenaren bij het uitvoeren van een bepaalde taak. De Gedragscode voor praktijkgericht onderzoek bestaat uit vijf regels. Dit zijn regels voor verantwoord gedrag van studenten en medewerkers tijdens het doen van praktijkgericht onderzoek in het hoger beroepsonderwijs in Nederland. De code beschrijft gewenst gedrag bij het uitvoeren van de 'taak' onderzoeken.

### Genade-zesje

Soms zijn studenten aan het eind van hun studie gekomen terwijl, bijvoorbeeld, uitsluitend één eerstejaars tentamen open staat. Het betreffende tentamen is al vele malen zonder succes door de student afgelegd. In veel gevallen wordt dan een mondeling afgenomen, waar studenten dan meestal nog een 'genade-zesje' voor krijgen.

### Gesloten toetsvragen

Een gesloten toetsvraag is een vraag die studenten kunnen beantwoorden door het simpelweg aankruisen van een antwoord. De mogelijke antwoorden staan onder de vraag. Er zijn verschillende typen gesloten toetsvragen. De bekendste is een vraag gevolgd door 2/3/4/5 mogelijke antwoorden (alternatieven). Uit doelmatigheidsoverwegingen heeft het de voorkeur studenten te laten kiezen uit drie antwoorden. Daarnaast zijn er gesloten vragen waarop studenten antwoorden met juist of onjuist, of ja en neen. Meestal bevat de vraag dan een stelling.

### Halo-effect

Het halo-effect treedt op wanneer beoordelaars hun oordeel mede laten beïnvloeden door andere prestaties van de student dan geuit op de te beoordelen prestatie. De beoordelaars hebben in dat geval de neiging om een 'goede' student bij een wat mindere prestatie wat hoger te beoordelen dan de prestatie van de student rechtvaardigt. De prestatie wordt overgewaardeerd.

### Hink-stap-sprong (*triple-jump-test*)

De hink-stap-sprong is oorspronkelijk bedoeld om te toetsen of studenten zich de methodiek hebben eigen gemaakt die van hen wordt verwacht in een probleemgestuurd onderwijscurriculum (PGO). Het is een mondelinge toets bestaande uit drie stappen. In stap 1 legt de docent een probleem, vraagstelling, gevals- of situatiebeschrijving voor die betrekking heeft op het voorafgaande onderwijs en die in zekere mate representatief is voor de latere beroepsuitoefening. In de tweede stap gaan studenten op zoek naar aanvullende informatie om de vragen te kunnen beantwoorden. Na zelfstudie komt de student in stap 3 terug bij de docent met een synthese van het probleem en met antwoorden. Dan volgt de beoordeling.

### Horn-effect

Het horn-effect is het tegenovergestelde van het halo-effect. Het horn-effect treedt op wanneer beoordelaars hun oordeel mede laten beïnvloeden door andere prestaties van de student dan geuit op de te beoordelen prestatie. De beoordelaars hebben in dat geval de neiging om een 'slechte' student bij een wat hogere prestatie wat lager te beoordelen dan de prestatie van de student rechtvaardigt. De prestatie wordt ondergewaardeerd.

### In- en aanvulvragen

In- en aanvulvragen zijn een vorm van open toetsvragen en bevatten een onvolledige zin, berekening of tekening die studenten moeten completeren. Lezen en beantwoorden van een dergelijke vraag kost weinig tijd, hooguit tien seconden. Dit type vragen kan automatisch worden nagekeken.

### Inhoudsvaliditeit (*representativiteit*)

De inhoudsvaliditeit van een toets, ook representativiteit genoemd, is de mate waarin de toets datgene toetst wat de docent beoogt te toetsen. De inhoud van de toets moet de beoogde doelstellingen representeren. Een hulpmiddel om een toets inhoudsvalide te laten zijn, is het hanteren van een specificatietabel of een blauwdruk.

### Itembank

Een itembank is een voorraad vragen ten behoeve van het samenstellen van een toets. Betrof de itembank vroeger vaak een kaartenbak, nu zijn de meeste itembanken elektronisch. Dat geeft uitgebreide beheermogelijkheden: allerlei eigenschappen van de toetsvraag kunnen eraan worden gekoppeld, zoals de vraagmoeilijkheid bij eerdere afnamen, het onderscheidingsvermogen bij eerdere afnamen, het aantal keren dat een vraag is gebruikt, het onderwerp, het cognitieve niveau, de taxonomiecode. Aan de hand van die eigenschappen kan de toetsverantwoordelijke een toets samenstellen met bepaalde eigenschappen.

### Juist/onjuist-vraag

Juist/onjuist-vragen komen in verschillende varianten voor. De meest bekendste variant bestaat uit een stelling waarvan studenten moeten aangeven of deze juist of onjuist is. De populariteit van de juist/onjuist-vraag is voornamelijk te danken aan de Amerikaanse toetsdeskundige Ebel. Hij stelt dat alle kennis kan worden uitgedrukt in proposities, waarbij een propositie een stelling of uitspraak is die juist of onjuist is.

### Korte-antwoordvragen

Op korte-antwoordvragen, een vorm van een open toetsvraag, moet met een citaat, met enkele woorden, een enkelvoudige zin, een getal, een (eenvoudige) tekening of formule worden geantwoord. De tijd voor het lezen en de beantwoording zal meestal niet meer dan een minuut vergen.

### Landelijke kennistoets

Een landelijke kennistoets wordt afgenomen bij studenten aan vergelijkbare opleidingen in het hele land, bijvoorbeeld bij alle studenten in de lerarenopleiding Geschiedenis. De behoefte aan de landelijke kennistoets is ontstaan vanwege twijfels over de kwaliteit van sommige opleidingen. Door landelijke kennistoetsen te ontwikkelen, vindt er een externe validering plaats van het eindniveau. De ontwikkeling van een landelijke kennistoets start met het formuleren van een kennisbasis: een systematische beschrijving van wat afgestudeerden van een bepaalde opleiding aan vakkennis moeten bezitten.

### Lange-antwoordvragen

Lange-antwoordvragen is een vorm van een open vraag. Studenten dienen lange-antwoordvragen te beantwoorden met samenhangende zinnen. Ook een toelichting, een gecompliceerde berekening, tekening of een bewijs bestaande uit verschillende stappen, zijn voorbeelden van antwoordmogelijkheden. Voor het lezen en beantwoorden van een dergelijke vraag is al gauw drie tot vijf minuten nodig.

### Meerkeuzevraag

Een meerkeuzevraag bestaat uit twee delen: de stam en de alternatieven. De stam is het gedeelte van de vraag dat het probleem bevat. De alternatieven zijn de antwoordmogelijkheden waaruit studenten kiezen. Niet-correcte alternatieven worden afleiders genoemd; sleutel is de benaming voor het correcte alternatief.

### Mondelinge toets

Bij een mondelinge toets vindt een gesprek plaats tussen aan de ene kant twee docenten en aan de andere kant een student. Beide docenten stellen in overleg het cijfer vast.

### Normeren

Normeren is het relateren van een toetsscore aan een criterium. Dat kan een relatief criterium zijn, bijvoorbeeld de scores van toetsdeelnemers (relatief normeren), of een vast criterium, bijvoorbeeld de maximaal te behalen punten (absoluut normeren). In de praktijk gebruiken docenten vaak compromismethoden.

### Normverschuiving

Normverschuiving is het effect dat optreedt wanneer een beoordelaar een beoordeling aanpast aan de geleverde prestaties van studenten. Beoordelaars kunnen bijvoorbeeld minder streng worden als na een aantal beoordelingen blijkt dat de meeste studenten eenzelfde vraag fout beantwoorden.

### Objectiviteit

Een toets is objectief wanneer de uitslag van de toets vrij is van onbedoelde invloeden van de nakijker of de ontwikkelaar van de toets. Het mag er niet toe doen wie de toets construeert of nakijkt.

### Onderscheidingsvermogen

Elke toetsvraag in een summatieve toets moet onderscheid maken – discrimineren – tussen studenten die de doelstellingen wel of niet beheersen. Dit wordt het onderscheidingsvermogen van de vraag genoemd, aangeduid met  $R_{it}$ . Andere benamingen zijn: het discriminerend vermogen of de discriminatie-index. De  $R_{it}$  kan waarden aannemen tussen ‘-1’ en ‘+1’. Een correct discriminerende vraag heeft een positieve  $R_{it}$ -waarde.

### Open boektoets

Bij een open boektoets mogen studenten boeken, artikelen en aantekeningen, of een specifieke selectie daaruit, meenemen en raadplegen. Vragen die een beroep doen op parate kennis, zijn voor deze toetsvorm ongeschikt omdat ze simpelweg kunnen worden opgezocht.

### Open toetsvraag

Een open toetsvraag is een vraag waarbij studenten zelf het antwoord moeten formuleren. Dat antwoord kan in lengte variëren van één woord of enkele woorden tot een volledig betoog. De open vraag heeft voordelen, maar geeft ook problemen. Voor een niet-onbelangrijk deel zijn de nadelen terug te voeren tot kwesties die met het formuleren te maken hebben. Open vragen zijn bij uitstek geschikt om hogere cognitieve vaardigheden te toetsen. Dat betreffen vaardigheden als toepassen, analyseren, synthetiseren, evalueren.

