Toetsbeleid en toetsrichtlijnen

2014-2018

Onderwijsinstituut Psychologie en KI

Radboud Universiteit Nijmegen

Totstandkoming

© Aan deze beleidsnotitie van het Onderwijsinstituut Psychologie en Kunstmatige Intelligentie hebben in de periode 2012-14 velen bijgedragen, waaronder het voormalige Tentamenteam Psychologie (inmiddels het Kwaliteitszorgteam). Conceptversies van dit rapport zijn besproken in het bachelor- en masteroverleg Psychologie en in het stafoverleg van Kunstmatige Intelligentie en zijn tevens gebruikt in de visitaties in de periode 2012/14. Eind 2013 hebben de examencommissies geadviseerd over een pre-final versie en zich uitgesproken voor één toetsbeleid voor het hele onderwijsinstituut, met waar nodig opleidingsspecifieke aanvullingen of uitzonderingen. Het toetsbeleid is vastgesteld door de onderwijsdirecteur op 15 maart 2014.

Toetsbeleid en toetsrichtlijnen 2014-2018

Onderwijsinstituut Psychologie en KI

Inhoudsopgave

Voorwoord

De laatste jaren hebben we binnen het onderwijsinstituut al veel verbeterd aan onze toetsing en beoordeling. Dit blijkt o.a. uit het volgende:

- Docenten variëren de toetsvormen door b.v. multiple-choice tentamens te combineren met open vragen en met tussentijdse werkstukken
- Docenten laten meerkeuze-toetsen vooraf door het kwaliteitszorgteam screenen en laten achteraf item-analyses uitvoeren door ITS
- Stage- en thesiscoördinatoren verbeteren de beoordelingsprotocollen voor stages en scripties
- De tweede beoordelaar beoordeelt de thesis onafhankelijk van de eerste beoordelaar ('blind vier ogen'-principe)
- De Examencommissies onderzoeken de kwaliteit van toetsen en van thesis-beoordelingen.

We hebben dus belangrijke stappen gezet. Er is echter meer nodig. Onze inspanningen tot dusverre zijn vooral gericht geweest op verbetering van de toetsing per programmaonderdeel. Aanleidingen om nu ook op curriculumniveau de toetsing beter te doordenken zijn:

- NVAO vereist een beschrijving van ons systeem van toetsing en beoordeling waarmee we kunnen aantonen dat onze studenten de beoogde eindkwalificaties realiseren[1].
- Het CvB vraagt in *Plan van aanpak Toetsing en Beoordeling, Radboud Universiteit, oktober 2013* aan alle opleidingen om hun toetsbeleid en toetsprogramma voor 1 maart 2014 te expliciteren.
- Opleidings- en leerlijncoördinatoren missen de 'tools' om binnen een opleiding, leerlijn of opleidingsjaar te sturen op samenhang en opbouw in toetsing.
- De druk op studenten neemt toe en daarmee hebben toetsen en beoordelingen grotere consequenties dan voorheen. Dit leidt ondermeer tot klachten van studenten over de tentamens.

Met het voor u liggende *Toetsbeleid en toetsrichtlijnen, 2014-2018 Onderwijsinstituut Psychologie en KI* willen we in dit 'meer' voorzien. In dit eerste hoofdstuk beschrijven we het toetsbeleid dat van toepassing is op de zes opleidingen[2]. Dit deel is het meest relevant voor hoofden van opleidingen en jaar- of leerlijncoördinatoren. Docenten/examinatoren dienen echter wel bekend te zijn met de hoofdlijnen hiervan.

In het tweede hoofdstuk beschrijven we concrete richtlijnen, vooral bedoeld voor docenten/examinatoren om hen te ondersteunen om het toetsbeleid in praktijk te brengen.

Voor vragen over het toetsbeleid of over de richtlijnen kunt u terecht bij het Kwaliteitszorgteam, kwaliteitszorg@psych.ru.nl.

[1] NVAO (2011). Beperkte opleidingsbeoordeling, p. 7.

[2] Bachelor Psychologie; Bachelor Kunstmatige Intelligentie; Master Psychologie; Master Artificial Intelligence; Research Master Behavioural Science; Research Master Cognitive Neuroscience.

Hoofdstuk 1: Toetsbeleid

1.1 Doelen en afstemming

Het toetsbeleid heeft als *doel* om (a) de gewenste systematiek van toetsing en beoordeling te beschrijven waarmee de opleidingen waarborgen dat studenten de eindkwalificaties realiseren; (b) studenten overzicht te bieden van de toetssystematiek en daarmee houvast voor het (bij)sturen van hun leeractiviteiten; (c) docenten en examinatoren te helpen om verantwoorde beslissingen te nemen rondom toetsing en beoordeling en (d) een kader te bieden voor evaluatie en eventuele bijsturing van de toetskwaliteit.

Het toetsbeleid staat niet op zichzelf maar dient *afgestemd* te zijn op:

- Opleidingsvisies van de zes opleidingen van het onderwijsinstituut
- Kwaliteitszorgbeleid van het onderwijsinstituut en het Kwaliteitshandboek RU (versie 3, januari 2013)
- Model Regels & Richtlijnen Examencommissies RU, d.d. 14 november 2011
- OER van de opleidingen.

Daarnaast is het toetsbeleid zoveel mogelijk gebaseerd op onderwijskundig en didactisch onderzoek, op evaluaties binnen de opleidingen en op *good practices* binnen de opleidingen, de RU en binnen andere universiteiten.

1.2 Richtinggevende principes

In deze paragraaf schetsen we vier breed gehanteerde uitgangspunten ten aanzien van toetsing en beoordeling en bespreken we de vijf functies die bij toetsing en beoordeling worden onderscheiden. Verder beschrijven we beknopt de drie meest relevante kwaliteitscriteria voor toetsing.

In paragraaf 1.3 vertalen we deze uitgangspunten, functies en kwaliteitscriteria naar concrete vuistregels voor de systematiek van toetsing en beoordeling.

1.2.1 Uitgangspunten

1. *Een verantwoord curriculum bestaat uit didactisch consistente cursussen die zijn afgestemd op de eindkwalificaties*

Het curriculum/programma dient zo in elkaar te steken dat studenten die tot de opleiding zijn toegelaten in staat zijn binnen de nominale studietijd de eindkwalificaties te bereiken. Biggs introduceerde voor de relatie tussen programma en eindkwalificaties de term *alignment*[3]. We vertalen dit als didactische consistentie [4].

[3] Biggs, J. (1999). Assessing for learning quality: Buckingham: SRHE and Open University Press. Verkregen van: http://teaching.polyu.edu.hk/datafiles/R131.pdf

[4] Ontleend aan Huisman, W. (2012). Didactische consistentie: zelfstudiemateriaal voor docenten. Verkregen van http://www.iowo.nl/icto/elem/63/

Vanuit curriculumperspectief dient het programma ten eerste een goede opbouw te vormen naar de eindkwalificaties (*verticale samenhang*). Waar gewenst worden hiervoor leerlijnen onderscheiden. Ten tweede dient het programma *per opleidingsjaar* samen te hangen. Deze *horizontale samenhang* uit zich in inhoudelijke integratie en in door het docententeam 'gedeelde' opvattingen over het *academisch niveau* dat in opleidingsjaar x wenselijk en haalbaar is.

Ook *per cursus* dient er sprake te zijn van consistentie: de cursusdoelen zijn afgeleid van de eindkwalificaties: ze passen én bij het jaarniveau én bij de leerlijn. Het geheel van colleges, literatuur, discussie, (zelfstudie)opdrachten, feedback, tussentijdse toetsen en eindbeoordeling is erop gericht dat studenten de cursusdoelen bereiken.

2. Leerdoelen, toetsing en beoordeling sturen het leerproces

Uit onderzoek blijkt dat de leeractiviteiten van studenten grotendeels worden gestuurd door de toetsing, en niet enkel door het onderwijsprogramma[5]. Studenten maken een inschatting van het gewenste toetsresultaat ("Moeten we dat weten voor de toets?") en passen daarop hun leeractiviteiten aan[6]. Wanneer een toets de vakkennis en de leerdoelen oppervlakkig toetst (bijvoorbeeld: het reproduceren van kennis in plaats van het zelf leggen van verbanden; of de in de hoorcolleges behandelde stof zonder de in de werkgroepen geoefende vaardigheden), dan zal de student oppervlakkig leren, ook al tracht de docent als maar te motiveren tot meer diepgaande bestudering.

Om *deep learning*[7] te stimuleren is het dus noodzakelijk dat de toetsen gaan over "wat we echt belangrijk vinden" en over hetgeen "ze eigenlijk moeten leren". Dit betekent dat docenten al voor het begin van de cursus nadenken over de toetsing: niet als sluitstuk van de cursus maar als fundament.

3. Een consistent toetsprogramma vertoont horizontale en verticale samenhang

Vanuit *curriculumperspectief* dient het toetsprogramma (= geheel van toetsen) ten eerste *per leerlijn* een goede opbouw te vormen naar de eindkwalificaties (verticale samenhang). Ten tweede dient het toetsprogramma *per opleidingsjaar* samen te hangen (horizontale samenhang). De verticale en horizontale samenhang vanuit curriculumperspectief zijn weergegeven in figuur 1.

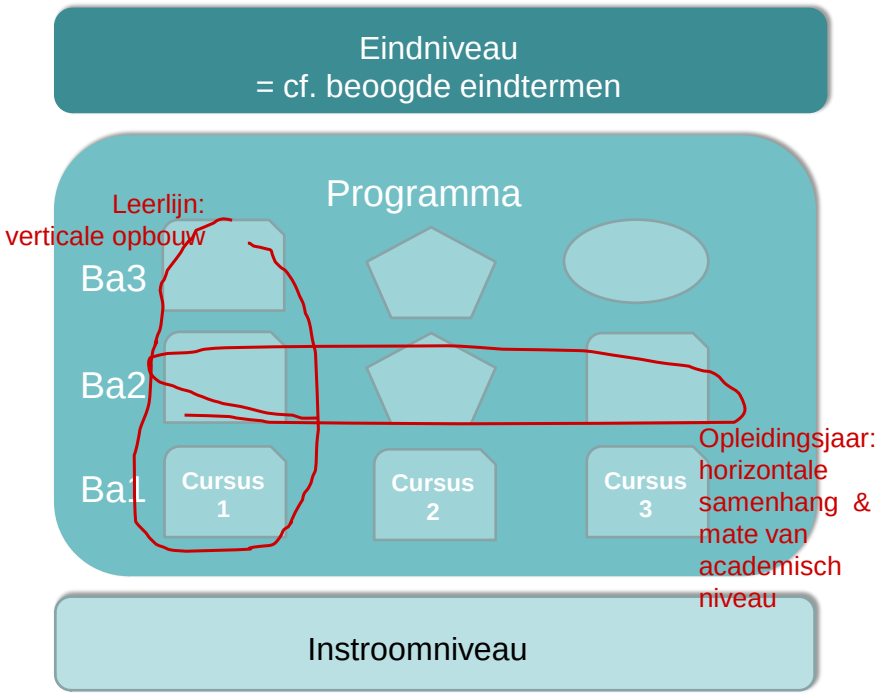
Vanuit *cursusperspectief* moet de toets een goede afspiegeling zijn van de leerdoelen van de cursus. In een consistente cursus zijn vervolgens ook de tussentijdse leertaken/ opdrachten een afspiegeling van de leerdoelen en dus van de toets. Daarmee is dus heel het cursusontwerp én leerdoel- én toetsgericht. Deze samenhangen worden geïllustreerd in figuur 2.

---

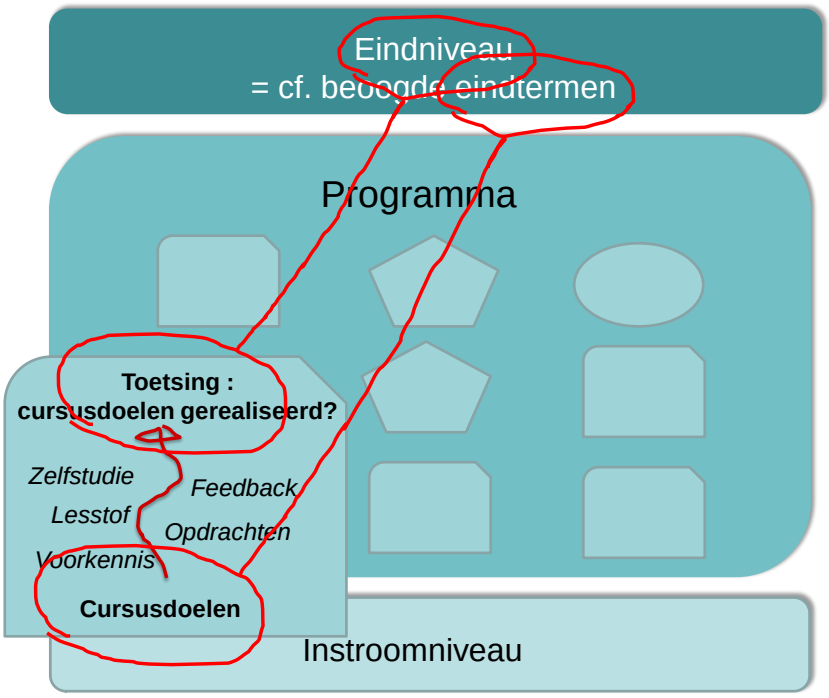
[5] Ramsden, P. (1992). *Learning to teach in higher education.* London: Routledge; Van der Vleuten, C.P.M. (1997). De intuïtie voorbij. *Tijdschrift voor Hoger Onderwijs, 15*(1), 34-46; Cilliers, F.J., Schuwirth, L.W., Adendorff, H.J., Herman, N., & van der Vleuten, C.P.M. (2010). The mechanism of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education, 15*(5), 695-715.

[6] Elton, L. (1987). *Teaching in higher education: Appraisal and training.* London: Kogan Page.

[7] Biggs, J. (1999). Assessing for learning quality: II. Practice. In: *Teaching for Quality Learning at University* (pp. 165-203). Buckingham: SRHE and Open University Press. Verkregen van: http://teaching.polyu.edu.hk/datafiles/R131.pdf.



Figuur 1*: Didactische consistentie vanuit curriculumperspectief.*



Figuur 2: *Didactische consistentie vanuit cursusperspectief.*

4. Het toetsprogramma gaat efficiënt om met de beschikbare tijd en middelen
Een consistent toetsprogramma ontwikkelen is geen kunst als tijd en geld onbeperkt zouden zijn. Maar dat is uiteraard niet het geval. Vandaar het vierde uitgangspunt: uitvoerbaarheid (efficiency), d.w.z. het toetsprogramma moet te realiseren zijn met de hoeveelheid tijd en geld die de opleidingen ter beschikking staat en moet de beschikbare middelen efficiënt gebruiken. Hierin schuilt misschien wel de grootste uitdaging voor het toetsbeleid.

1.2.2. Functies

Naast de hierboven beschreven vier uitgangspunten zijn in de literatuur vijf functies van toetsen en beoordelen terug te vinden die als belangrijke richtinggevers dienen voor toetsbeleid. Deze functies zijn:

1. *Verbetering of ontwikkeling*, ook wel formatief toetsen genoemd: de student ontvangt tussentijds feedback op over hetgeen al goed (genoeg) gaat en over hetgeen nog verbetering behoeft. Feedback is een krachtig middel om het leren van studenten te beïnvloeden[8]. Duidelijke en ontwikkelingsgerichte *feedback* stelt de student in staat zichzelf te verbeteren in een gewenste richting.
2. *Beoordeling*, ook wel summatief toetsen genoemd: op basis van vooraf bepaalde criteria wordt een oordeel onvoldoende of voldoende (in meerdere gradaties) uitgesproken. Er is sprake van een duidelijke cesuur (‘omslagpunt’) tussen onvoldoende en voldoende, of bekwaam en niet-bekwaam.
3. *Selectie*: een voldoende-beoordeling geeft toegang tot bijvoorbeeld een volgend programmaonderdeel, een vervolgopleiding of beroep.
4. *Kwalificatie*: de beoordeling voldoende levert een kwalificatie (diploma) op waaraan nadere rechten (titulatuur, registratie in beroepsregister etc.) zijn gekoppeld.
5. *Feedback inzake de onderwijskwaliteit*: de opleiding of docent maakt uit de toetsresultaten op in hoeverre het onderwijs adequaat is geweest.

1.2.3 Kwaliteitscriteria

In de onderzoeksliteratuur en in de kwaliteitszorg worden tot slot drie kwaliteitscriteria gehanteerd voor toetsing en beoordeling.

1. Validiteit
De belangrijkste kwaliteitseis is inhoudsvaliditeit, d.w.z. valide inzake de inhoud van de leerdoelen. Een valide toets meet wat deze moet meten. Zou dit niet het geval zijn, dan moeten we onze uitspraken over het presteren van de student in twijfel trekken. Validiteit is een noodzakelijke voorwaarde, die vooraf gaat aan betrouwbaarheid.

[8] Hattie, J. & H. Timperley (2007). The power of feedback. *Review of Educational Research*, vol. 77 (1), p. 81-112.

2. Betrouwbaarheid
Een betrouwbare toets is een toets die consistent meet. Deze consistentie moet hoog zijn, zodat er een adequaat oordeel over de prestaties van de student uitgesproken kan worden. Een toetsing is consistent als het resultaat in een andere situatie of op een ander tijdstip of door een andere beoordelaar hetzelfde zou zijn. Voor de kwaliteit van een toets is betrouwbaarheid een noodzakelijke maar geen voldoende voorwaarde.

3. Inzichtelijkheid
Een toets is voor de student inzichtelijk als hij/zij vooraf weet aan de hand van welke beoordelingscriteria hij/zij op welke wijze wordt beoordeeld en als de student zich kan voorstellen hoe deze toets bijdraagt aan de eigen professionele vakuitoefening.

1.3 Systeem van toetsing en beoordeling

In deze paragraaf vertalen we de hierboven beschreven richtinggevende principes naar concrete vuistregels voor de systematiek van toetsen en beoordelen van de zes opleidingen binnen het onderwijsinstituut. Deze vuistregels geven enerzijds duidelijke sturing; anderzijds laten zij voldoende ruimte voor docenten/ examinatoren om, gegeven de concrete situatie, een optimale invulling te kiezen.

Aan de orde komen zes vuistregels op curriculumniveau en elf vuistregels op cursusniveau. De vuistregels op curriculumniveau zijn met name bedoeld voor *opleidingscoördinatoren, leerlijn-/domeincoördinatoren en jaarcoördinatoren*. De vuistregels op cursusniveau zijn bedoeld voor *docenten/examinatoren.* Docenten/ examinatoren dienen echter wel op de hoogte te zijn van het geheel van vuistregels.

1.3.1 Zes vuistregels op curriculumniveau

Vuistregel 1: Het beoogd niveau per opleidingsjaar is geëxpliciteerd
Idealiter geeft het opleidingsprogramma en dus ook het toetsprogramma van de verschillende opleidingsjaren een opbouw tot de eindkwalificaties te zien. Deze opbouw zou geëxpliciteerd moeten zijn en bekend bij docenten en studenten. De vraag is of de huidige globale typeringen van het academisch niveau per opleidingsjaar (bijvoorbeeld bij Psychologie: B1: inleidend; B2: verbredend; B3: verdiepend; Ma: specialiserend en praktijkgericht; Resma: specialiserend en onderzoeksgericht) docenten voldoende houvast bieden voor de precieze niveau-bepaling van hun toetsen; en of ze studenten voldoende ‘richtinggevoel’ geven voor de academische ontwikkeling die zij geacht worden te realiseren. Binnen het onderwijsinstituut streven we ernaar om het beoogde academisch niveau per opleidingsjaar te expliciteren[9].

Vuistregel 2: Per opleidingsjaar een bij het beoogde niveau passende toetsmix
Idealiter kent de opleiding in B2 een andere combinatie en/of een andere weging van toetsvormen dan in B1, en in master weer anders dan in de bachelor etc. Zo zouden in de

[9]Hiervoor zijn verschillende modellen beschikbaar, zoals de taxonomie van Bloom (1956; Anderson & Krathwohl, 2001), de zogeheten piramide van Miller (1999) en de SOLO taxonomie van Biggs (1982; 2007).

hogere opleidingsjaren bijvoorbeeld meer werkstukken en (semi-)authentieke opdrachten gebruikt dienen te worden. In geval van meerdere toetsvormen per cursus zou in de loop der opleidingsjaren sprake kunnen zijn van een verandering in de weging tussen toetsvormen (voorbeeld: in B1 telt de kennistoetsing middels een meerkeuze-tentamen voor 70% mee in het eindcijfer en practicum-opdrachten voor 30%; in B2 zou de weging kunnen verschuiven naar 50/50%).

Vuistregel 3: Integratieve toetsing per opleidingsjaar
Het toetsprogramma dient ook horizontaal samen te hangen, d.w.z. *per opleidingsjaar*. De vuistregel luidt dat er per opleidingsjaar sprake is van enige vorm van integratieve toetsing, waarin studenten kennis en vaardigheden uit verschillende vakken/ leerlijnen integreren. Bijvoorbeeld: binnen de Ba Psychologie is de horizontale samenhang zichtbaar in de integratieve functie van de Kernthema's en van OP1, OP2 en OP3.

Vuistregel 4: Goede spreiding van tentamens
Een verdere randvoorwaarde is dat de roostering van de toetsing zodanig is, dat de student een goede prestatie kan leveren in de voorbereiding naar het tentamen. Dit betekent dat de werkdruk voor studenten goed verdeeld is over het onderwijsprogramma en dat de toetsingen elkaar niet overlappen. In de praktijk kan dit bijvoorbeeld betekenen dat een aantal semestervakken geen eindtoetsing kent, maar bestaat uit summatieve toetsopdrachten die al tijdens de cursus worden afgerond.

Vuistregel 5: Afweging tussen de kosten en de baten per opleidingsjaar
De ene toetsvorm is tijdsintensiever en dus duurder dan de andere. Bij het kiezen van het toetsmoment en de -vorm zouden de kosten afgewogen moeten worden tegen de inhoudelijke baten: weegt de benodigde investering (tijd, geld) op tegen de informatie die met de toets wordt verkregen? Toetsing van welke kennis, vaardigheden en attitude is ons op welk moment in de opleiding het meest waard? Voor het bewaken van een doelmatige inzet van middelen is het ondermeer van belang om:

- Na te gaan hoeveel tijd verschillende toetsmomenten en –vormen nu precies kosten. Voorbeeld: een veelgehoord argument is dat meerkeuzevragen minder tijd kosten. Dit heeft betrekking op het nakijken. Het maken, onderhouden en vernieuwen van valide en betrouwbare meerkeuzevragen vergt echter een flinke tijdsinvestering, vanwege de minimale hoeveelheid vragen die nodig is en vanwege de gewenste psychometrische kwaliteit.
- Na te gaan of er slimme en creatieve mogelijkheden zijn om docenttijd van de ene naar de andere docenttaak te 'realloceren', bijvoorbeeld door inzet van peer feedback en ICT.
- De meeste tijd en het meeste geld te investeren in de toetsmomenten die voor het behalen en beoordelen van het Ba- en Ma-eindkwalificaties het belangrijkste zijn.
- De inzet van tijd te spreiden over de verbeter- en de beoordelingsfunctie van toetsen, bijvoorbeeld door meer tijd te besteden aan feedback *tijdens* de cursus dan aan inzage *na* de cursus.

Vuistregel 6: Voldoende spreiding en opbouw in formatieve toetsing
De zesde vuistregel heeft specifiek betrekking op de formatieve toetsing en dus op de ontwikkelingsfunctie van toetsing (zie 1.2.2). Idealiter ontvangt elke student in *elke cursus minimaal één keer tussentijds ontwikkelingsgerichte feedback.* Formatieve toetsen dienen daarbij een betrouwbare afspiegeling te vormen van de summatieve toets. Omdat feedback geven echter tijdsintensief is, is dit wellicht niet in alle programmaonderdelen te realiseren. In die situaties geldt dat in een opleidingsjaar de feedback-momenten doelgericht worden gekozen en *gespreid* over het opleidingsjaar.

Verder is van belang dat ook de formatieve toetsing in de loop van de opleiding goed wordt *opgebouwd*. Bijvoorbeeld: de feedback door de docent wordt geleidelijk aangevuld met (en mogelijk deels vervangen door) peer feedback tussen studenten onderling en zelfreflectie door de student. Studenten maken zich zo de academische attitude steeds meer eigen door als het ware de 'stemmen van rolmodellen' te internaliseren en geleidelijk aan de eigen prestaties en ontwikkeling zelf te leren beoordelen. Het kunnen ontvangen, verwerken en geven van (peer) feedback en het ontwikkelen van zelfreflectief vermogen zijn daarmee in meerdere cursussen belangrijke (sub)leerdoelen.

1.3.2 Elf vuistregels op cursusniveau

Vuistregel 7: Validiteit van toets bewaken d.m.v. toetsmatrix
De vuistregels op curriculumniveau zijn voornamelijk gericht op het vergroten van de validiteit en zijn in die zin voorwaardelijk voor de validiteit op cursusniveau.