### Opstelvragen/betoogvragen

De opstel/betoogvraag is een vorm van een open vraag. Typische antwoorden op opstelvragen/betoogvragen zijn: een samenhangende tekst, gestructureerd bijvoorbeeld naar inleiding, midden en slot, of een zeer gedetailleerde tekening of berekening die beide als een afgerond geheel worden beschouwd. Het lezen en beantwoorden van een dergelijke vraag is een tijdrovende aangelegenheid. Vijftien tot dertig minuten voor een dergelijke opdracht is niet ongebruikelijk.

### P-waarde

De p-waarde van een toetsvraag is de proportie studenten die de vraag correct heeft beantwoord. De p-waarde kan waarden aannemen tussen ‘0’ (geen enkele student heeft de

vraag correct beantwoord, en '1' (alle studenten beantwoordden de vraag correct). Indien de p-waarde wordt uitgerekend voor een meerkeuzevraag, moet men rekening houden met de raatkans. Een lage p-waarde geeft aan dat de vraag moeilijk was ( $p < \text{gokkans}$ ), een hoge p-waarde geeft aan dat de vraag makkelijk was (of dat de studenten respectievelijk niet goed of juist wel goed geleerd hebben).

### Peer-assessment

Bij peer-assessments beoordelen medestudenten elkaars werk en relateren dat werk aan de leerdoelen. Peer-assessment zorgt ervoor dat medestudenten worden betrokken bij het toetsproces door samen beoordelingscriteria op te stellen, werk van elkaar te bekijken en elkaar te voorzien van feedback. Cruciaal voor peer-assessment is dat studenten inzicht hebben in groepsdynamische processen.

### Peer review

Peer review is een toetsvorm waarbij studenten kritiek geven op een paper van een medestudent. Het doel van deze toetsvorm is na te gaan of studenten in staat zijn een paper kritisch te lezen en te beoordelen.

### Performance assessment

Performance assessment is een verzamelterm voor een aantal methoden waarmee gedrag kan worden uitgelokt bij studenten met als doel vast te stellen in hoeverre doelstellingen van hogere orde zijn verworven. De kwaliteit van een performance assessment wordt op de eerste plaats bepaald door de representativiteit van de assessmenttaken ten opzichte van domeinen. Het vaststellen van de beheersing van complexe doelen, bijvoorbeeld competenties, vraagt om een assessmentprogramma dat op systematische wijze dient te worden ontworpen.

### Piramide van Miller

De piramide van Miller wordt veel gebruikt bij het formuleren van toetsvragen. De piramide is horizontaal verdeeld in vier lagen die de complexiteit van een competentie uitbeelden, oplopend van relatief eenvoudig tot complex. Iedere laag symboliseert een niveau van competentie: weten, toepassen, demonstreren en doen.

### Plagiaat/plagiëren

Plagiëren op toetsen betekent dat studenten ten onrechte de suggestie wekken dat een tekst(fragment) van henzelf afkomstig is. De grens tussen plagiëren en correct citeren is niet scherp te trekken en is afhankelijk van de discipline.

### Psychometrische analyse

Een psychometrische (getalsmatige) analyse wordt uitgevoerd nadat de toets is afgenomen. De analyse geeft inzicht in enkele belangrijke kwaliteitsaspecten van een toets: de moeilijkheid, het discriminerend vermogen van de vraag, de (on)aantrekkelijkheid van de afleiders en de betrouwbaarheid van de toets als geheel. Deze informatie is van belang bij de beoordeling van de inhoudelijke kwaliteit van de toets. Om een psychometrische

analyse te kunnen uitvoeren, is het vereist een elektronisch bestand aan te maken, bijvoorbeeld in Excel.

### Relatief normeren

Bij het relatief geven van cijfers wordt de toetsscore van een student vergeleken met de toetsscores van de overige toetsdeelnemers. Of een student een hoog cijfer verdient, is dus afhankelijk van de toetsprestaties van anderen. Het hanteren van een relatieve manier van cijfergeven wordt vooral ingegeven door de gedachte dat studenten niet de dupe mogen worden van een toets die kwalitatief niet aan de maat is geweest.

### Representativiteit

Zie inhoudsvaliditeit

### Reproductiekennis

Wanneer kennisvragen zijn te beantwoorden door herinnering of herkenning, meet de vraag reproductiekennis. De vraag bevat een bekende probleemsituatie, zij het vaak in een geparafraseerde vorm.

### Restriction-of-range

Restriction-of-range is een effect dat optreedt bij het beoordelen. Het wordt veroorzaakt door de vrijheid in de beoordelingstaak die willekeurig of onwillekeurig leidt tot bepaalde verdelingen van de beoordelingen die algemeen menselijke of persoonlijke neigingen uitdrukken. Zo zal de ene beoordelaar wel alle schaalwaarden (scores van 1 tot 10) gebruiken, terwijl een andere beoordelaar steeds waarden zal toekennen in de buurt van het midden.

### Rubrics (rubrieken)

Rubrics beschrijven aan de hand van een set criteria de verschillende niveaus van kwaliteit, gerelateerd aan de leerdoelen. Rubrics kunnen worden gebruikt voor zowel formatieve doeleinden (om 'werk in ontwikkeling' te beschrijven) als voor summatieve doeleinden (als beoordelingscriteria om een leeruitkomst te beoordelen). Als beoordelingsrubrieken worden gebruikt bij de nabespreking van toetsen om studenten feedback te geven, zorgt het werken met rubrics voor een hogere motivatie en het bijsturen van het leren. Het werken met rubrics (rubrieken) is een effectieve methode om formatief toetsen te ondersteunen.

Er zijn twee soorten rubrics. Bij holistische rubrics worden complexe vaardigheden zo veel mogelijk als één geheel beoordeeld. De beoordelaar geeft dan een oordeel over een product op basis van een overall indruk. In een analytische rubric wordt gestreefd naar zo eenduidig en enkelvoudig mogelijke criteria, zodat de beoordelaar alle relevante beoordelingsaspecten afzonderlijk van een kwalificatie kan voorzien.

### SafeAssign

SafeAssign is een zogenaamd buildingblock van Blackboard. Het controleert teksten op plagiaat door de tekst te vergelijken met vele bestanden op internet. De uitkomst wordt gegeven in een percentage.

### Script Concordance Test (SCT)

In een Script Concordance Test worden problemen en situaties beschreven in de vorm van een casus met bijbehorende vragen. De vragen hebben de vorm van te toetsen hypothesen of suggesties voor een verdere aanpak, én aanvullende casusinformatie. Deelnemers wordt gevraagd te beoordelen in hoeverre de gesuggereerde hypothese of aanpak bevestigd c.q. verworpen moet worden op basis van de totale casusinformatie in het vignet en de aanvulling. Een casusvignet bevat de eerste informatie over de casus. Het geheel aan informatie geeft geen totaalbeeld van de casus. In principe wordt die informatie gegeven die in de praktijk ook over het (klinische) probleem bekend is op het moment waarop de overweging of beslissing zich voordoet.

Belangrijke kenmerken van een SCT zijn dat deze uitgaat van authentieke omstandigheden en problemen en omstandigheden waarbij meerdere opties mogelijk zijn, en dat wordt gevraagd naar een schatting van waarschijnlijkheden en onderbouwing daarvan. De rationale voor het hanteren van een dergelijk test is dat het oordelen en beslissen in situaties van relatieve onzekerheid, op basis van slechts beperkte informatie of onder tijdsdruk, voor velen – zeker in hogere functies en beroepen – onderdeel van hun professionele praktijk is.

### Selectie

Wanneer, voorafgaand aan het doorlopen van het onderwijsprogramma, moet worden nagegaan of een student geschikt is voor de opleiding, heeft de toets (ingangs)selectie tot doel. De functie is primair die studenten toe te laten tot de opleiding die geschikt zijn om de opleiding te voltooien. Maar er kunnen ook andere overwegingen een rol spelen bij selecteren. Door middel van het limiteren van het aantal toe te laten studenten, kan men proberen het aantal beroepsbeoefenaren te beperken, of andersom: meer beroepsbeoefenaren door meer personen toe te laten tot de opleiding.

### Sequentie-effect

Het sequentie-effect treedt op wanneer een beoordelaar de beoordeling ten onrechte mede laat leiden door een of meer voorafgaande beoordelingen. Bijvoorbeeld, beoordelaars kijken toetsvragen in een bepaalde volgorde na. Een groot aantal slechte antwoorden dat wordt gevolgd door een correct antwoord, kan er dan toe leiden dat het correcte antwoord onevenredig hoog wordt beoordeeld.

### Signifisch effect

Het significisch effect treedt op wanneer beoordelaars letten op verschillende aspecten, of als ze de aspecten verschillend wegen in hun oordeel. Dit effect treedt bijvoorbeeld op wanneer beoordelaars schrijfproducten beoordelen. De ene beoordelaar let vooral op de grammaticale (on)juistheden, en de andere op structuur of inhoud.

### Sleutel

Een meerkeuzevraag bestaat uit twee delen: de stam en de alternatieven. De stam is het gedeelte van de vraag dat het probleem bevat. De alternatieven zijn de antwoordmogelijk-



heden waaruit studenten kunnen kiezen. De niet-correcte alternatieven worden afleiders genoemd; het correcte alternatief is de sleutel.

### Specificatietabel (toetsmatrijs)

Een specificatietabel, ook wel toetsmatrijs genoemd, dient ervoor te zorgen dat in de toets, naast een spreiding van de vragen over belangrijke onderwerpen die in de blauwdruk staan, ook op alle beoogde (cognitieve) niveaus wordt getoetst. Per onderwerp staat aangegeven op welk niveau docenten toetsen.

### Summatieve toets

Een summatieve toets heeft de functie een oordeel over de kennis en vaardigheden van een student uit te spreken. Aan de resultaten van summatieve toetsen zijn consequenties verbonden, meestal in de vorm van een cijfer. Summatieve toetsen moeten voldoen aan hoge eisen van betrouwbaarheid en validiteit. Voorbeelden van summatieve toetsen zijn opdrachten, overhoringen, proefwerken, tentamens, (praktijk)examens. Het belangrijkste doel van summatieve toetsen is meten. Tegenover de summatieve toets staat de formatieve toets.

### Toetsbeleid

Toetsbeleid is een samenhangend stelsel maatregelen en voorzieningen met als specifiek doel om de kwaliteit van toetsing en examinering te bewaken en te bevorderen. Dat samenhangend stelsel kan worden beschreven aan de hand van verantwoordelijkheden van de belangrijkste spelers in de organisatie, hun activiteiten en de documenten waaruit hun activiteiten blijkt.

Een toetsbeleid kan op verschillende niveaus worden geformuleerd: instelling, faculteit en opleiding. In het algemeen geldt dat naarmate het beleid lager in de organisatie is geformuleerd, hoe concreter het is.

### Toetsbeleidsplan (facultair)

Het toetsbeleidsplan bevat het toetsbeleid van het faculteitsbestuur. Het plan dient binnen de kaders te blijven die het instellingsbestuur heeft vastgesteld. Onderwerpen hierin zijn: 1. De visie op toetsing; 2. Richtlijnen voor het format van het toetsplan dat geldt voor een onder het faculteitsbestuur vallende opleiding; 3. Wijze waarop de toetsplannen van opleidingen moeten worden ingebed in de kwaliteitszorg in het algemeen; 4. Het minimum aan toetsdeskundigheid voor docenten en het minimum aan bijscholingsverplichting; 5. De wijze waarop het niveau van de eindwerkstukken wordt bewaakt; 6. Minimumeisen die aan alle tentamens en toetsen worden gesteld; 7. De wijze waarop de opleidingsdirecteuren de examencommissie faciliteren in het uitvoeren van haar wettelijke taken.

### Toetsfolder

Opleidingen kunnen hun toetsen transparant maken door een toetsfolder te laten verschijnen. In een dergelijke folder staat informatie over de toets: onderwerpen, soort vragen plus voorbeeldvragen, beoogd niveau, c.q. verhouding kennis-, toepassings- en inzichtvragen, compensatiemogelijkheden, en uitslagbepaling.

### Toetskader

Het instellingsbestuur formuleert een toetskader dat een resultante is van een analyse van externe ontwikkelingen en de eigen visie op besturen en opleiden. Het toetskader houdt rekening met de wettelijke kaders en politieke en maatschappelijke vragen. Met het toetskader bepaalt het instellingsbestuur de ruimte die de faculteiten hebben met betrekking tot hun eigen toetsbeleid. Een instellingsbreed toetskader zal zich beperken tot hoofdlijnen. Een toetskader zal de volgende ingrediënten bevatten: 1. De eisen die de wettelijke kaders en maatschappelijke en politieke context stellen ten aanzien van de toetsing in het hoger onderwijs; 2. De onderwijsvisie van de instelling en aansluitend de visie op toetsing; 3. Het beleid dat op instellingsniveau wordt vastgesteld en dat gerelateerd is aan toetsen, zoals eisen ten aanzien van deskundigheid van docenten; 4. Maatregelen die het instellingsbestuur voor de hele instelling hetzelfde wil hebben geregeld om 'in controle' te kunnen zijn wat betreft de kwaliteit van toetsing; 4. Onderwerpen die een andere laag in de organisatie (formeel) niet kan regelen; bijvoorbeeld centraal overleg tussen de examencommissies; 5. Onderwerpen die vanuit doelmatigheid beter op hoger organisatieniveau kunnen worden geregeld.

### Toetsproces

Het afnemen van toetsen impliceert meer dan het construeren van toetsvragen; het is een proces dat verschillende fasen en stappen kent. Elk toetsproces is in essentie te herleiden tot een viertal fasen waarin steeds een ander aspect op de voorgrond staat: doelbepaling, meten, cijfergeven en beslissen. Deze ontleding van het toetsproces geldt voor alle toetsen met een summatief doel. De grenzen tussen de fasen zijn niet precies te trekken omdat bij het concretiseren van een fase rekening wordt gehouden met de andere fase, bewust of onbewust. Elke fase is onder te verdelen in stappen.

### Toetsprogramma

Een toetsprogramma bestaat uit een combinatie van toetsvormen dat (bewust) is samengesteld, passend bij de inhoud, opbouw en structuur van het curriculum. Qua inhoud zijn de toetsen in een toetsprogramma complementair. Ze vullen elkaar aan wat betreft de dekking van de onderwijsdoelen. Qua opbouw en structuur volgen de toetsen het curriculum.

### Transparantie

In een toetssituatie spreekt men van transparantie indien alle informatie voorhanden is die de best denkbare voorbereidings- en antwoordstrategie voor studenten mogelijk maakt. Over het algemeen kan aan de transparantie-eis worden voldaan door uitgebreid voorlichting te geven in het voorafgaande onderwijs, bijvoorbeeld met een toetsfolder, en/of via een toetsinstructie bij de toets.

### Validiteit

De validiteit van een toets heeft te maken met de betekenis, de bruikbaarheid en de juistheid van de conclusies die uit de toetsscores worden getrokken. Al naargelang de aard van de conclusies zijn er drie typen validiteit te onderscheiden: inhouds-, begrips- en criteriumvaliditeit (ook wel genoemd 'voorspellende' validiteit genoemd). Inhoudsvaliditeit is

de mate waarin de toets als geheel representatief is voor datgene wat men met de toets wil nagaan. Begripsvaliditeit is de mate waarin de toetsscore kan worden beschouwd als een maat voor het bedoelde theoretische begrip. Criteriumvaliditeit is de mate waarin de toetscore samenhangt met soortgelijke metingen buiten de toets. Een hulpmiddel om ervoor te zorgen dat de toets inhouds- en begripsvalide is, is de specificatietabel.

### Voortgangstoets

Een voortgangstoets is een vaak omvangrijke toets die gericht is op cognitieve doelstellingen van het onderwijsprogramma. Een voortgangstoets wordt afgenomen bij alle studenten van alle jaargroepen. De ontwikkeling (groei) in kennis door de studie jaren heen wordt gemeten. Voortgangstoetsing biedt de mogelijkheid om vast te stellen hoe ver een student wat betreft kennisniveau verwijderd is van het einddoel, en ook welke kennisgroei zich in een bepaalde periode heeft voorgedaan. Voortgangstoetsing gaat toetsgericht studeren tegen en bevordert een continue wijze van studeren.

### WCOPYFIND

WCOPYFIND is een free access softwareprogramma dat dient ter opsporing van plagiaat. Het vergelijkt elektronisch ingeleverde papers met elkaar en spoort identieke tekstfragmenten op. De output bestaat uit een lijst van het aantal overeenkomende woorden tussen steeds twee papers.

### Werkplekleren

Werkplekleren is het leren in de beroepspraktijk. De stage is de bekendste vorm. Daarnaast is er de duale opleiding, waarbij de student een belangrijk deel van de opleiding op de werkplek volgt, maar waarbij de kwaliteit van de werkplek door de onderwijsinstelling wordt gegarandeerd. Ook de leerafdeling is een vorm van werkplekleren. Kenmerkend voor leerafdelingen is dat een student samen met andere studenten op een afdeling aan het werk gaat, waarbij zij als team alle werkzaamheden uitvoeren die normaliter tot de taken van die afdeling behoren. Bij het toetsen van werkplekleren worden onder andere de volgende instrumenten gehanteerd: portfolio, 360 graden feedback, persoonlijk ontwikkelingsplannen, ontwikkelingsgesprekken, performance assessments en stageverslagen.