Op cursusniveau moet de toets representatief zijn voor de cognitieve leerdoelen (inhoudsvaliditeit) en dekkend voor wat betreft de metacognitieve leerdoelen van de cursus. Inhoudsvaliditeit betekent een representatieve steekproef van de leerstof; dat is iets anders dan "de toets moet de hele leerstof dekken" of "één vraag per hoofdstuk uit het handboek".

De leerdoelen dienen op hun beurt te passen bij het afgesproken jaar*niveau* en – voor zover de opleiding die onderscheidt - binnen één of meerdere *leerlijnen*. Om de validiteit van de toetsing inzichtelijk te maken is met ingang van het academisch jaar 2014-2015 een zogeheten *toetsmatrix* (of alternatieven daarvoor) verplicht (zie 2.2).

Vuistregel 8: Meerdere toetsmomenten en -vormen per cursus
De betrouwbaarheid van toetsing neemt toe naarmate er sprake is van meerdere en gevarieerde metingen binnen een cursus[10]. Als vuistregel hanteren we binnen het onderwijsinstituut dat elke cursus *minimaal twee toetsmomenten of -vormen* hanteert, in ieder geval vanaf B2. Dit stelt docenten beter in staat om de verscheidenheid van leerdoelen te toetsen; en het biedt studenten de kans om hun bekwaamheid middels gevarieerde metingen aan te tonen.

In de studiegids en (online) studiehandleiding dient duidelijk beschreven te zijn wat het relatieve gewicht is van elk toetsonderdeel voor het uiteindelijke cijfer. Deze *weging* moet in logische verhouding staan tot de (prioriteiten binnen de) leerdoelen en tot de geprogrammeerde tijdsinvestering van de student gericht op het betreffende leerdoel.

[10] Milius, J. (2007). *Schriftelijk tentamineren. Een draaiboek voor docenten in het hoger onderwijs.* Utrecht: Universiteit Utrecht (Ivlos).

Vuistregel 9: Toetsen worden jaarlijks substantieel 'ververst'
Het in identieke vorm hergebruiken van oude tentamens is uit den boze, aangezien deze via sociale media snel circuleren. Dit speelt met name bij meerkeuze- en open vragen tentamens. Tentamens dienen daarom jaarlijks substantieel vernieuwd te worden. Dat kan door een voldoende ruime vragen-pool te ontwikkelen, antwoordalternatieven te veranderen, door broertje/zusje-vragen te maken, door vragen en antwoordalternatieven husselen, door casuïstiek anders te gebruiken etc.

Vuistregel 10: Toetsvormen en weging worden vooraf bekend gemaakt aan studenten
Het criterium van inzichtelijkheid vereist dat studenten niet voor verrassingen komen te staan. Binnen het onderwijsinstituut kunnen studenten bij het maken van hun studieplanning nagaan (via de studiegids) welke toetsvormen en welke weging van de toetsonderdelen binnen een bepaalde cursus worden gehanteerd. Dit betekent dat de verantwoordelijk docent ervoor zorg draagt dat deze gegevens ten tijde van het maken van de studiegids (d.w.z. uiterlijk 1 mei voor het eerste semester en uiterlijk 1 december voor het tweede semester) bekend zijn; en dat deze daarna niet meer gewijzigd worden.

In geval van werkstukken, stage en scripties dienen ook de beoordelingscriteria voor aanvang van de cursus bekend te worden gemaakt. Dit gebeurt vlak voor het begin of aan het begin van de cursus via de (online) studiehandleiding.

Vuistregel 11: Studenten worden individueel beoordeeld
Studenten worden individueel beoordeeld. Ook in het geval van cursussen, stages of scripties waarbij studenten in een groep samenwerken zijn er manieren om studenten individueel te beoordelen. Bijvoorbeeld: eigen observaties van de docent, logboeken van studenten, mondelinge tentamens, mondelinge presentaties, en mondelinge en schriftelijke reflectie door individuele studenten.

Vuistregel 12: Beoordelaars hanteren (bij open vragen en werkstukken) een transparant beoordelingsmodel
Er zijn meerdere redenen om met beoordelingsmodellen te werken. Ten eerste om studenten in staat te stellen hun leeractiviteiten (bij) te sturen en zichzelf of *peers* van feedback te voorzien; ten tweede om de betrouwbaarheid te vergroten en interbeoordelaarsverschillen te verminderen. Een derde reden is dat op deze basis de psychometrische kwaliteit van de toets bewaakt kan worden.

De precieze vorm van het beoordelingsmodel verschilt per toetsvorm. In het geval van open vraag-tentamens is een nakijk-model nodig (zie 2.3.2); werkstukken, logboeken, reflectieverslagen, stages en scripties vergen een uitgebreider beoordelingsschema (of *rubric*[11]) (zie 2.3.3).

[11] Stevens, D.E. & Levi, A.J. (2013). Introduction to rubrics. An assessment tool to save grading time, convey effective feedback, and promote student learning (2nd edition). Sterling, Virginia: Stylus Publishing. Voor gratis online resources zie bijvoorbeeld: iRubric (http://www.rcampus.com/indexrubric.cfm). Ook Blackboard kent rubrics; zie Handleiding Blackboard.

Vuistregel 13: Beoordelaars hanteren (bij meerkeuze-tentamens) de methode van 'absolute beoordeling' of van 'absolute beoordeling met een relatieve referentie'

Binnen het onderwijsinstituut wordt de 'methode van relatieve beoordeling' (d.w.z. min of meer vaste percentages van het aantal studenten dat 'moet slagen' of 'moet zakken', ongeacht de prestaties van de studenten) niet wenselijk geacht. Cijfers worden bij voorkeur berekend middels de methode van 'absolute beoordeling'; is dit niet mogelijk dan wordt de methode 'absolute beoordeling met relatieve referentie' gebruikt[12]. Beide methoden worden verder uitgewerkt in paragraaf 2.3.1 en in bijlage II.

Vuistregel 14: Bij meerkeuzevragen-tentamens wordt gecorrigeerd voor gokken en wordt de psychometrische kwaliteit van de toets onderzocht .

Bij meerkeuzevragen-tentamens wordt er gecorrigeerd voor gokken. Dit wordt aangegeven in de toetsinstructies aan studenten. De implicatie hiervan is namelijk dat studenten een vraag die ze niet weten beter kunnen gokken dan openlaten, omdat ze anders dubbel worden 'gestraft'.

Psychometrische analyse van meerkeuze-tentamens is verplicht. Het ITS levert de docent hiertoe een zogeheten toetsrapport op basis waarvan de docenten kan besluiten hoe om te gaan met meerkeuzevragen van onvoldoende kwaliteit (zie ook 2.4.3).

Vuistregel 15: Toetsen en beoordelingen worden door een collega bekeken (peer review)

De betrouwbaarheid van toetsing wordt vergroot door de toets en de toetsmatrix aan een vakgenoot voor te leggen ('vier ogen principe'). Dat kan een docent zijn die betrokken is bij de cursus, of een deskundige op het gebied van toetsing, bijvoorbeeld het kwaliteitszorgteam. Zij kunnen met een 'frisse blik' bekijken of alle leerdoelen op het gewenste niveau aan de orde komen en wijzen op eventuele tekstuele onduidelijkheden (bijv. antwoordalternatieven die teveel op elkaar lijken etc.).

In het geval dat meerdere docenten toetsvragen aanleveren, is het van belang dat de eindverantwoordelijke docent over het eindproduct beslist en de validiteit, de betrouwbaarheid en de algehele moeilijkheidsgraad bewaakt.

Ook over de beoordelingen vindt *peer review* plaats, zeker wanneer de slaag- en zakpercentages opvallend afwijken van vorige jaren en/of wanneer besloten moet worden om van absolute beoordeling over te gaan op absolute beoordeling met relatieve component, of wanneer op grond van afwijkende scores in de psychometrische analyse besloten moet worden over het 'laten vervallen van vragen', het 'goedkeuren van meerdere alternatieven' of anderszinse 'bijstellingen van de normen'. In deze gevallen overlegt de eindverantwoordelijk docent met een onafhankelijke collega of bijvoorbeeld met de leerlijncoördinator.

[12] Gruijter, D.N.M. de (2008). *Toetsing en Toetsanalyse*. Leiden: Universiteit Leiden.

Vuistregel 16: Stagebeoordeling door interne beoordelaar; scriptiebeoordeling door twee beoordelaars op onafhankelijk wijze van elkaar

Stage- en thesehandleidingen zijn beschikbaar voor de start van de stage en these. Hierin zijn de beoordelingscriteria geëxpliciteerd - mede in hun relatie tot de eindkwalificaties – en is de weging en de vertaling naar een eindcijfer inzichtelijk. Vooraf dient duidelijk te zijn of stage en these integraal of afzonderlijk worden beoordeeld. Indien beide integraal worden beoordeeld dan dient vooraf bekend te zijn of het cijfer voor de stage gecompenseerd kan worden door het cijfer voor de these.

Bij stages en theses vindt er standaard een tussentijdse evaluatie/ beoordeling plaats. In de stage- en thesehandleidingen is duidelijk wie deze evaluatie/ beoordeling uitvoert en aan de hand van welke criteria. Tevens is geëxpliciteerd wat er gebeurt als het tussentijdse oordeel onvoldoende is.

De eindbeoordeling van stages is in handen van een stagebeoordelaar vanuit de opleiding; de externe stagebegeleider kan hiervoor informatie aandragen. De beoordelaar beoordeelt aan de hand van een standaard beoordelingsformulier.

Theses worden door twee beoordelaars onafhankelijk van elkaar beoordeeld (blind vier-ogen-principe) aan de hand van een standaard beoordelingsformulier. Vooraf dient duidelijk te zijn of alleen het product of product én proces worden beoordeeld. Indien beide worden beoordeeld dan dienen voor beide vooraf beoordelingscriteria bekend te zijn, plus hun relatieve bijdrage aan het eindcijfer.

Discrepanties tussen beoordelaars worden besproken. Indien geen consensus ontstaat, wordt een derde beoordelaar ingeschakeld die beslist. Discrepanties van 1,5 punt of meer en discrepanties tussen voldoende en onvoldoende worden geregistreerd en jaarlijks geanalyseerd door de coördinator. Deze analyse kan leiden tot nadere toelichting op de beoordelingscriteria of tot aanscherping ervan. Ook kan het aanleiding vormen tot een hernieuwd collegiaal gesprek tussen beoordelaars over de gezamenlijke visie op (interpretatie van) de beoordelingscriteria.

Om de scheiding tussen begeleiding en beoordeling bij scripties en theses duidelijk te markeren, wordt in de stage/scriptiehandleiding en/of in een stage/scriptieovereenkomst duidelijk omschreven hoeveel conceptversies de student kan inleveren, wanneer de deadline is voor de te beoordelen eindversie en wat de deadline is voor een eventuele herkansing.

Vuistregel 17: Examinatoren evalueren de toets en de beoordeling in het teacher report

Binnen zes weken na de eindtoets evalueert de verantwoordelijk docent/ examinator in het teacher report het proces van toetsing en beoordeling. Hierin wordt algemene informatie over de toets gegeven, zoals de toetsvormen, de weging, en het slagingspercentage na de eerste kans. Ook wordt aangegeven wat de sterktes en zwaktes zijn geweest met betrekking tot de toetsing en wat eventuele oplossingen hiervoor zijn. Het teacher report, inclusief de toetsmatrix, wordt besproken in de OLC.

Afhankelijk van de evaluatie in het teacher report wordt waar nodig het toetsontwerp bijgesteld. In het geval van toetsen met gesloten of open vragen is het raadzaam de psychometrische gegevens in combinatie met de vraag vast te leggen in een database en te bewaren voor volgende toetsingen.

1.4 Verantwoordelijkheden

Om de kwaliteit van het toetsings- en beoordelingsproces te borgen, dient de organisatie van de toetsing optimaal geregeld te zijn. Daarvoor is transparantie in verantwoordelijkheden nodig. We streven naar een toetscultuur, waarin ieder de eigen verantwoordelijkheden op zich neemt en we elkaar daarop kunnen aanspreken.

In tabel 1 wordt aangegeven wie op welk niveau verantwoordelijk is voor de didactische consistentie tussen toets en onderwijs en daarmee voor de validiteit, betrouwbaarheid en inzichtelijkheid van de toetsing en beoordeling.