### Zelfbeoordeling (self-assessment)

Zelfbeoordeling is een methode waarin studenten hun eigen werk relateren aan de leerdoelen. Het bestaat uit drie stappen: studenten beoordelen hun eigen werk, identificeren verschillen tussen hun werk en de gewenste uitkomsten, en voeren nieuwe activiteiten uit in de richting van de einddoelen. Zelfbeoordeling staat veelal niet op zichzelf, maar wordt gecombineerd met een aantal andere toetstaken. Zelfbeoordeling draagt bij aan het ontwikkelen van de zelfregulatievaardigheden van studenten.

# Referenties

A hand holding an open book with glowing pages, set against a blue background with light trails.

- Akker, J. van den (1999). Principles and methods of development research. In J. van den Akker, R.M. Branch, K. Gustafson, N. Nieveen, & T. Plomp (eds.), *Design Approaches and Tools in Education and Training* (pp. 1–14). Dordrecht: Kluwer.
- Allal, L., & Lopez, L. M. (2005). Formative assessment of learning: A review of publications in French. In J. Looney (Ed.), *Formative assessment: Improving learning in secondary classrooms* (pp. 241–264). Paris, France: Organisation for Economic Cooperation and Development.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, G.L., & Herr, K. (1999). The new paradigm wars: Is there room for rigorous practitioner knowledge in schools and universities? *Educational researcher*, 28(5), 12–21.
- Anderson, L. W., Kratwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*.
- Andriessen, D., J. Onstenk, P. Delnooz, H. Smeijsters, & S. Peij (2010). *Gedragscode praktijkgericht onderzoek voor het HBO. Gedragscode voor het voorbereiden en uitvoeren van praktijkgericht onderzoek in het Hoger beroepsonderwijs in Nederland. Advies van de commissie gedragscode praktijkgericht onderzoek in het HBO, HBO-Raad*.
- ARG (2002). *Assessment for learning: 10 principles*. Cambridge UK: University of Cambridge School of Education Author.
- Arnold, L., & Willoughby T.L. (1990). The quarterly profile examination. *Acad.Med*, 65(8), 515–516.
- Arter, J. A., & Chappuis, J. (2006). *Creating & recognizing Quality Rubrics*. Boston: Pearson.
- Arts, J., Jaspers, M., & Joosten-ten Brinke, D. (ingediend). *Written feedback as a form of feedback: So much to gain*.
- Baartman, L.K.J., Bastiaens, Th.J., Kirschner, P.A. (2005). *Kwaliteitscriteria voor Competentie Assessment Programma's*. *Examens. Tijdschrift voor de toetspraktijk*, 2, 13–15.
- Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment: presenting quality criteria for competency assessment programs. *Studies in Educational Evaluations*, 32(2), 153–170.
- Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Determining the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation*, 33, 258–281.
- Baartman, L.K.J. (2008). 'Assessing the assessment'. Development and use of quality criteria for Competence Assessment Programmes. *Proefschrift, Universiteit Utrecht*.
- Baartman, L. K. J., & de Bruijn, E. (2011). Integrating knowledge, skills and attitudes: Conceptualising learning processes towards vocational competence. *Educational Research Review*, 6, 125–134.
- Baartman, L. K. J., & Gulikers, J. T. M. (in druk). *Beoordeling als fundament van goed beroepsonderwijs: een analyse van 11 toetsprogramma's in het mbo en hbo*. *Pedagogische Studietoetsen*.
- Barnes, C.M., & Morgeson, F.P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. *Human Performance*, 20, 259–274.
- Barreveld, S. (2012). *Docenten volgen nog steeds de methode*. *Didactief*, 30–31.
- Bax, A.E. (2004). *Loon naar werken. Beoordelingsmethoden voor het toekennen van individuele cijfers aan groepsproducten*. *Examens. Tijdschrift voor de toetspraktijk*, 4, 18–21.
- Bennett, R. E. (2011). *Formative assessment: A critical review*. *Assessment in Education: Principles, Policy & Practice*, 18, 5–25.
- Bentum, J. van (2012) *Projectplan EVC in de Jeugdzorg*. Den Bosch: Avans Hogeschool.
- Berg, I. van den, Pilot, A., & Admiraal, W. (2005). *Peer assessment als leermiddel. Voorbeelden uit het hoger onderwijs*. *IVLOS, Universiteit Utrecht*.
- Berkel, A. van (2003). *Het ontwikkelen van een systeem voor EVC*. Leuven: KH Leuven (congresmap EVC-studiedag 21-11-03).
- Berkel, A. van (2012). *Kritische reflectie op competentietoetsen in het hbo*. *OnderwijsInnovatie*, 3, 17–26.
- Berkel, H.J.M. van (1984). *De diagnose van toetsvragen*. Amsterdam: Universiteit van Amsterdam (proefschrift).
- Berkel, H.J.M. van, & Bax, A.E. (Eds.)(1993). *Beoordelen in het Onderwijs*. Houten: Bohn Stafleu Van Loghum.
- Berkel, H.J.M. van, & Starren, J. (1993). *Beoordelen met studietoetsen*. In H.J.M. van Berkel, & A.E. Bax (Eds.), *Beoordelen in het onderwijs*. Houten: Bohn, Stafleu & Van Loghum.
- Berkel, H.J.M. van (1996). *Beroepszaken*. *Studiehuisreeks*, 7, 42–61.
- Berkel, H.J.M. van (1999). *Zicht op toetsen*. Assen: Van Gorcum.
- Berkel, H.J.M. van, & Bax, A.E. (2006). *Toetsen in het hoger onderwijs*. 2e druk. Houten: Bohn Stafleu Van Loghum.

- Berkel, H.J.M. van (2011). Selecteren aan de poort? Niet doen. *Thema* (5), 43–47.
- Biggs, J. B. (1999). *Teaching for quality learning at university*. Open University Press.
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 347–364.
- Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: the SOLO taxonomy* New York: Academic Press.
- Birenbaum, M., Kimron, H., & Shilton, H. (2011). Nested contexts that shape assessment for learning: School-based professional learning community and classroom culture. *Studies in Educational Evaluation*, 37(1), 35–48.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.
- Bloom, B. S. (1956). *Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain*. New York: McKay Bloom's Taxonomy.
- Bloom, B.S. (Eds.) (1972). *Taxonomy of educational objectives*. New York: David McKay.
- Bolhuis, S., & Simons, R.J. (2011). Naar een breder begrip van leren. In Poell, R., & Kessels, J. (red). *Handboek Human Resource Development*. Springer: Dordrecht.
- Bolhuis, S. (2012). Praktijkonderzoek als professionele leerstrategie in onderwijs en opleiding. In S. Bolhuis, & Q. Kools (Red.). *Praktijkonderzoek als professionele leerstrategie in onderwijs en opleiding*. Fontys Hogescho- len. ► [www.hbo-kennisbank.nl](http://www.hbo-kennisbank.nl).
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061.
- Boud, D. (1995). Assessment and learning: contradictory or complementary? In P. Knight (Ed.) *Assessment for Learning in Higher Education*, London: Kogan Page, p. 35–48.
- Boud, D., & Falchikov, N. (2007). *Rethinking Assessment in Higher Education: Learning for the Longer Term*. Oxon: Routledge.
- Boud, D. (2010). *Student assessment for learning in and after courses*. Sydney: University of Technology.
- Brennan, R.L. (Ed.) (2006). *Educational measurement*. New York: American Council on Education.
- Breukers, H., Doorten, M., Joosten-ten Brinke, D., Loth, F., Moerkerke, G., De Roode, F., Schlusmans, K. (2011). *Handboek toetsen en tentamens bij de Open Universiteit*. Heerlen: Open universiteit. ► <http://hdl.handle.net/1820/3884>.
- Bronneman-Helmers, R. (2006). *Duaal als ideaal?* Den Haag: Sociaal en Cultureel Planbureau.
- Brookhart, S. M., Moss, C. M., & Long, B. A. (2008). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education: Principles, Policy & Practice*, 17, 41–58.
- Bügel, K., & Sanders, P.F. (red.) (1998). *Richtlijnen voor de ontwikkeling van onpartijdig toetsen*. Arnhem: Cito.
- Burg, J. ter, & Kempers-Warmerdam (2013). De logica van mondelinge toetsing. *Examens*, 10 (1), 27–30.
- Camp, A. J., Donkers, J., & Van de Kamp, F. (2012). Het VGTogether project: samenwerken in voortgangstoetsing. Paper presented at the NVMO 2012, Maastricht, Nederland.
- Cauley, K. M., & McMillan, J. H. (2009). Formative assessment techniques to support student motivation and achievement. *Clearing House*, 83(1), 1–6.
- Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. P. M. (2000). The Script Concordance Test, a Tool to Assess the Reflective Clinician. *Teaching and Learning in Medicine*, 12, 189–195.
- Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Academic Medicine*, 75, 182–190.
- Charlin, B., & van der Vleuten, C. P. M. (2004). Standardized Assessment of Reasoning in Contexts of Uncertainty: The Script Concordance Approach. *Evaluation & the Health Professions*, 27(3), 304–319.
- Chi, M.T.H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R.J. Sternberg (Ed.) *Advances in the psychology of human intelligence*. Hillsdale New Jersey, USA: Lawrence Erlbaum Associates (p. 7-76).
- Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, 24(2), 205–249.
- Cluitmans, J. (2000). *Toetsing en zelfstandig leren. Gevarieerd toetsen binnen probleemgestuurd onderwijs en projectonderwijs in HBO en MBO*. Nuenen: OAB Dekkers.
- Cohen-Schotanus, J., Vleuten, C.P.M. van der, & Bender, W. (1996). Een betere cesuur bij tentamens. *Onderzoek van Onderwijs*, 3, 54–55.
- Colley, H., Hodkinson, P., & Malcolm, J. (2002). *Informality and formality in learning: a report for the Learning and Skills Research Centre*. London: Learning and skills Research Centre. Also available in the informal education archives: ► [www.infed.org/archives/e-texts/colley\\_informal\\_learning.htm](http://www.infed.org/archives/e-texts/colley_informal_learning.htm).
- Commissie Toekomstbestendig Hoger Onderwijs Stelsel. (2010). *Differentiëren in drievoud*. Den Haag: Ministerie van OCW.