Tabel 1. *Verantwoordelijken inzake toetsing en beoordeling binnen onderwijsinstituut PsyKI*

Niveau	Verantwoordelijk	Borgend & controlerend	Adviserend
Curriculum-niveau	Onderwijsdirecteur (bij KI, BS en CNS gedelegeerd aan hoofden opleidingen en bij master Psy aan mastercoördinatoren)	Examencommissie	Opleidingscommissie
Leerlijn/ Jaarniveau	Verantwoordelijke voor leerlijn of jaar	Examencommissie	Opleidingscommissie
Cursus-niveau	Eindverantwoordelijk docent	Examencommissie	Opleidingscommissie

De *eindverantwoordelijk docent* is verantwoordelijk voor de didactische consistentie op cursusniveau. Deze docent treedt tevens op als door de Examencommissie aangewezen examinator en is als zodanig verantwoordelijk voor de validiteit, betrouwbaarheid, inzichtelijkheid en uitvoerbaarheid van de toetsing en de beoordeling. In cursussen waarbij meerdere docenten betrokken zijn, vervult de eindverantwoordelijk docent bij de constructie van het tentamen een coördinerende maar ook sturende en beslissende rol. Waar gewenst kan een eindverantwoordelijk docent zich laten ondersteunen door een cursuscoördinator.

De *jaarcoördinator* zorgt ervoor dat de toetsen binnen het betreffende opleidingsjaar voldoen aan de vuistregels op curriculumniveau. De jaarcoördinator overlegt daartoe per semester met de eindverantwoordelijke docenten over de afstemming van onderwijs en toetsing, aan de hand van de teacher reports.

Op opleidingsniveau is de *onderwijsdirecteur* ervoor verantwoordelijk dat de toetsprogramma's van de opleidingen beschreven en regulier onderhouden worden en dat de toetsprogramma's voldoen aan de vuistregels en/of dat verantwoord wordt waarom afwijking van de vuistregels geboden is. De onderwijsdirecteur laat zich hierin bijstaan door hoofden/coördinatoren van de opleiding (KI, CNS en BS; master Psychologie en binnen de bachelor Psychologie door leerlijn-/domeincoördinatoren). Verder zorgt de onderwijsdirecteur c.q. hoofd/coördinator van de opleiding, in samenspraak met de Examencommissies, voor toerusting van docenten inzake toetsing en beoordeling, en voor passende ondersteuning door een *kwaliteitszorgteam* of vergelijkbare ondersteuning.

De *Opleidingscommissie* adviseert de onderwijsdirecteur/ hoofden van de opleiding over alle aspecten van onderwijskwaliteit, inclusief de kwaliteit van toetsing. Teacher reports en bijbehorende toetsmatrixen worden in de OLC besproken.

De *Examencommissie* speelt een cruciale rol in het bewaken en bevorderen van de kwaliteit van toetsing en beoordeling. De examencommissie doet dit door, op basis van het toetsbeleid en op basis van vragen van examinatoren of studenten, verdere richtlijnen op te stellen voor de uitvoering van het toetsbeleid. Verder borgt de examencommissie de realisatie van het eindniveau, door (a) de toetsprogramma's van de opleidingen te screenen aan de hand van de zes vuistregels op curriculumniveau; en door (b) steekproefsgewijs toetsmateriaal en enkele bijbehorende beoordelingen te screenen aan de hand van de elf vuistregels op cursusniveau.

Deze screening van toetsprogramma's en toetsmaterialen gebeurt in de bacheloropleidingen eenmaal per accreditatieperiode (1x per 6 jaar) en in de masteropleidingen tweemaal per accreditatieperiode (1x per 3 jaar), zodanig dat leerlijnen of opleidingsjaren alternerend 'aan de beurt zijn'.

Daarnaast screenen de examencommissies steekproefsgewijs jaarlijks 10% van de beoordelingen van de bachelor- en mastertheses (en indien van toepassing: van de stagebeoordelingen) uit het voorgaande academisch jaar, gespreid naar eindcijfers 6 tot en met 10.

Voor hun screeningsactiviteiten hanteren de examencommissies een meerjarenplanning die bekend wordt gemaakt aan docenten. Deze planning sluit bij voorkeur aan op de kwaliteitszorg- en innovatiecyclus.

Hoofdstuk 2: Toetsfasen en -toetsrichtlijnen

Dit hoofdstuk vormt een verdere uitwerking van de vuistregels op cursusniveau. Dit hoofdstuk is dan ook met name bedoeld voor eindverantwoordelijke docenten c.q. examinatoren. We beschrijven eerst de toetsfasen die een examinator doorloopt. In de tweede paragraaf worden deze toetsfasen voorzien van concrete richtlijnen om de examinatoren te ondersteunen bij het nemen van verantwoorde beslissingen inzake toetsing en beoordeling, in lijn met het toetsbeleid.

2.1 Toetsfasen

De *verantwoordelijk docent/ toetsontwikkelaar/ examinator* doorloopt voor elke toets vijf fasen. Meestal doen docenten dat vrij impliciet. Naar mate de toetskwaliteit meer aandacht krijgt en naar mate meerdere partijen bij de toetsing betrokken zijn, wordt het belangrijk om een gedeelde visie te hebben op deze fasen en op de planning daarvan. Hieronder worden deze fasen globaal uitgewerkt.

De *eerste fase* vindt plaats voor aanvang van de cursus. Het product uit deze fase is een algemeen ontwerp van de toets, waarin geëxpliciteerd is welke leerdoelen worden getoetst en op welke manier. Dit wordt ook wel de toetsmatrix genoemd. Ook wordt in deze fase al een beslissing genomen over de weging van de (verschillende) toetsonderdelen en de cesuur voldoende/onvoldoende, zodat dit naar studenten gecommuniceerd kan worden.

De *tweede fase* bestaat uit de constructie van de toets. De toetsmatrix wordt hierbij verder uitgewerkt tot toetsvragen en een beoordelingsmodel. De kwaliteit van deze producten wordt gecontroleerd door collega's en op basis van deze feedback aangepast tot een definitieve toets. Ter afsluiting van deze fase worden alle praktische zaken rondom toetsing geregeld (toetsinstructies, voldoende exemplaren et cetera), zodat de toets afgenomen kan worden in de *derde fase*.

In de *vierde fase* vindt het nakijken van de toets plaats. Hierbij worden op zo objectief mogelijke wijze punten aan de gegeven antwoorden toegekend en worden punten omgezet in cijfers.

De *vijfde* fase, tot slot, betreft de evaluatie van de toets. Hierin analyseert de docent de betrouwbaarheid van de toets en voert hij/ zij zo nodig wijzigingen door in de puntentoekenning of normering. De informatie die uit deze stap wordt verkregen en de bevindingen op basis van cursusevaluatie kunnen leiden tot aanpassing van het toetsontwerp of de leerdoelen van de cursus. Op deze wijze loopt fase 5 over in fase 1 en begint de toetscyclus weer opnieuw, zie Figuur 3. De evaluatie van de toets wordt beschreven in het teacher report.

Figuur 3. *Toetscyclus*

In Figuur 4 is een schema opgenomen voor het toetsproces, uitgaande van een onderwijsperiode van 10 weken (week 1-8: onderwijs; week 9-10: tentamen en hertentamen). De figuur laat zien hoe de toetsing een integraal onderdeel kan worden van het onderwijsproces: naast de leerdoelen en de toetsvorm worden bij aanvang van de cursus ook de weging, de toetsmatrix en initiële beoordelingscriteria al vastgesteld.

Fase 3: De toetsafname
week 9

Voor aanvang van de cursus *week 0*	Loop van de cursus *weken 1 t/m 6* *weken 7 t/m 8*	Afsluiting van de cursus *weken 9 t/m 14*
Fase 1: Het cursusontwerp 1. Leerdoelen bepalen en toetsmatrix maken 2. Bepalen van toetsvorm 3. Beslissingen over weging toetsonderdelen en cesuurbepaling.	**Fase 2: De toetsconstructie** 4. Ontwerpen van concepttoets o.b.v. kwaliteitscriteria 5. Ontwerpen van beoordelingsmodel 6. Kwaliteitscontrole vooraf: Collegiale feedback. 7. Definitieve toets afronden.	**Fase 4: Nakijken** 8. Punten toekennen 9. Cijfer bepalen **Fase 5: Evaluatie** 10. Kwaliteitscontrole achteraf: Toetsanalyse 11. Inzage en nabespreking 12. Teacher Report over toetsing opstellen 13. Toetsontwerp bijstellen en/of toetsdatabase bijwerken.

Figuur 4: Toetsfasen, uitgaande van 8 onderwijsweken en 2 tentamenweken per onderwijsperiode

2.2 Richtlijnen inzake cursusontwerp

In deze paragraaf gaan we in op de richtlijnen bij de cursusontwerp (zie Figuur 4). In 2.3 zoomen we per toetsvorm in op de toetsconstructie en de beoordeling; en in 2.4 op de toetsevaluatie.

1. Leerdoelen (opnieuw) bepalen en toetsmatrix opstellen

Voor het bepalen van de leerdoelen bekijkt de verantwoordelijk docent:

- de eindkwalificaties waaraan de cursus bijdraagt
- de leerdoelen van voorgaande cursussen in de leerlijn
- het beoogde niveau van het opleidingsjaar waarin de cursus plaatsvindt.

Bij het formuleren van de leerdoelen kan Bijlage I behulpzaam zijn.

Een *toetsmatrix op cursusniveau* is een tweedelig schematisch overzicht. Hierin wordt ten eerste de relatie tussen de leerdoelen en (één of meerdere) eindkwalificaties zichtbaar gemaakt[13].

Tabel 2: Relatie tussen leerdoelen en eindkwalificaties

	Eindkw 1	Eindkw2	Eindkw3	Eindkw 4	Eindkw 5	Eindkw 6
Leerdoel						
1	X					
2		X				
3	X					
4	X					
5						X
6						X
……						

Ten tweede wordt in de *toetsmatrix op cursusniveau* de relatie weergegeven tussen de leerdoelen en de nog te construeren tentamenvragen/ opdrachten/ cases. Hiervan geven we twee voorbeelden: een eenvoudige toetsmatrix (tabel 3) of een meer uitgebreide matrix waarin ook het cognitieve niveau van de leerdoelen wordt geëxpliciteerd (tabel 4).

[13] Dit deel van de toetsmatrix vormt de verbindende schakel met het toetsprogramma (ook wel: toetsmatrix op opleidingsniveau), dat onder de verantwoordelijkheid van de opleidingscoördinator ressorteert.

Tabel 3: Een eenvoudige toetsmatrix

	Leerdoel 1	Leerdoel 2	Leerdoel 3	Leerdoel 4
Nummer van Vraag/ taak				
1	x			
2		x		
3	x		x	
4	x			
5		x		
6			x	
7	X			x

Tabel 4: Een meer uitgebreide toetsmatrix; de getallen in de matrix vertegenwoordigen verschillende opgaven; de percentages het aandeel in het eindcijfer resp. het cognitief niveau

Leer-doelen	Kennis & Inzicht	Toepassen	Analyseren	Synthese	Evalueren	Aandeel
LD 1	1,4,17,8	2, 3 etc.				***40 %***
LD 2	5,7,12		10, 16			***20 %***
LD 3	Etc.					***10 %***
LD 4						***10 %***
LD 5						***10 %***
LD 6						***10 %***
	30 %	***15 %***	***25 %***	***15%***	***15%***	***100 %***

2. Bepalen van de toetsvorm(en)

Het is afhankelijk van het doel van de cursus en van de gekozen onderwijsvorm welke toetsvormen zich goed lenen voor welke cursus. Veel voorkomende toetsvormen zijn: toetsen met gesloten vragen, toetsen met open vragen, werkstukken, presentaties en opdrachten waarin studenten vaardigheden demonstreren (zoals interview- en gespreksvaardigheden, statistische berekeningen, ontwerpvaardigheden etc.). Deze laatste groep vatten we samen onder de noemer *assessments*. In tabel 4 worden de voor- en nadelen van deze toetsvormen schematisch weergegeven en wordt aangegeven welke toetsvormen zich voor welke leerdoelen lenen[14] [15] [16].