- Commissie externe validering examenkwaliteit hoger beroepsonderwijs. (2011). *Vreemde ogen dwingen*, Den Haag: HBO-raad.
- Crisp, G. T. (2007). *The e-assessment handbook*. Continuum.
- Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational research*, 58, 438–481.
- Crombag, H.F.M. (1981). Wat niet moet bij het samenstellen van tentamens. *Onderzoek van Onderwijs*, 10 (1), 6–7.
- Davies, W.M. (2009). Groupwork as a form of assessment: common problems and recommended solutions. *Higher Education*, 58, 563–584.
- Dierick, S., Watering, G. van de, & Muijtens, A. (2002). De actuele kwaliteit van assessment: ontwikkelingen in de edumetrie. In F. Dochy, L. Heylen, & H. Van der Mosselaar, (Red.). *Assessment in het onderwijs, nieuwe toetsvormen en examinering in studentgericht onderwijs en competentiegericht onderwijs*, (pp. 91-122). Utrecht: Lemma.
- Dochy, F. J.R.C., & Segers. M.S.R. (1999). Innovatieve toetsvormen als gevolg van constructiegericht onderwijs: op weg naar een assessment-cultuur. In P. de Boeck, & M. Lacante (Eds.) *Meer kansen creëren voor het hoger onderwijs*. Dordrecht: Kluwer.
- Donk, C. van der, & Lanen, B van (2009). *Praktijkonderzoek in de school*. Uitgeverij Coutinho b.v.
- Dousma, T., Horsten, A., & Brants, J. (1997). *Tentamineren*. Groningen: Wolters-Noordhoff.
- Duvekot, R. (2012) *Evaluatie van de EVC.Pilot bij Transferium Jeugdzorg*. Leiden: Hogeschool Inholland.
- Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of Educational Measurement*. Englewood Cliffs, N.J.: Prentice Hall (second edition).
- Elsen, M., Wolters, L., Nedermeijer, J., & Pilot, A. (2001) *Geschiktheid van toetsvormen voor het meten van academische competenties*. *Tijdschrift voor Hoger Onderwijs* 19, 33–43.
- Elshout-Mohr, M., Oostdam, R., & Dietze, A. (2001). *Assessment van competenties*. *Velon Tijdschrift voor Lerarenopleiders*, 22, 48–54.
- Eraut, M., & du Boulay, B. (2000). *Developing the Attributes of Medical Professional Judgement and Competence*. ► [www.cogs.susx.ac.uk/users/bend/doh](http://www.cogs.susx.ac.uk/users/bend/doh).
- Elstein, A.S., Shulman, L.S., & Sprafka, S.A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press, Cambridge MA, USA.
- Fastre, G., Van der Klink, M.R., Sluijsmans, D., & Van Merriënboer, J.J.G. (2012). *Towards an integrated model for developing sustainable assessment skills*. *Assessment and Evaluation in Higher Education*, 1–20.
- Fitzpatrick, R., & Morrison, E.J. (1971). *Performance and Product Evaluation*. In E.L. Thorndike (Ed.), *Educational Measurement* (2nd edition), pgs. 237-270. Washington, DC: American Council on Education.
- Freeman, A, van der Vleuten, C, Nouns, Z., & Ricketts, C (2010) *Progress testing internationally*. *Medical Teacher*, 32(6), 451–5.
- Geerligts, T. (1993). *Beoordelen van scripties*. In H.J.M. van Berkel en A.E. Bax (Eds.) *Beoordelen in het Onderwijs*. Houten: Bohn Stafleu Van Loghum.
- Gibbs, G., & Simpson, C. (2004). *Conditions under which assessment supports students' learning*. *Learning and Teaching in Higher Education*, 1, 3–31.
- Gibbs, G. (2010). *Using assessment to support student learning*. Leeds Metropolitan University.
- Groot, A.D. de (1978). *Acceptabiliteit van selectiemethoden*. In A.G. Vroon, & S.E.M. Everwijn (Eds.), *Handboek voor de Onderwijspraktijk*. Deventer: Van Loghum Slaterus.
- Groot, A.D. de (1981). *Methodologie, grondslagen van onderzoek en denken in de gedragswetenschappen*. 's Gravenhage: Mouton.
- Groot, A.D. de, & Naerssen, R.F. van (1969). *Studietoetsen*. Den Haag: Mouton.
- Guile, D., & Griffiths, T. (2001). *Learning through Work Experience*. *Journal of Education and Work*, 14, 113–131.
- Gulikers, J. (2005). *Authentieke toetsing: de beroepspraktijk in het vizier*. *OnderwijsInnovatie*, 2, 17–24.
- Gulikers, J. T. M. (2006). *Authenticity is in the eye of the beholder. Beliefs and perceptions of authentic assessment and the influence on student learning*. Unpublished doctoral dissertation, Open University of the Netherlands, Heerlen.
- Gulikers, J., Bastiaens, T., & Kirschner, P. (2004). *A five-dimensional framework for authentic assessment*. *Educational Technology Research and Development*, 52(3), 67-85 Haan, J., & Staa, A. van (2013). *Gezamenlijk toetsen. Een leerschool voor gezamenlijke toetsing*. *HO management* 5 (1) 18–20.
- Haladyna, T.M. (1991). *Generic questioning strategies for linking teaching and testing*. *Educational Technology: Research and development*, 39, 73–81.
- Haladyna, T.M. (2004). *Developing and validating multiple-choice test items*. London: Erlbaum.
- Haladyna, T.M., & Downing, S.M. (1989). *Validity of a taxonomy of multiple-choice item-writing rules*. *Applied Measurement in Education*, 2, 51–78.