[14] Van Berkel, H., & Bax, A. (2006). *Toetsen in het hoger onderwijs.* Houten: Bohn Stafleu van Loghum.
[15] Academisch Centrum Tandheelkunde Amsterdam (1998). *Handleiding tentaminering*. Amsterdam: Academisch Centrum Tandheelkunde. Verkregen van http://www.onderwijs.acta.nl/studieweb/docentenwegwijzer/2_tentoe_handleiding_tentaminering.pdf
[16] Vrije Universiteit Amsterdam (2006). *Handleiding toetsen en beoordelen.* Verkregen van http://www.fsw.vu.nl/nl/Images/088%20toetsen_beoordelen_2007_tcm30-36518.pdf

Tabel 5. *Voor- en nadelen per toetsvorm en koppeling met leerdoelen*

Toetsvorm	Voordelen	Nadelen	Leerdoelen
Gesloten vragen	- Weinig nakijktijd - Efficiënt bij grotere studentaantallen	- Constructie is arbeidsintensief	- Cognitieve vaardigheden op het niveau van kennis, begrip en toepassing.
Open vragen	- Uitdagend voor studenten - Stimuleert creativiteit en originaliteit - Efficiënt bij kleine studentaantallen	- Nakijken is arbeidsintensief - Studenten verwerken feedback mogelijk te oppervlakkig - Doet een beroep op de taalvaardigheid van de student	- Cognitieve vaardigheden op het niveau van kennis, begrip, toepassing, analyse, synthese en evaluatie.
Werkstukken	- Uitdagend voor studenten - Stimuleert creativiteit en originaliteit - Geschikt voor integratieve toetsing van meerdere vaardigheden	- Betrouwbaarheid van beoordelingen vereist extra aandacht - Nakijken is arbeidsintensief - Studenten verwerken feedback mogelijk te oppervlakkig - Toetst vaardigheden die mogelijk geen leerdoel zijn (bv. taalvaardigheid, organisatorische vaardigheden)	- Hogere cognitieve vaardigheden zoals toepassen, analyse, synthese en evaluatie - Schriftelijke communicatieve vaardigheden, zoals het schrijven van een onderzoeksverslag, paper of essay
Assessments	- Betere afspiegeling van latere beroepspraktijk	- Betrouwbaarheid van beoordelingen vereist extra aandacht - Tijdrovend, zowel m.b.t. organisatie als afname	- Toepassen en integreren van vaardigheden die niet schriftelijk zijn te toetsen.

Gesloten vragen. Toetsen met gesloten vragen hebben het grote voordeel dat het nakijken van de toets snel kan gebeuren. Een nadeel van deze toetsvorm is dat de constructie van de toets relatief veel tijd en aandacht vraagt. De toetsvorm is daarmee vooral geschikt bij grotere studentenaantallen, omdat de tijdsinvestering die voor de constructie moet worden gedaan dan wordt terugverdiend door het snelle nakijkwerk.

De gesloten toetsvraag leent zich erg goed om de aanwezigheid van kennis te toetsen. Gesloten vragen kunnen ook zo worden gemaakt dat ze hogere *cognitieve* vaardigheden toetsen, maar dit vraagt meer aandacht voor het constructieproces.

Open vragen. Bij een open vraag, die al dan niet wordt ingeleid met informatie over de context, dient de student veelal een eigen stellingname te onderbouwen. Het toetsen met open vragen kan zo de creativiteit, de redeneervaardigheden en de zelfstandigheid van studenten stimuleren die de academische attitude kenmerken. Toetsen met open vragen is daarom te verkiezen boven gesloten vragen.

Het nadeel van het gebruik van open vragen is het intensieve nakijkwerk. Daarom zijn toetsen met open vragen vooral geschikt voor kleinere studentenaantallen. Bij grote aantallen kunnen combinaties van gesloten en open vragen uitkomst bieden. In dat geval

kunnen eerst de gesloten vragen worden nagekeken om studenten die zullen zakken uit te sluiten voor het nakijken van de open vragen.

Een bijkomend nadeel van toetsen met open vragen is dat deze ook een beroep doen op de taal- en redeneervaardigheid van de student, terwijl dat niet altijd een specifiek doel van de cursus hoeft te zijn. Dit nadeel is echter relatief aangezien taal- en redeneervaardigheid voor academische studenten sowieso vereist zijn. Het is dan wel aan te raden deze vereisten ook in de leerdoelen op te nemen.

Werkstukken. Bij toetsen met werkstukken maakt de student, al dan niet in groepsverband, een tekst die wordt beoordeeld. Voorbeelden hiervan zijn het paper, het essay en het onderzoeksverslag. Werkstukken zijn erg geschikt voor het toetsen van leerdoelen waarbij het toepassen en integreren van verschillende kennisdomeinen centraal staat. Zo zal de student kennis moeten kunnen selecteren, combineren en toepassen, en over voldoende schrijf- en redeneervaardigheid moeten beschikken.

Een potentieel nadeel van het gebruik van werkstukken voor het beoordelen van studenten is de lagere betrouwbaarheid van de beoordelingen en de mogelijkheid van interbeoordelaarsverschillen. Dit kan verminderd worden (echter nooit volledig uitgesloten worden) door het gebruik van robuuste beoordelingsschema's. Vermijd echter teveel details.

Verder moeten bij werkstukken die door een groep studenten zijn gemaakt, extra maatregelen worden getroffen om studenten (ook) individueel te beoordelen. Een laatste nadeel voor de docent is dat werkstukken vaak intensief zijn om na te kijken.

Assessments. Met een assessment wordt handelen in een gesimuleerde (beroeps)praktijk getoetst en beoordeeld. Hierbij valt te denken aan het toetsen van vaardigheden als het afnemen van interviews, het geven van presentaties en het gebruiken van bepaalde computerprogramma's. Dit type toetsen zal vooral gebruikt worden om leerdoelen te toetsen die inhouden dat de student bepaalde dingen moet kunnen die niet goed schriftelijk zijn te demonstreren. Een voorbeeld van een dergelijk leerdoel is "je kunt een mondelinge presentatie geven".

Assessments kennen als mogelijk nadeel de lage betrouwbaarheid. Daarbij is een assessment een tijdrovende klus die bovendien de nodige extra organisatie en faciliteiten vraagt. Voor bepaalde leerdoelen vormen assessments echter de enige goede toetsmogelijkheid.

2.3 Richtlijnen inzake toetsconstructie en beoordeling, per toetsvorm

In deze paragraaf wordt beschreven hoe er per toetsvorm zo goed mogelijk kan worden voldaan aan de validiteits- en betrouwbaarheidseisen en wat er komt kijken bij de beoordeling. Achtereenvolgens komen aan de orde: gesloten vragen (2.3.1); open vragen (2.3.2), werkstukken (2.3.3), assessments (2.3.4) en stage en thesis (2.3.5). Assessments gebruiken we hier als een verzamelterm voor toetsing middels min of meer authentieke beroepstaken, zoals presentaties, interviews, testafname, gespreksvoering, statistische bewerkingen, formeel modelleren, software-ontwerp of -evaluatie etc.

Om zicht te bieden op de hoofdlijnen, bieden we eerst een samenvattend overzicht (tabel 5) van de richtlijnen per toetsvorm die in acht genomen moeten worden om de validiteit en betrouwbaarheid te waarborgen.

Tabel 6. *Overzicht van richtlijnen per toetsvorm*

Toetsvorm	**Validiteit**	**Betrouwbaarheid**	**Beoordeling**
Gesloten vragen	- Toetsmatrix - Plausibele afleiders - Juiste constructie - Peer review, zowel vooraf als achteraf - Analyse psycho-metrische gegevens.	- Voldoende items - Analyse van psychometrische gegevens	- Absolute cesuur of absolute cesuur met relatieve referentie
Open vragen	- Toetsmatrix - Juiste constructie	- Voldoende vragen - Nakijk-model - Beoordelingsmethode die beoordelaars-effecten vermindert	- Beoordelingsschema - Tweede beoordelaar; anoniem beoordelen
Werkstukken	- Toetsmatrix - Duidelijke opdracht - Peer review.	- Beoordelingsschema: criteria en procedures - Peer review.	- Tweede beoordelaar; anoniem beoordelen.
Assessments	- Toetsmatrix - Duidelijke opdracht - Peer review.	- Beoordelingsschema: criteria en procedures - Peer review.	- Tweede beoordelaar - Oefenen met beoordelingsschema - Evt. opnemen van assessment.

2.3.1 Toetsen met gesloten vragen

Hoeveelheid vragen
De minimaal vereiste hoeveelheid vragen bij een meerkeuze-toets wordt ten eerste bepaald door de leerdoelen: elk leerdoel moet tenminste één keer worden getoetst. Naarmate een leerdoel meer gewicht in een cursus heeft, zullen er logischerwijze ook meer vragen over gesteld worden. Ook geldt dat hoe meer vragen er in een toets zitten, hoe beter de student kan demonstreren dat hij/zij aan de leerdoelen voldoet. De toets is dan namelijk een betere steekproef uit het leerdomein. In tabel 6 is het minimaal aantal vragen per meerkeuze-variant weergegeven voor een betrouwbare toets[17].

Tabel 6. *Minimaal aantal vragen per meerkeuze-variant.*

Meerkeuze-variant	*Aantal vragen*
Vierkeuze	40
Driekeuze	60
Tweekeuze	80

[17] Berkel, H. van, Bakx, A. & Joosten-Tenbrinke, D. (2013). Toetsen in het hoger onderwijs (3de druk). Houten: Bohn, Stafleu, van Loghum.

Aantal antwoordalternatieven

De kwaliteit van de meerkeuzevraag hangt verder in hoge mate af van de kwaliteit van de afleiders. Het advies is om de beslissing voor vierkeuze- of driekeuze-vraag te laten afhangen van het aantal goede afleiders dat geconstrueerd kan worden, mede in relatie tot de tijdsinvestering. Docenten blijken vaak te kiezen voor vierkeuze-vragen vanuit het idee dat daarbij de gokkans kleiner is; de gokkans is echter vaak groter dan wordt aangenomen doordat de kwaliteit van de afleiders bij vierkeuze-vragen vermindert[18].

Het is mogelijk om in één tentamen twee-, drie- en vierkeuze-vragen door elkaar te gebruiken. Het is wel goed om dit van tevoren kenbaar te maken aan studenten. Bij het nakijken dient uiteraard gecorrigeerd te worden voor deze verschillende gokkansen. Deze tentamens kunnen ook gewoon door de ITS geanalyseerd worden, mits de docent bij de antwoordsleutel aangeeft hoeveel antwoordalternatieven elke vraag heeft.

Zorgvuldige formulering van alternatieven

Zorg ervoor dat de antwoordalternatieven goed worden geconstrueerd:

1. De antwoordalternatieven zijn allemaal op hetzelfde aspect gericht.
2. De antwoordalternatieven hebben ongeveer dezelfde lengte.
3. De antwoordalternatieven zijn even genuanceerd.
   Voorbeeld waarin alternatief A genuanceerder is dan B:
   A) de groep cellen die geactiveerd wordt tijdens de vorming van geheugen
   B) neuronen
4. Orden de antwoordalternatieven neutraal, zodat de volgorde geen impliciete aanwijzingen levert voor wat het juiste antwoord is.
5. Stel één vraag tegelijk en geef ook maar één antwoord per antwoordalternatief.
6. Geef geen onnodige informatie in de antwoordalternatieven; verwijder gegevens die niet onderscheidend zijn.
7. De antwoordalternatieven moeten elkaar uitsluiten. Voorkom overlap in de antwoordalternatieven.
   Voorbeeld van overlap:
   A) mensen met psychiatrische problemen
   B) mensen met een as-II stoornis
   C) mensen met een narcistische persoonlijkheidsstoornis
8. Eén antwoordalternatief moet verdedigbaar correct zijn, de anderen verdedigbaar incorrect.
9. Vermijd logische of inhoudelijke cues.
   Voorbeeld waarin het goede antwoord uit de zinsconstructie kan worden afgeleid:
   "Wie eens steelt is altijd een dief." Wat betekent dit spreekwoord?
   A) Als je iemand bedriegt, word je zelf ook bedrogen.
   B) Als je niet eerlijk bent, kun je dat nooit meer vergeten.
   C) Als je slecht bent, zul je ook slecht over anderen denken.
   D) Wie eenmaal een misstap begaat, is altijd onbetrouwbaar.
10. Vermijd in de antwoordalternatieven absolute (nooit, altijd etc.) en te open (kan, soms, misschien) begrippen of termen.
11. Hou het zo simpel mogelijk en vermijd ingewikkeld of overdrachtelijk taalgebruik.

[18] E.C. Paes, E.C. & Cate, O. ten. (2009*).* Meerkeuzevragen met drie, vier of vijf alternatieven: wat is beter? *Tijdschrift voor Medisch Onderwijs,* 28*(3).*

Juist/ onjuist-vragen

Voor juist/ onjuist-vragen (stellingen) gelden dezelfde richtlijnen als voor meerkeuze-tentamens. Er zijn nog enkele aanvullende aanwijzingen: zorg ervoor dat de stelling echt in zijn geheel juist is en dat de formulering nauwkeurig is. De student moet immers een gehele stelling als juist/ onjuist beoordelen, en kan dus over elk onderdeel van de formulering twijfelen. En verder: beperk eventueel de moeilijkheid van een stelling door de instructie te geven om alleen een specifiek woord of zinsdeel dat gecursiveerd is op correctheid te beoordelen.

Bij vragen van het onderstaande format kan de validiteit in het gevaar komen.

A. Alleen 1 is juist
B. Alleen 2 is juist
C. 1 en 2 zijn beide juist
D. 1 en 2 zijn beide onjuist.

Dit zijn in feite twee juist/onjuist-vragen die in een vierkeuze-vraag zijn gestopt. Deze vraagvorm wordt afgeraden. Ten eerste omdat deze niet eerlijk is ten opzichte van studenten: als ze terecht aangeven dat stelling 1 correct is, maar van stelling 2 geen idee hebben, krijgen ze voor stelling 1 geen punten. Ten tweede komt een groter aantal vragen de betrouwbaarheid van de toets ten goede en is het opsplitsen van de vragen dus gewenst. Een derde reden is dat de feedback aan de student minder nauwkeurig wordt: als een student de vraag fout heeft, weten we immers niet wat de student op het niveau van elke afzonderlijke stelling eventueel wel weet. Om deze redenen is het dus beter om een vraag van deze vorm te splitsen in twee afzonderlijke juist/onjuist vragen.

Nakijken van toetsen met gesloten vragen

De ruwe score van studenten op de toets met gesloten vragen moet omgezet worden in een cijfer. Allereerst is het van belang om de cesuur te bepalen: dus welke score krijgt een 5,6 (af te ronden naar 6) en dus voldoende; en welke score krijgt een 5,4 en dus onvoldoende. Op basis daarvan kan de verdere range van cijfers worden vastgesteld.

Hierbij kunnen twee methodieken worden gebruikt: de absolute beoordeling of de relatieve beoordeling. Beide methodieken kennen voor- en tegenstanders. Bij de *relatieve beoordeling* wordt – ongeacht het niveau van de studentengroep – een vaststaande spreiding aangehouden, bijvoorbeeld: de beste 30% haalt 7 of hoger; de slechtste 30% behaalt een onvoldoende; en de middengroep slaagt met 6 of 7. Bij de *absolute beoordeling* wordt de cesuur gehanteerd ongeacht het feitelijke niveau van de studentengroep. De owi-richtlijn is om *de absolute beoordeling te hanteren*. Dit is mogelijk wanneer meerkeuze-tentamens zijn samengesteld op basis van een goede steekproefmethode uit een toetsdatabase die psychometrisch gecontroleerd is.

In bepaalde gevallen is absolute beoordeling niet goed mogelijk: bijvoorbeeld als er nieuwe toetsvragen zijn ontwikkeld of als psychometrische gegevens anderszins ontbreken. In dit geval heeft de samensteller van de toets geen controle over de moeilijkheidsgraad van de uiteindelijke toets. Daardoor is het mogelijk dat de ene toets veel moeilijker is dan de andere. Een absolute cesuur is dan niet verantwoord, aangezien de eis voor slagen dan ongelijk is voor verschillende groepen studenten. In dat geval is de owi-richtlijn dat de examinator een tussenvorm kiest, namelijk de zogeheten *absolute*

methode met relatieve referentie. Hieronder verstaan we een methode waarbij de cesuur wordt gerelateerd aan de gemiddelde score van de beste 5%. De berekeningswijze van 'absolute beoordeling' en van 'absolute beoordeling met een relatieve referentie' staan beschreven in bijlage II.

Een tweede reden om te kiezen voor *absolute methode met relatieve referentie* is dat voor een goed presterende student de maximale score op een meerkeuzetoets lager is dan de theoretisch hoogst haalbare score: door de correctie voor gokken kan de student nooit de maximale score (het cijfer 10) behalen, ook niet als hij/ zij wellicht wel alle vragen correct beantwoord heeft. Door te relateren aan de 5% beste studenten is de theoretisch hoogst haalbare score wel mogelijk.

Absolute beoordeling met relatieve referentie is alleen te verantwoorden als de groepen studenten groot zijn (groter dan 400)[19], ze van jaar tot jaar vergelijkbaar zijn en de kwaliteit van onderwijs gelijk is gebleven. Uit onderzoek blijkt dat wanneer voor de moeilijkheidsgraad wordt gecorrigeerd, het slagingspercentage minder fluctueert over de toetsing van verschillende jaargangen heen, en dat een groter percentage slaagt[14]. Verder blijkt, uit hetzelfde onderzoek, dat er geen kennisverlies is bij toetsing waar cesuur met relatieve referentie is toegepast: studenten blijken aan het eind van hun opleiding op hetzelfde niveau van leren te zitten als studenten van opleidingen die de toetsing met absolute cesuur hebben.

Norm handhaven bij herkansingen

Omdat de populatie studenten bij een herkansing niet hetzelfde is als bij een eerste afname van de toets (de herkansingsgroep bevat immers vooral de relatief zwakke studenten en is een kleinere groep waarbij er procentueel gezien meer studenten zakken), is het in principe niet wenselijk om de cesuur afhankelijk te stellen van de prestatie van de groep. Dat zou immers betekenen dat de herkansing waarschijnlijk 'soepeler' beoordeeld zou worden dan de eerste afname. Dit vergroot de kans dat studenten ten onrechte een voldoende halen. Anderzijds kan ook bij herkansingen de moeilijkheidsgraad verschillen. Om aan dit probleem het hoofd te bieden, is er binnen het onderwijsinstituut voor gekozen om bij een herkansing uit te gaan van eenzelfde moeilijkheid als bij de eerste afname en dus dezelfde norm te handhaven[20]. De voorwaarde hierbij is dat de herkansing op dezelfde manier wordt samengesteld als de eerste kans. Dit betekent dat in het geval van het 'absoluut beoordelen met relatieve referentie' de gemiddelde score van de beste 5% bij de herkansing hetzelfde wordt gesteld als de gemiddelde score van de beste 5% bij de eerste afname.

2.3.2 Toetsen met open vragen

Voor open vragen geldt, net als voor gesloten vragen, dat deze vragen gerelateerd moeten worden aan de leerdoelen om zo de validiteit van de toets te waarborgen. Ook hier vormt de toetsmatrix (zie 2.2) een goed hulpmiddel.

Verder wordt de *validiteit* bij open vragen ook weer in hoge mate bepaald door de kwaliteit van de formulering van de vraag. Hierbij dient rekening te worden gehouden met de volgende punten:

[19] Sanders, P. (2011). *Toetsen op school.* Arnhem: Cito.
[20] Gruijter, D.N.M. de (2008). *Toetsing en Toetsanalyse*. Leiden: Universiteit Leiden.

1. Een probleem dat zich bij open vragen kan voordoen is dat de vraag niet *eenduidig* is; de vraag lokt dan verschillende antwoorden uit die allemaal te verdedigen zijn. Om dit te voorkomen kan het helpen om andersom te werken: formuleer eerst het gewenste (model)antwoord en daarna de bijbehorende vraag.
2. Ook kan het van belang zijn om antwoordrestricties aan de vraag toe te voegen; studenten zijn bijvoorbeeld vaak beknopter of uitgebreider in hun antwoord dan de docent voor ogen had. Informatie over de gewenste lengte van het antwoord kan hierbij uitkomst bieden. Wees hierbij wel concreet; formuleringen als “noem enkele voorbeelden van…” zijn te vaag. Ook kan een voorgestructureerde antwoordruimte handig zijn.
3. Zorg dat duidelijk is op welk deel van de vraag het antwoord betrekking moet hebben. Een vraag als “leg uit waarom Freud bij de behandeling van kleine Hans achter zijn patiënt plaatsnam” kan bijvoorbeeld gelezen worden met de nadruk op *waarom*, met de nadruk op *Freud*, met de nadruk op *kleine Hans* of op *achter*. Afhankelijk hiervan zullen de antwoorden waarschijnlijk verschillen. Dit kan worden opgelost door het onderdeel waarop men de nadruk wil leggen bijvoorbeeld te onderstrepen of te cursiveren.
4. Als u wil dat de student een onderbouwing/toelichting/voorbeeld geeft, zet dit dan duidelijk in de toelichting bij de vraag. Vergelijkbaar: geef aan dat de student kennis van de leerstof x en y moet gebruiken bij het beantwoorden.
5. Zorg dat de vraag taalkundig juist en zo eenvoudig mogelijk is geformuleerd. Vermijd overdrachtelijk taalgebruik.
6. Formuleer de vraag positief (vertel wat de student wél moet doen, niet wat deze moet laten).
7. Controleer of de vraag voldoende informatie bevat om hem optimaal te kunnen beantwoorden en of overbodige informatie vermeden is.
8. Stel geen strikvragen.
9. Zorg voor een overzichtelijke layout etc.

Om de *betrouwbaarheid* van de toets te waarborgen is het bij het gebruik van open vragen net als bij gesloten vragen van belang voldoende vragen te stellen; zodat alle leerdoelen gedekt zijn en elk leerdoel zo mogelijk op meerdere manieren bevraagd wordt.

Verder zijn de *beoordelingscriteria* van belang voor de betrouwbaarheid van de toets. Bij open vragen dient voorafgaand aan de correctie duidelijk te zijn op grond van welke criteria de vraag beoordeeld wordt en welke prestatie tot welk oordeel leidt. Dit verhoogt de objectiviteit van de beoordeling. De volgende richtlijnen waarborgen de kwaliteit van de beoordelingscriteria:

1. Formuleer een nakijk-model tegelijkertijd met het construeren van de toets. Hierbij moet eenduidig en zorgvuldig zijn vastgelegd welke prestatie van de student welke waardering krijgt. Vaak gebeurt dit door per vraag een zogenaamd *modelantwoord* te formuleren. Het nakijk-model is vooral bedoeld voor de beoordelaars en voor de onderbouwing van de cijfers tijdens de inzage.
2. Een goed nakijk-model is echter meer dan een modelantwoord. Het nakijk-model bevat tevens de aan de antwoorden te koppelen puntenverdeling. En bijvoorbeeld ook voorbeelden van voor de hand liggende foute antwoorden.

3. Expliciteer in het nakijk-model ook algemene beoordelingsinstructies, bijv. hoe om te gaan met gedeeltelijk juiste antwoorden, fouten die doorwerken in volgende vragen, taal- en spellingfouten, onleesbaar handschrift en overschrijding van de voorgeschreven maximum antwoordlengte.
4. Test het nakijk-model vooraf op bruikbaarheid en volledigheid. Dit kun je doen door de toets af te nemen bij iemand die de betreffende stof beheerst, maar voor wie de toets zelf nieuw is. Ook kunnen de beoordelaars het nakijk-model tussentijds evalueren op het moment dat ongeveer een derde van de toetsen is nagekeken: is het nakijk-model volledig? Zijn alle criteria bruikbaar?

De betrouwbaarheid van een toets met open vragen is ook te vergroten door een handige beoordelingsmethodiek te kiezen en door optredende beoordelaarseffecten te verminderen. Tabel 8 beschrijft veel voorkomende beoordelaarseffecten en geeft tip voor kleine aanpassingen in de beoordelingsmethodiek die deze effecten (zoveel als mogelijk) tegengaan.

Tabel 8. *Voorkomende beoordelaarseffecten en maatregelen om die tegen te gaan.*

Beoordelaarseffect	**Maatregel**
Normverschuiving: de beoordelaar wordt tijdens de beoordeling steeds milder of strenger.	De nakijk-volgorde van toetsonderdelen variëren. Het op alfabetische volgorde nakijken wordt sterk afgeraden.
Sequentie-effect: de beoordelaar beoordeelt - na een groot aantal slechte prestaties - een relatief goede prestatie buitenproportioneel hoog (en andersom).	Kijk open vragen bij voorkeur per vraag na en niet per student. Hierdoor wordt het sequentie-effect en meestal ook het halo- en contaminatie-effect verminderd. Bovendien valt zo meestal ook tijdwinst te behalen, omdat de beoordelaar niet steeds hoeft om te schakelen tussen de vragen.
Halo-effect: het beeld dat de beoordelaar van de student heeft (bijvoorbeeld “hij/zij is een goede student”) beïnvloedt de beoordeling.	De toets wordt nagekeken door een beoordelaar die de student niet kent. Anoniem (op studentnummer) beoordelen.
Contaminatie-effect: de beoordelaar ziet de prestaties van de studenten als een afspiegeling van de kwaliteit van het zelfgegeven onderwijs en is geneigd hoger te beoordelen dan realistisch is.	Ontkoppel doceren en beoordelen zoveel mogelijk.

2.3.3 Toetsen met werkstukken

Net als bij andere toetsvormen kan bij het beoordelen van werkstukken een toetsmatrix worden gemaakt. In plaats van de toetsopgaven worden in dit geval de beoordelingcriteria uit het beoordelingsschema in de matrix gekoppeld aan de leerdoelen die horen bij de opdracht.