- Harlen, W., & Crick, R.D. (2003) Testing and Motivation for Learning Assessment in Education, 10, 2, 169-207 DOI: 10.1080/0969594032000121270.
- Harris, R. (2004). Anti-plagiarism strategies for research papers, VirtualSalt.
- Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77, 81-112.
- HBO-Raad (1999). Protocol Proefaccreditering, richtlijnen voor het accrediteringsproces. Den Haag: HBO-Raad. (Uitgave in de reeks: Proefaccreditering hbo, nr. 1)
- HBO-Raad/ Vereniging Hogescholen (2011). Geslaagd! Handreiking examencommissies. Advies voor hogescholen voor het sterker neerzetten van hun examencommissies. ► [www.hbo-raad.nl](http://www.hbo-raad.nl) (bekeken op 1-6-2013)
- Hennissen, P. (2011). In de nesten werken. Opleiden van leraren op de werkplek (lectorale rede). Heerlen: Zuyd Hogeschool.
- Hofstee, W.K.B. (1977). Cesuurprobleem opgelost. Onderzoek van Onderwijs, 6, 6-7.
- Hofstee, W.K.B. (1999). Principes van beoordelen. Lisse: Swets & Zeitlinger.
- Holleman, J.W., Oost, H., Milius, J., Berg, I. van den, & Admiraal, W. (1999). Gevraagd: academicus (m/v). (Mededeling 61). Utrecht: Universiteit Utrecht, IVLOS.
- Holmes, L. (2013). Competing perspectives on graduate employability: possession, position or process? Studies in Higher Education, 38, 538-554.
- Inspectie van het Onderwijs (2003). Zicht op toetsen. ► [www.onderwijsinspectie.nl/actueel/publicaties](http://www.onderwijsinspectie.nl/actueel/publicaties) (bekeken op 1-6-2013).
- Jankowicz, A.D. (2005). Business Research Projects. London: Thomson Learning.
- Jaspers, M., & Zijl, E. van (2011). Kwaliteit van Toetsing in het Hoger Onderwijs. Naar een versterkte rol van examencommissies. Fontys Hogescholen, dienst Onderzoek en Onderwijs.
- Jellema, F.A. (2006). Measuring training effects: The potential of 360-degree feedback (proefschrift). Enschede: Universiteit Twente.
- Johnson, D. W., Johnson, R. T., & Smith, K. A. (2007). The State Of Cooperative Learning In Postsecondary And Professional Settings. Educational Psychology Review, 19(1), 15-29.
- Joosten-ten Brinke, D. (2011). Eigentijds toetsen en beoordelen. Lectorale rede. Tilburg: Fontys Lerarenopleiding Tilburg.
- Kane, M.T. (1992). The validity of assessments of professional competence. (ERIC Document Reproduction Service No. ED 343 958).
- Kane, M.T. (2006). Validation. In R.L. Brennan (Eds). Educational Measurement (4th ed.), pp. 17-63. Westport, CT: Praeger Publishers.
- Keller, J. M. (1983). Development and use of the arcs model of motivational design. Enschede, Twente Univ. of Technology. ► [www.ittheory.com/keller1.htm](http://www.ittheory.com/keller1.htm).
- Kicken, W. (2008). Portfolio use in vocational education (proefschrift). Heerlen: Open Universiteit Nederland.
- Klarus, R. (1998). Competenties erkennen. Een studie naar modellen en procedures voor leerwegaafhankelijke beoordeling van beroepscompetenties. Cinop/KUN: 's-Hertogenbosch/Nijmegen (dissertatie).
- Klarus, R., Schuler, Y.D., & Wee, E. ter (2000). Bewezen geschikt. Wageningen: STOAS.
- Kleij, F. van der, Wools, S., & Straetmans, G. (2011). Mystery Guest. Mogelijkheden van mystery guest beoordelingen binnen het hbo. Verkregen op 20 februari 2013 van: ► [www.saxion.nl](http://www.saxion.nl).
- Klink, M. van der, Boon, J., & Schlusmans, K. (2012). All by myself. Employees' informal learning experiences. International Journal of Human Resources Development & Management, 12(1/2) 77-91.
- Kloppenburger, R.T. H. M. (2011). Bekwaam beoordeeld. inhoud, functie en kwaliteit van competentiegerichte assessments in social work opleidingen. Proefschrift, Universiteit Utrecht.
- Lampe, T., & Hendriks, P. (2010) Leren en beoordelen op de werkplek. Onderzoek naar de kwaliteit van praktijkstages bij hbo-opleidingen. Deventer: Saxion.
- Lappia, J. (2010). Opzetten, uitvoeren en evalueren van effectieve leerwerkarrangementen. In Streumer, J.N. (red). De kracht van werkpleklernen, 45-80. Utrecht: Boom Lemma.
- Lejk, M., Wyvill, M., & Farrow, S. (1996). A Survey of Methods of Deriving Individual Grades from Group Assessments. Assessment & Evaluation in Higher Education, 21, 267-280.
- Linn, R.L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13, 1, 5-15.
- Lord, F. M. (1963). Formula scoring and validity. Educ Psych Meas, 23(663), 672.
- Luken, T.P. (2004). Zijn competenties meetbaar. Dilemma en uitweg bij het werkbaar maken van het competentiebegrip. Tijdschrift voor Hoger Onderwijs, 22, 38-53.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. Language Testing, 12, 1, 54-71.



- Mabry, L. (1999). *Portfolio Plus. A Critical Guide to Alternative Assessment*. Corwin Press, Inc.: Thousand Oaks, California.
- Maesen, P.E.A.M. van der, & Schakel, L. (1999). Wat zijn competenties niet? *Opleiding & Ontwikkeling*, 12, 11–17.
- McHarg, J., Bradley, P., Chamberlain, S., Ricketts, C., Searle, J., & McLachlan, J. C. (2005). Assessment of progress tests. *Medical Education*, 39(2), 221–227.
- Merriënboer, J.J.G. van, Klink, M.R. van der, Hendriks, M. (2002). *Competenties: van complicaties tot compromis. Over schuifjes en begrenzers*. Den Haag: Onderwijsraad.
- Messick, S. (1994). Validity of psychological assessment. validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Milius, J., Oost, H., & Holleman, W. (2001). *Werken aan academische vorming; Ideeën voor actief leren in de onderwijspraktijk*. IVLOS, Universiteit Utrecht.
- Miller, G.E. (1990). The assessment of clinical skills/competence/performance. *Academic medicine*, 65, 63–67.
- Ministerie van OC&W (2008). *Krachtig meesterschap, Kwaliteitsagenda voor het opleiden van leraren 2008–2011*. Den Haag: Ministerie van OC&W.
- Moerkerke, G., & Dochy, F.J.R.C. (2000). Het toetsen van complexe vaardigheden. In G. ten Dam e.a. (Eds.) *Onderwijskunde Hoger Onderwijs, handboek voor docenten*. Assen: Van Gorcum.
- Moerkerke, G. (2012). *Eindrapport IPO project Facultaire Toetsbeleidsplannen*. Heerlen: Open Universiteit.
- Molkenboer, H. (2011). Van examens krijg je uitslag. Een recept voor oplossingen. *Expertise*, 14–16.
- Mueller, J. (z.j.). *Authentic Assessment Toolbox*. Verkregen op 14 maart 2013 van: ► <http://jfmuller.faculty.noctrl.edu/toolbox/index.htm>.
- Muijtjens, A. M. M., van Mameren, H., Hoogenboom, R. J. I., Evers, J. L. H., & van der Vleuten, C. P. M. (1999). The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Medical Education*, 33(4), 267–275.
- Muijtjens, A. M.M., Timmermans, I., Donkers, J., Peperkamp, R., Medema, H., Cohen-Schotanus, J., et al. (2010). Flexible electronic feedback using the virtues of progress testing. *Medical Teacher*, 32(6), 491–495.
- Mulder, R.V. de, & Noortwijk, C. van (2000). Beoordelen van klassieke open vraag- en essaytentamens. Computerondersteund nakijken van open vragen. In G.W.H. Heijnen, & S. Meeder (Eds.) *Toetsen en ICT in het hoger onderwijs. Stand van zaken en trends in Nederland*. Utrecht: SURF-Educatie. (► [www.edusite.nl/toetsdag/toetsbundel.pdf](http://www.edusite.nl/toetsdag/toetsbundel.pdf)).
- Mulder, M. (in druk) Conceptions of professional competence. In S. Billett, C. Harteis, H. Gruber (Eds.) *International handbook on research into professional and practice-based Learning*. Dordrecht: Springer.
- Mulder, M., Gulikers, J., Biemans, H., & Wesselink, R. (2009). The new competence concept in higher education: error or enrichment? *Journal of European Industrial Training*, 33, 755–770.
- Nedermeijer, J., & Pilot, A. (2000). *Beroepscompetenties en academische vorming in het Hoger Onderwijs*. Groningen: Wolters-Noordhoff.
- Newby, L., & Winterbottom, M. (2011). Can research homework provide a vehicle for assessment for learning in science lessons?. *Educational Review*, 63(3), 275–290.
- Nijhof, W.J. (2006). *Het leerpotentieel van de werkplek (afscheidsrede)*. Enschede: Universiteit Twente.
- Nitko, A.J., & Brookhart, S.M. (2007). *Educational Assessment of students*, 5th Ed. Upper Saddle River (NJ): Pearson Education Inc.
- NVAO (2011). *Beoordelingskaders accreditatiestelsel hoger onderwijs Opleidingsbeoordeling (beperkt)*. ► [www.nvao.net/beoordelingskaders](http://www.nvao.net/beoordelingskaders) (bekeken op 1-6-2013).
- Onderwijsraad (2004). *Examinering in het hoger onderwijs*. Den Haag: Onderwijsraad.
- Oost, H. (2004). *Een onderzoek begeleiden*. Amersfoort: Thieme Meulenhoff.
- Palm, T. (2008). *Performance Assessment and Authentic Assessment: A Conceptual Analysis of the Literature*. Practical Assessment, Research & Evaluation, 13, 4. ► <http://pareonline.net/getvn.asp?v=13&n=4>.
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144.
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or Fundamental? *Theory into Practice*, 48, 1–11.
- Reenalda, M. (2009). *Effecten van dualisering in het HBO (proefschrift)*. Enschede: Universiteit Twente.
- Rexwinkel, G.B., Haenen, J.P.P., & Pilot, A. (2012). Evaluating the level of degree programmes in higher education. *Conceptual Design. Assessment and Evaluation in Higher Education*. DOI: 10.1080/02602938.2012.755150.
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249–255.
- Roelofs, E. (2006). Een procesmodel voor de beoordeling van competent handelen. *Tijdschrift voor Hoger Onderwijs*, 24, 152–167.
- Romme, S., & Nijhuis, J. (2000). *Samenwerkend Leren in Afstudeerkringen*. Groningen: Wolters-Noordhoff.

- Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *J Educ Meas*, 14, 15–22.
- Ruijter, C.T.A., Bruggencate, G.C. ten, Spil, A.A.M., & Wouw, H.M.W.J. van der (1995). Beoordelen in projectonderwijs. Enschede: Universiteit Twente, Onderwijskundig Centrum.
- Ruiz-Primo, M.A., & Shavelson, R.J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, 33, 1045–1063.
- Ruiz-Primo, M. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies In Educational Evaluation*, 37(1), 15–24.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78.
- Sadler, R. D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144.
- Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist*, 24(2), 113–142.
- Schön, D. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.
- Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": An exploratory study of the consequential validity of assessment. *Studies in Educational Evaluation*, 23, 349–371.
- Schlusmans, K., Slotman, R., Nagtegaal, C., & Kinkhorst, G. (Eds.) (1999). *Competentiegerichte leeromgevingen*. Utrecht: Lemma.
- Schmidt, H., & Moust, J. (1998). *Probleemgestuurd onderwijs. Praktijk en theorie*. Groningen: Wolters-Noordhoff.
- Schoemaker, P.J.H. (1995). Scenario Planning: a Tool for Strategic Thinking. *Sloan Management Review*, 25–39.
- Schön, D. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.
- Servicecentrum Limburg Competent (2012). EVC. ► [www.limburgcompetent.nl](http://www.limburgcompetent.nl).
- Shavelson, R.J., Young, D.B., Ayala, C.C., Brandon, P., Furtak, E.M., Ruiz-Primo, M.A., Tomita, M., & Yin, Y. (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.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.
- Sluijsmans, D. M. A. (2013). Verankerd in leren: Vijf bouwstenen voor professioneel beoordelen in het hoger beroepsonderwijs. Heerlen: Zuyd Hogeschool.
- Sluijsmans, D. M. A., Joosten-ten Brinke, D., van der Vleuten, C. P. M. (2013). Toetsen met leerwaarde. Een reviewstudie naar de effectieve kenmerken van formatief toetsen. Den Haag: NWO.
- Slusser, S. R., & Erickson, R. J. (2006). Group Quizzes: An Extension of the Collaborative Learning Process. *Teaching Sociology*, 34(3), 249–262. doi:10.1177/0092055x0603400304.
- Stapert, R.S. (2012). Beoordelingskader eindwerkstuk HBO- bachelor master. Hobéon Certificering, september 2012.
- Stiggins, R.J. (1987). NCME Instructional module on Design and development of performance assessments. *Educational Measurement: Issues and Practice*, 6, 3, 33–42.
- Stiggins, R.J. (1992). Het ontwerpen en ontwikkelen van performance-assessment-toetsen. In J.W.M. Kessels, & C.A. Smit (Eds.) *Opleiders in Organisaties*, Kluwer Bedrijfswetenschappen. *Capita Selecta*, afl. 10.
- Stokking, K. Schaaf, van der, M., Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British educational research journal*, 30 (1), p. 93–116.
- Straetmans, G.J.J.M. (1995). Assessment in opleiding en onderwijs met competentietoetsen. In F.J.R.C. Dochy, & T.R. Rijke (Eds.) *Assessmentcenters. Nieuwe toepassingen in opleiding en onderwijs*. Utrecht: Lemma.
- Straetmans, G.J.J.M. (1998). Toetsing van competenties. In *Handboek effectief opleiden*. Den Haag: Delwel.
- Straetmans, G.J.J.M. (2006). *Bekwaam Beoordelen en Beslissen. Beoordelen in competentiegerichte beroepsopleidingen*. Lectorale Rede. Deventer: Saxion.
- Straetmans, G.J.J.M., Roelofs, E., & Peters, M. (2011). Vaststellen van didactische bekwaamheid tijdens de LIO-stage. *Tijdschrift voor Lerarenopleiders (VELON / VELOV)*, 32, 1, 4–11.
- Straetmans, G.J.J.M. (in voorbereiding). Beoordelen van competentie. In Sanders, P.F. (Red.), *Toetsen op School: Hoger Onderwijs*. ► [www.cito.nl](http://www.cito.nl).
- Streumer, J.N. (red) (2010). *De kracht van werkplekleren*. Utrecht: Boom Lemma.
- Streumer, J., & Van der Klink, M. (2004). *Leren op de werkplek*. 's Gravenhage: Reed Elsevier Business.
- Stuurgroep EVC (2013). *Validering van leerresultaten uit non-formeel en informeel leren. Voorstel voor een toekomstbestendig valideringssysteem*. Utrecht: Kenniscentrum EVC.
- Swanson, D.B., Norcini, J.J., & Grosso, L.J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education*, 12, 220–246.

- Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal*, 27(5), 615–631.
- Tijdschrift voor Hoger Onderwijs (2013). 31(1-2) Themanummer praktijkgericht onderzoek.
- Verhoeven, N. (2011) Wat is onderzoek? Praktijkboek methoden en technieken voor het hoger onderwijs. Boom Lemma uitgevers, Den Haag (vierde druk).
- Verschuren, P (2010). Onderzoek in het HBO; soorten, scholingsniveaus en methodologische profielen. Radboud Universiteit, Nijmegen.
- Verschuren, P. (2011, juni 27). Praktijkgericht onderzoek in het HBO: welk niveau is toereikend? Opgeroepen april 2013, van ► [www.han.nl](http://www.han.nl).
- Vleuten, C.P.M. van der, Besseling, C.W.M., Tromp, J.H.M., & Verbraeck, A. (1986). Voortgangsevaluatie als middel voor kwaliteitsbewaking. In N.A.J. Lagerweij (Ed.): *Onderwijsverbetering als opdracht*. Lisse: Swets & Zeitlinger.
- Vleuten, C.P.M. van der, Verwijnen, G.M., & Wijnen, W.H.F.W. (1996). Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher*, 18, 103–109.
- Vleuten, C.P.M. van der, & Driessen, E.W. (2000). *Toetsing in probleemgestuurd onderwijs*. Groningen: Wolters Noordhoff.
- Vleuten, C.P.M. van der (1995). Toetsing van medische competentie. In J.C. Metz, A.J.J.A. Scherpbier, & C.P.M. van der Vleuten (Eds.) *Medisch onderwijs in de praktijk*. Assen: Van Gorcum.
- Vleuten, C.P.M. van der, Schuwirth, L. T. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice & Research Clinical Obstetrics and Gynaecology*, 24, 703–719.
- Vleuten, C.P.M. van der, Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., et al. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34, 205–214.
- Vliet, R.G. van (z.j.). *Docentenvademecum*. Website van het Onderwijs Service Centrum Technische Universiteit Eindhoven. Geraadpleegd op het world wide web (14-03-2001): ► [www.osc.tue.nl/vademecum/mainvadem.htm](http://www.osc.tue.nl/vademecum/mainvadem.htm).
- Wainer, H. (2000). *Computerized adaptive testing : a primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Walliman, N. (2001 of 2005). *Your Research Project*. London: Sage.
- Walraven, I. (1994). *Problemen oplossen*. Utrecht: Lemma.
- Watkins, D., Dahlin, B., & Ekholm, M. (2005). Awareness of the backwash effect of assessment: A phenomenographic study of the views of Hong Kong and Swedish lecturers. *Instructional Science*, 33, 283–309.
- Webb, Eugene J., Donald T. Campbell, Richard D. Swartz, Lee B. Sechrest (1966). *Unobtrusive Measures; non-reactive research in the social sciences*. Chicago: Rand McNally.
- Wet op het hoger onderwijs en wetenschappelijk onderzoek (WHW). (1992). ► <http://wetten.overheid.nl> (bekeken op 1-6-2013).
- Wet Versterking Besturing (wijzigingswet WHW). (2010) Nr. 119 Staatsblad ► <https://zoek.officielebekendmakingen.nl> (bekeken op 1-6-2013).
- Wetenschappelijke Raad voor het Regeringsbeleid (WRR)(1995). *Hoger Onderwijs in Fasen*. Den Haag: SDU. (WRR-rapport 47).
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37, 3–14.
- Wijnen, W.H.F.W. (1971). *Onder of boven de maat*. Lisse, Swets en Zeitlinger.
- Wijnen, W.H.F.W., & Van der Vleuten, C.P.M. (1985). Toetsing: hordeloop of voortgangscontrole? *Universiteit en Hogeschool*, 31, 270–279.
- Wijnen, W.H.F.W. (1993). Waarden en normeren van beoordelingen. In H.J.M. van Berkel en A.E. Bax (Eds.) *Beoordelen in het Onderwijs*. Houten: Bohn Stafleu Van Loghum.
- Wijngaarden, P. van (2008). *Over de brug. Discours over werken, onderzoeken en leren in het duaal wetenschappelijk onderwijs*. Rotterdam: Performna.
- Wingate, U. (2010). The impact of formative feedback on the development of academic writing. *Assessment & Evaluation in Higher Education*, 35(5), 519–533.
- Wolters, L. (1998). *Toetsvormen voor academische vaardigheden*. Leiden: Universiteit Leiden / ICLON. (ICLON-rapport 59)
- Wrigley, W., Van der Vleuten, C. P. M., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher*, 34, 683–697.