Voor het verhogen van de validiteit, de betrouwbaarheid en de inzichtelijkheid is het van belang om een concrete beschrijving te geven van het verwachte eindproduct (*wat* wordt beoordeeld?), de beoordelingscriteria (*waarop* wordt beoordeeld?) en de scoring (*hoe* wordt gewogen en becijferd?). Daardoor heeft de student voldoende houvast bij het

uitvoeren van de opdracht (het maken van het werkstuk) en de docent bij het beoordelen daarvan. Het geheel van beoordelingscriteria en beoordelingsprocedures, wordt het beoordelingsschema genoemd.

Net als bij andere toetsvormen dienen de opdracht en het beoordelingschema bij het gebruik van werkstukken bij voorkeur vooraf 'getest' te worden, bijvoorbeeld door in peer review de opdracht en het nakijkschema te evalueren. Hierbij dienen de volgende vragen besproken en afgestemd te worden:

1. Is de opdracht helder geformuleerd? Is de opdracht eenduidig? Is de opdracht taalkundig juist geformuleerd?
2. Bevat de opdracht alle informatie die de student nodig heeft om deze goed uit te kunnen voeren? Bevat de opdracht onnodige informatie (ruis)?
3. Is het nakijkschema eenduidig?
4. Sluiten opdracht en beoordelingscriteria aan bij de bijbehorende leerdoelen?
5. Is er onderlinge overeenstemming over de manier waarop het gebruik van het schema tot een cijfer leidt?

Onderscheiden van begeleiden en beoordelen

In de praktijk is de begeleider bij het schrijven van een werkstuk (bijvoorbeeld de thesisbegeleider) vaak ook (mede) beoordelaar. Deze dubbelrol brengt echter het risico met zich mee dat de begeleider bij de beoordeling het beeld van- en de relatie die hij/ zij heeft opgebouwd met de student meeweegt in de beoordeling (halo-effect). Een ander risico is dat hij/zij mede zichzelf beoordeelt (contaminatie-effect). Deze effecten zijn sterker als niet alleen het product wordt beoordeeld, maar ook het proces.

Om deze effecten tegen te gaan verdient het aanbeveling werkstukken door een andere beoordelaar dan de begeleider te laten beoordelen. Bij stages en scripties is een onafhankelijke beoordeling door een tweede beoordelaar verplicht (blind vier-ogenprincipe). Wanneer deze personen verschillen van oordeel is de procedure dat beide beoordelaars met elkaar in overleg gaan om, op basis van de afgesproken criteria, tot een eensgezind oordeel te komen. Lukt dit niet, dan wordt de kwestie overgedragen aan de coördinator van de betreffende cursus die vervolgens het uiteindelijke cijfer zal bepalen.

Individueel beoordelen van groepsproces en -product

Bij werkstukken die in zijn geheel door een groep studenten zijn gemaakt, is het lastiger om individuele cijfers toe te kennen. Omdat dit toch gewenst is kan de docent bijvoorbeeld zijn/haar observaties systematisch noteren, studenten een logboek laten bijhouden van werkzaamheden, vragen en ideeën of studenten individuele (deel)presentaties of reflectieverslagen laten maken. Ook kunnen studenten een gedeelte van het werkstuk individueel schrijven en een gedeelte als groep.

Bij groepswerk is het belangrijk om het groepsproces los te zien van het groepsproduct, het werkstuk. Het groepsproces wordt alleen beoordeeld als 'samenwerken' deel uitmaakt van de leerdoelen. Als dit het geval is dan is het van belang de bijbehorende beoordelingscriteria vroegtijdig te expliciteren, en hieraan ook tussentijds aandacht te besteden in feedback door de docent of in *peer feedback* door de groepsleden. Daarnaast kan de docent er ook voor kiezen om bij de beoordeling *peer*

assessment (studenten beoordelen elkaar) in te zetten. Studenten dienen hierop grondig voorbereid te worden.

2.3.4 Toetsen met assessments

Het toetsen aan de hand van *assessments* (min of meer authentieke beroepssituaties) komt in grote lijnen overeen met het toetsen middels werkstukken, zoals hierboven beschreven. Een belangrijk verschil is echter dat een assessment ter plekke wordt beoordeeld, terwijl een werkstuk kan worden beoordeeld in het tempo en op het moment dat de beoordelaar dat wenst. Om deze reden is het belangrijk om een goed beoordelingsschema te hebben en om de beoordelaars in te werken in het gebruik ervan. Een hulpmiddel is het opnemen van het assessment in beeld en/of geluid. Zo kan de beoordelaar de assessment later terugzien en/of -luisteren als dat voor de beoordeling nodig is. Dit helpt ook bij de feedback aan de student.

2.4 Richtlijnen inzake toetsevaluatie

Tot nu toe is beschreven hoe de kwaliteit van de toets *voor* afname gewaarborgd dient te worden. Psychometrische analyse *na* afname van een toets levert veel informatie op over de kwaliteit van de toetsvragen en van de toets als geheel, zowel wat betreft de validiteit als de betrouwbaarheid. Deze informatie draagt daarom bij aan de kwaliteit van het oordeel over de prestatie van de student en geeft verder aanwijzingen voor verbetering van de toetskwaliteit.

Meerkeuze-toetsen worden psychometrisch geanalyseerd door het ITS. In principe lenen alle toetsvormen zich voor psychometrische analyse en interpretatie. Op dit moment zijn van psychometrische analyse van andere toetsvormen echter nog geen voorbeelden voorhanden.

Hieronder wordt de psychometrische analyse van meerkeuze-toetsen toegelicht voor wat betreft validiteit (2.4.1) en betrouwbaarheid (2.4.2); ook gaan we in op de vraag op hoe psychometrische waarden geïnterpreteerd kunnen worden en welke beslissingen de examinator in overweging kan nemen (2.4.3).

2.4.1 Validiteit

F-waarde

De *f-waarde* (frequentie) per antwoordalternatief geeft het absolute aantal studenten aan dat voor het betreffende antwoordalternatief heeft gekozen. Daarmee is het een maat voor de kwaliteit van de afleiders. Idealiter zijn de studenten die de vraag fout hebben beantwoord, gelijk verdeeld zijn over de afleiders. Dit geeft immers aan dat alle afleiders evenveel kans hebben om gekozen te worden door iemand die de stof niet beheerst.

In de praktijk zien we echter vaak dat één of meerdere alternatieven erg weinig of soms zelfs nooit worden gekozen. In dat geval is blijkbaar zonder veel kennis van de stof te bepalen dat dit alternatief in ieder geval niet juist is. De vraag meet zodoende niet alleen in welke mate de student de stof beheerst, maar blijkbaar ook nog iets anders. De validiteit wordt daarmee dus lager.

Treft u in een toets afleiders die (te) weinig worden gekozen, verwijder deze afleider dan uit de toets. Gebruikt u de vraag hierna gewoon met een afleider minder, vergeet dan niet de gokkans voor de vraag aan te passen. Een nieuwe afleider verzinnen kan natuurlijk ook.

P-waarde

De *p-waarde* is een indicatie van de moeilijkheidsgraad per toetsitem en geeft informatie over het selecterende vermogen van de toets. De *p*-waarde wordt bepaald door de proportie van de studenten die de vraag correct hebben beantwoord en is daarom altijd een getal tussen de 0 en de 1. Doorgaans worden twee verschillende *p*-waardes gegeven: de *p* en de *p'*. Beide geven aan welk deel van de studenten de vraag goed heeft beantwoord, waarbij bij *p'* is gecorrigeerd voor de gokkans. Een *p* van 0.80 geeft aan dat 80% van de studenten de vraag goed heeft beantwoord. Een deel van hen zal de vraag echter goed hebben gegokt. De *p'* waarbij voor dit gokken is gecorrigeerd wordt in geval van een vierkeuze-vraag dan 0,73. (zie: bijlage II). De vraag is nu te zien als een vraag die door 73% van de studenten goed is beantwoord zonder te gokken.

Door in een toets opgaven van verschillende moeilijkheidgraad op te nemen, is het mogelijk om beter te kunnen differentiëren tussen verschillende prestaties van studenten. Aan de *p'*-waarde is te zien in hoeverre dit gelukt is. Een vraag met een *p'*-waarde van 1 is door iedereen goed gemaakt; deze vraag heeft dus geen/ weinig selecterend vermogen. Voor een vraag met een *p'*-waarde van 0 geldt hetzelfde, in die zin dat nu niemand de vraag goed heeft beantwoord. Een *p'* van 0.5 lijkt wenselijk aangezien het de maximale bijdrage levert aan de selectieve (summatieve) functie van de toets.

Voor de juiste interpretatie van de *p'*-waarde is het van belang om rekening te houden met de aard van de groep studenten die de vraag heeft beantwoord. Omdat de *p'*-waarde groepsafhankelijk is zal deze waarde over het algemeen lager uitvallen bij groepen met een algemeen lager niveau (bijvoorbeeld herkansers) of bij kleinere groepen waarbij toeval een grotere rol speelt.

2.4.1 Betrouwbaarheid

Coëfficiënt alpha

De *coëfficiënt alpha* is de maat voor interne consistentie, en geeft dus aan wat de betrouwbaarheid van de toets als geheel is. De interne consistentie van een toets geeft de mate aan waarin de opgaven van die toets onderling in statistische zin samenhangen. Coëfficiënt alpha ligt altijd tussen 0 en 1 en hoe hoger hoe beter (vanaf 0.7 is doorgaans acceptabel). Een lage coëfficiënt alpha kan worden verhoogd door (een volgende keer) meer vragen met een hoge R_{ir} (zie verder) in het tentamen op te nemen. Voor toetsen die al afgenomen zijn en waarbij de coëfficiënt alpha te laag gevonden wordt, kan deze worden verhoogd door vragen met een negatieve R_{ir} uit de toets te verwijderen. Als er hierdoor erg veel vragen uit de toets moeten worden verwijderd, zal de toets uiteraard minder representatief worden voor de leerdoelen en zal voor een volgende toetsafname een andere toets moeten worden samengesteld.

Item-restcorrelatie

De *item-restcorrelatie* (R_{ir}-waarde) is een maat voor het discriminerend vermogen van een item. Idealiter moet elke toetsvraag een zo goed mogelijk onderscheid maken tussen studenten met een hoge en lage eindscore. Door middel van de R_{ir} kan gecontroleerd worden of de toetsvragen aan dit criterium voldoen. De R_{ir}-waardes van de afzonderlijke items hangen samen met de betrouwbaarheid van de toets als geheel. De R_{ir} wordt bepaald door de score op het item te relateren aan de eindscore op de gehele toets (gecorrigeerd voor de score op het desbetreffende item) en bestaat uit een getal tussen de -1 en 1. Een sterke, positieve R_{ir} geeft aan dat studenten die deze vraag goed hebben beantwoord, gemiddeld hoger hebben gescoord op de toets dan studenten die deze vraag fout hebben. De R_{ir} moet in elk geval positief zijn met een streefwaarde van > 0,2. Is de R_{ir} (sterk) negatief, dan heeft de vraag wel een discriminerend vermogen maar op een verkeerde manier. Het betekent grofweg dat de vraag over het algemeen goed is gemaakt door de slechte studenten en slecht door de goede studenten.

2.4.3 Toetskwaliteit verbeteren

Op basis van de kwaliteitsindicatoren die hierboven zijn beschreven is het mogelijk om de kwaliteit van een toets *na* afname te verhogen (tabel 9). Dit vereist een juiste interpretatie van de psychometrische analyses.