**Websites**

- ▶ <http://bit.ly/QM-example-test-Use-IE-to-view2>.
- ▶ <http://course1.winona.edu/shatfield/air/rubrics.htm> voor een verzameling voorbeelden van 'academic rubrics'.
- ▶ [http://en.wikipedia.org/wiki/Digital\\_badges](http://en.wikipedia.org/wiki/Digital_badges).
- ▶ <http://plagiarism.bloomfieldmedia.com/z-wordpress/software/wcopyfind>.
- ▶ <http://plagiarism.phys.virginia.edu/Wsoftware.html>. (geraadpleegd op 2 februari 2006).
- ▶ <http://rubistar.4teachers.org>.
- ▶ <http://tlt.its.psu.edu/plagiarism/links>.
- ▶ <http://www.celt.iastate.edu/pdfs-docs/teaching/RevisedBloomsHandout.pdf>.
- ▶ [http://www.infed.org/archives/e-texts/colley\\_informal\\_learning.htm](http://www.infed.org/archives/e-texts/colley_informal_learning.htm).
- ▶ <http://www.kenniscentrumevc.nl>.
- ▶ <http://www.limburgcompetent.nl/evc>.
- ▶ <http://www.rijksoverheid.nl/hoger-onderwijs>.
- ▶ <http://www.10voorde.leraar.nl>.
- ▶ [http://www.umuc.edu/distance/odell/cip/links\\_plagiarism.html](http://www.umuc.edu/distance/odell/cip/links_plagiarism.html).
- ▶ <http://www.vandermaesenkoch.nl/ecartoon/nl>.

# Register



A

B

C

De vet gedrukte paginanummers verwijzen naar een definitie van het begrip.

360 graden feedback 191, 235

## A

absoluut normeren 76  
 accreditatie 30, 31, 98, 158, 212, 215, 226, 244  
 adaptief toetsen 94  
 afleider 9, 25, 114, 115  
 afstudeerkring 270  
 alternatieven 19, 97, 104, 114  
 alternative assessment 182  
 analytisch beoordelen 196, 245, 265, 279  
 antwoordmodel 11, 96, 104, 110, 133, 194  
 arbeidsproof 257  
 assessmentcenter 258  
 assessmentprogramma ► toetsprogramma 188  
 authentiek 60, 144, 184, 190, 210, 221, 224, 259, 261

## B

beantwoordingstijd 118, 131  
 begripsvaliditeit 17  
 beoordelaarseffect 133, 139, 216, 248, 264, 285  
 beoordelingscriterium 12, 59, 84, 99, 102, 187, 223, 224, 226, 233, 265  
 beoordelingsformulier 35, 266  
 beoordelingsinstrument 180, 206, 234, 238  
 beoordelingskader 30, 244  
 beoordelingsvoorschrift 18, 278  
 betoogvragen ► opstelvragen 131  
 betrouwbaarheid 18, 20, 23, 47, 58, 59, 95, 115, 117, 118, 133, 139, 140, 144, 161, 176, 187, 197, 212, 216, 234, 248, 256, 261, 264, 277, 284  
 bezwaar en beroep 50, 115, 162  
 bias 58, 138, 178, 192  
 blauwdruk ► specificatietabel 7  
 borging 37, 209  
 bruikbaarheid 17, 102, 165, 245, 256

## C

casus 8, 94, 109, 188, 222, 246, 259  
 casustoetsen 144  
 cesuur 12, 33, 50, 74, 97, 140, 163, 176, 200  
 checklist 127, 133, 140, 265, 282, 283  
 cijfergeven 3, 78, 200  
 classificatie ► taxonomie 6  
 College van Beroep voor de Examinens 115, 165  
 competentie 8, 48, 83, 152, 158, 180, 184, 185, 187, 188, 190, 206, 216, 230, 252, 261  
 compromismethode bij beoordelen 77  
 conceptmap 102  
 constructieregel 117, 134, 153  
 constructieve alignment 223  
 contaminatie-effect 139, 140  
 correctievoorschrift 133  
 creatief toetsen 102  
 criteriumgericht interview 58, 210, 246  
 criteriumvaliditeit 17

## D

digitaal toetsen 92  
 documentanalyse 247  
 doelmatigheid 115, 158, 270  
 Dublin-descriptoren 36, 213, 218, 244

## E

EVC 206  
 examencommissie 13, 30, 37, 50, 72  
 examiner 38, 190, 268, 277  
 externe validering 42

## F

feedback 5, 40, 56, 61, 84, 88, 95, 98, 144, 160, 198, 223, 234, 235, 248, 249, 264, 276, 279  
 formatief toetsen 82, 84, 95, 228  
 formula scoring 173  
 functie van toetsen 5, 16, 56, 95, 158, 199, 256, 261

## G

genade-zesje 159  
 gesloten toetsvorm 8  
 gesloten vragen 20, 94, 114  
 groepswork 87, 274

## H

halo-effect 55, 98, 139, 140, 163, 264  
 hink-stap-sprong 109  
 holistisch beoordelen 196, 245, 265, 279  
 horn-effect 98, 139, 140

## I

in- en aanvulvragen 130  
 inhoudsvaliditeit 7, 17, 133, 160  
 interview 103, 210, 246  
 intrabeoordelaarsbetrouwbaarheid 248  
 itembank 46, 96, 172

## J

ja/nee-vragen 114, 116  
 juist/onjuist-vragen 114–116

## K

korte-antwoordvragen 130  
 kwaliteitsbewaking 33, 179  
 kwaliteitscriterium ► kwaliteits-eis 37  
 kwaliteitseis 16, 58, 160, 188, 201, 211, 283

## L

landelijke kennistoets 42, 177  
 lange-antwoordvragen 130

**M**

meerkeuzevragen 19, 25, 78, 114, 147, 171  
 methodemix 238, 261  
 mondeling 103, 109, 158, 238, 257

**N**

nakijken 2, 11, 18, 97, 98, 140, 285  
 natuurgetrouwheid 184  
 Nederlands-Vlaamse Accreditatie  
 Organisatie (NVAO) 30  
 normeren 74  
 normverschuiving 139, 140  
 NVAO ► Nederlands-Vlaamse  
 Accreditatie Organisatie 30, 215

**O**

objectiviteit 10, 18, 45, 161, 277, 284  
 observatie 246  
 onderscheidingsvermogen 10, 16, 20, 22  
 ontwerp 13, 57, 82, 96, 170, 187, 215, 265  
 ontwikkelingsgesprek 236  
 open boektoets 110  
 open vragen 97  
 opstelvragen 131

**P**

PDCA-cyclus 31  
 peer review 103  
 peer-assessment 87, 223, 258, 277, 278, 282, 285  
 performance assessment 182, 183, 223, 237  
 Persoonlijk Ontwikkelingplan (POP) 236  
 piramide van Miller 8, 54, 221  
 plagiaat 66, 98  
 POP ► Persoonlijk Ontwikkelingplan 236  
 presentatie 105  
 proces 183, 186, 227, 243, 268, 276, 278, 281  
 product 183, 243, 268, 274, 276, 281  
 psychometrische analyse 16, 19, 45, 94  
 p-waarde 20

**R**

relatief normeren 74  
 relevantie 10, 19, 39, 45, 106, 141, 245  
 representativiteit 7, 19, 27, 118, 160, 184, 283  
 reproductie 7, 119, 183  
 restriction-of-range 139  
 rollenspel 107  
 rubric-applicatie 99  
 rubrics 88, 99, 191, 194, 227, 247

**S**

SafeAssign 69  
 samenwerken 87, 97, 256, 270, 274  
 scoring 14, 146, 173, 178, 197  
 script concordance test 259  
 scriptie 264, 268, 272  
 selectie 5, 21, 59, 88, 95, 188, 256, 279  
 self-assessment ► zelfbeoordeling 20  
 sequentie-effect 139, 140  
 signifisch effect 139  
 sleutel 11, 23, 114, 146, 172  
 specificatietabel 2, 7, 16, 17, 39, 43, 44, 97, 117, 145, 155, 164  
 stageverslag 237, 264  
 summatief toetsen 4, 88, 199

**T**

taxonomie 6, 199, 255  
 toetsafname 11, 76, 77, 92, 172  
 toetsbeleid 30  
 toetscommissie 39, 61, 170, 179  
 toetsdialoog 85  
 toetsfolder 164, 167  
 toetskader 33  
 toetsmatrjjs ► specificatietabel 7  
 toetsproces 2, 16, 37, 87, 92  
 toetsprogramma 36, 52, 88, 220  
 toetservicesysteem 45  
 toetsvorm 2, 8, 52, 57, 96, 99, 117, 158, 167, 195, 221, 255, 256, 261, 278  
 transparantie 11, 18, 59, 88, 164, 207, 247, 256, 270, 285

**U**

uitvoerbaarheid 256

**V**

validiteit 17, 55, 58, 145, 187, 199, 234, 258, 260, 284  
 volledigheid 184  
 voorspellend 219  
 voortgangstoets 170  
 vragenbank ► itembank 97  
 vragenlijst 104

**W**

WCOPYFIND 70  
 werkplekleren 230  
 werkstuk 264  
 werkwoord 119, 123  
 WHW ► Wet op het Hoger  
 Onderwijs en Wetenschappelijk  
 onderzoek 37, 211

**Z**

zak/slaaggrens ► cesuur 3  
 zelfbeoordeling 86, 198, 200, 223, 258, 266, 281, 282