Tabel 9. *Interpretatie van mogelijke combinaties p' en R_{ir} -waarden en gewenste maatregelen.*

	R_{ir} is negatief	**R_{ir} is lager dan 0.15**	**R_{ir} is hoger dan 0.15**
***p'* lager dan 0.1**	Dit is een slechte vraag: Sleutel correct? *> verwijder de vraag uit de toets als de sleutel correct is toegepast.*	Dit is een vraag waar iets mee mis lijkt te zijn: Sleutel correct? Detailvraag die niet overeenkomt met de leerdoelen? Formulering van vraag eenduidig? Ander alternatief ook plausibel? *> verwijder de vraag of reken meerdere antwoordalternatieven goed als daar aanleiding voor is.*	Dit is een moeilijke vraag, en maakt alleen onderscheid tussen de negens en tienen: Instinkertje? Te moeilijk/complex? *> de vraag handhaven, zorg er alleen voor dat er niet teveel van dit soort vragen zijn.*
***p'* tussen 0.1 en 0.8**	Dit is een slechte vraag: Sleutel correct? *> verwijder de vraag uit de toets als de sleutel correct is toegepast.*	Deze vraag is niet te moeilijk of makkelijk, maar er zou iets anders kunnen zijn waarom de vraag niet discrimineert: Is een ander alternatief ook waarschijnlijk? *> evt. meerdere alternatieven goed rekenen*	Prima vraag. *> de vraag handhaven*
***p'* hoger dan 0.8**	Dit is een slechte vraag: Sleutel correct? *> verwijder de vraag*	Deze vraag is te makkelijk en discrimineert niet. Weggever (op te lossen met boerenverstand)? Zijn de afleiders wel plausibel	Dit is een makkelijke vraag, en maakt alleen een onderscheid tussen de enen en tweeën.

	uit de toets als de sleutel correct is toegepast.	genoeg? *> behoeft geen directe actie, maar moet wel aangepast worden in de database*	*> de vraag handhaven, zorg er alleen voor dat er niet teveel van dit soort vragen inzitten.*

Er zijn echter een aantal voorwaarden waarmee rekening moet worden gehouden bij het aanpassen van een toets op basis van de psychometrische gegevens:

1. Het is van belang om de kwaliteit van een vraag of een gehele toets steeds in combinatie met de inhoud van de toetsopgaven zelf te interpreteren, en dus nooit alleen op basis van de psychometrische gegevens. Deze kwantitatieve analyses gaan namelijk altijd gepaard met marges op basis van toeval, en mogen zodoende niet als absoluut worden opgevat.
2. Houd rekening met het aantal toetsdeelnemers. Hoe minder deelnemers, hoe minder betrouwbaar de kwaliteitsindicatoren zijn. Dit speelt met name een rol bij herkansingen. In dit specifieke geval is de steekproef namelijk een niet-heterogene groep, en moeten de beslissingen op basis van de psychometrische gegevens dus met extra grote voorzichtigheid worden genomen.
3. Bij het nemen van beslissingen over het aanpassen van de toets, is het belangrijk om de *p*- en R_{ir} –waardes in combinatie te interpreteren. Tabel 9 geeft een overzicht van alle mogelijke combinaties en de bijbehorende geadviseerde maatregelen. Een toelichting op de belangrijkste combinaties:

Combinatie van een erg lage p'-waarde en een lage R_{ir} : deze combinatie is verdacht. Controleer de inhoud en antwoordsleutel van deze vragen altijd op juistheid en verwijder de vraag als daar aanleiding voor is.

Combinatie van een lage p'-waarde met een negatieve R_{ir} : reden om de vraag uit de toets te verwijderen (ervan uitgaande dat de antwoordsleutel correct is gebruikt). Ten eerste is de vraag te moeilijk: ze discrimineert alleen tussen de slechte studenten en dus niet tussen de goed en slecht presterende studenten. Ten tweede scoren de beste studenten slechter op deze vraag dan de slechte studenten, en dat is dus precies wat je *niet* wilt.

Combinatie van en lage p'-waarde met een voldoende positieve $R_{ir:}$ een dergelijke vraag kan behouden blijven. Waak er echter wel voor dat er niet teveel van dit soort moeilijke vragen in de toets zitten, want uiteindelijk wil je vooral de vijven van de zessen kunnen onderscheiden.

Slechte vragen verwijderen of alle alternatieven goed rekenen?
Een alternatief voor het uit de toets verwijderen van de vraag is het goed rekenen van meerdere antwoordalternatieven. Kies voor dit laatste op het moment dat de inhoud van de vraag en de antwoordalternatieven dit rechtvaardigen, bijvoorbeeld als één van de afleiders bij nader inzien toch niet echt fout is. Wanneer toch gekozen wordt voor het verwijderen van een vraag uit de toets, houd er dan rekening mee dat bij minder vragen de betrouwbaarheid van de gehele toets verlaagd kan worden, en zorg er ook voor dat de toets nog steeds voldoende representatief is voor de leerdoelen van de cursus.

Bijlage I: Leerdoelen opstellen (of verbeteren)

Leerdoelen zijn:

- expliciet (duidelijk en concreet). *Zie hieronder bij: SMART, RUMBA*
- beperkt in aantal en hiërarchisch geordend: het belangrijkste doel wordt eerst genoemd; gebruik eventueel subdoelen (bijv. 2a, 2b)
- op het niveau dat past bij de opleidingsfase en binnen de leerlijn
- geformuleerd in gedragstermen (handelingen) of als cognitieve prestatie
- ook nog relevant na de studie, dus relevant voor het verdere leven van de studenten als professional, academicus en mens.

1. Redactionele tips bij het veranderen van al bestaande cursusdoelen

- Begin met: Na afloop van de cursus
- Gebruik *je* in plaats van *de student.*
- Gebruik een actief werkwoord (zie hieronder): je analyseert, je onderbouwt, je vergelijkt, je beoordeelt etc.

2. Heeft u momenteel teveel cursusdoelen?

- Breng het aantal (hoofd)doelen terug tot 3 á 5
- Zet het belangrijkste doel bovenaan; in de oude formulering staat die veelal als laatste.
- Een deel van de doelen is waarschijnlijk voorwaardelijk voor het kunnen realiseren van de hoofddoelen. Formuleer deze voorwaardelijke doelen als subdoelen (bijv. als 2a, 2b). Zo ontstaat een duidelijke ‘doelenhiërarchie’ die studenten en uzelf helpt om meer leerdoelgericht te werken.
- Ga nu na of het u echt om deze doelen en deze volgorde te doen is (in deze cursus, op dit niveau)?

3. Heeft u momenteel cursusdoelen die beginnen met “Kennis en inzicht in…”?

- Kennis en inzicht zijn zeer zeker nodig. Maak echter wel concreet om wélke kennis het in deze cursus gaat, en om welke niet.
  Dus niet: kennis van een vakgebied A
  Wel: kennis van drie gangbare persoonlijkheidstheorieën, te weten X, Y en Z
- Vraag uzelf af welk soort cognitieve prestatie of welk soort academisch gedrag u verwacht van een student die over deze kennis beschikt. Wat moet een student doen/ laten zien om u ervan te overtuigen dat hij/ zij deze kennis heeft én ermee weet te werken?
  Uw antwoord op deze vragen bevat waarschijnlijk een concreter cursusdoel, waarin kennis en inzicht (*Knowledge* en *Understanding*-niveau bij Bloom) voorwaardelijk is om te komen tot een hoger cognitief niveau. Gebruik bijvoorbeeld één van de werkwoorden van Analyse, Evaluatie of Synthese.
  Zie hieronder bij: Toelichting Academisch niveau.

4. Heeft u momenteel cursusdoelen met daarin 'benoemen', 'herkennen', 'vergelijken', 'beschrijven'?

- Deze werkwoorden duiden erop dat uw cursusdoelen op *Knowledge*- en *Understanding*-niveau liggen. Voor inleidende cursussen kan dat een passend niveau zijn.
- Voor vervolgcursussen is een dergelijk niveau veelal niet voldoende. Als u dit wilt verhelpen, is het handig om te proberen om de volgende zin voor uzelf af te maken: De student moet benoemen/ herkennen/ vergelijken/ beschrijven zodat hij/ zij daarna X en Y kan. Het afmaken van deze zin leidt waarschijnlijk tot een cursusdoel van een hoger niveau; zeker als u in de zin één van de werkwoorden van Analyse, Evaluatie of Synthese gebruikt.
  Zie hieronder bij: Toelichting Academisch niveau

Concreet, met behulp van SMART of RUMBA

U maakt uw cursusdoelen concreter door te letten op de zogeheten SMART-criteria:

- Specifiek
  Is de formulering in concrete en begrijpelijke termen? Is de context duidelijk?
- Meetbaar
  Als het doel bereikt is moet het resultaat meetbaar zijn. Is dat terug te vinden in de formulering van het doel? Hebben studenten bijv. een idee wat inhoudelijk van hen gevraagd zal worden bij toetsing?
- Acceptabel
  De doelformulering zal door studenten erkend worden als zinvol en relevant gezien de eigen leerbehoeften in relatie tot de opleiding.
- Realistisch
  Is de formulering zodanig dat gesproken kan worden van een haalbaar doel gezien het niveau, bijvoorbeeld de voorkennis, van de studenten?
- Tijdgebonden
  Zullen de doelen te halen zijn in de tijd die ervoor gereserveerd is? Maar ook: de cursus is op het goede moment geplaatst gezien het curriculum.

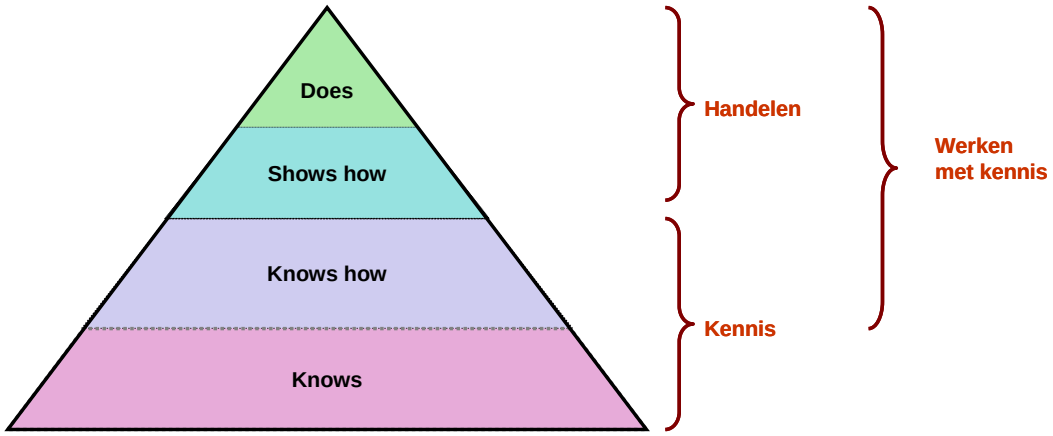
Er doen meerdere lijstjes de ronde, bijvoorbeeld RUMBA: *relevant, understandable, measurable, behavioral.* Alle leiden tot een vergelijkbaar resultaat doordat ze u helpen een middenweg te vinden tussen leerdoelen die te abstract, onbegrijpelijk of saai zijn; en leerdoelen die te populair of te toepassingsgericht te zijn. Ook helpen ze u na te gaan of de doelen het goede niveau hebben.

Academisch niveau

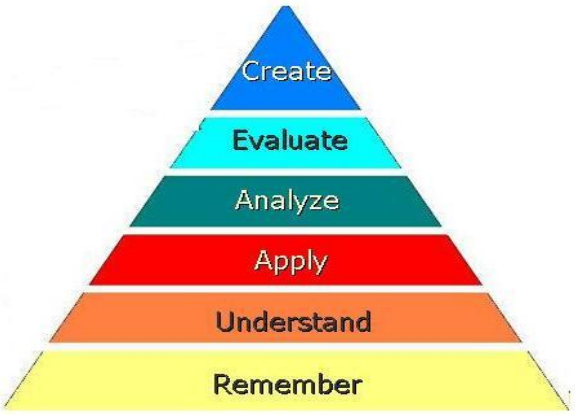
Hoger onderwijs is erop gericht om studenten betekenisgericht of toepassingsgericht te laten leren, in plaats van reproductiegericht of ongericht. Dat betekent dat docenten de studenten laten *werken met kennis*, m.a.w. docenten richten hun cursusdoelen (en daarna de toetsing) op de zogeheten hogere orde-denkvaardigheden (metacognitieve vaardigheden).

We gebruiken hier twee hulpmiddelen om cursusdoelen op het juiste academisch niveau te formuleren: de piramide van Miller (1990) en de (herziene) taxonomie van Bloom (Anderson & Krathwohl, 2001). Deze staan hieronder afgebeeld.

Academische cursusdoelen dienen (minimaal) op het niveau van **'werken met kennis'** te liggen (Miller) ofwel op het niveau van **Toepassing en hoger** (Apply, Analyze, Evaluate en Create) (Anderson & Krathwohl).
Bij de verschillende cognitieve niveaus horen typerende werkwoorden (zie hieronder). Deze kunt u gebruiken om uw leerdoelen actief te formuleren: de student doet iets



Piramide van Miller (1990)



Revised Bloom Taxonomy, Anderson & Krathwohl, 2001

Toepassen

Het uitvoeren of gebruiken van een procedure, model, theorie op (voor de student) nieuwe en concrete situaties of problemen. Kenmerkende vragen: hoe werkt dat hier dan? Wat is hierbij nodig?
Kenmerkende *werkwoorden* zijn: kiezen, demonstreren, construeren, verrichten, voorspellen, vertalen, gebruiken, uitvoeren, implementeren etc.

Analyseren

Het opsplitsen van een groter geheel in onderdelen waaruit het is samengesteld, het stap voor stap uitzoeken welke verschillende aspecten aan een probleem, gedachtegang of theorie zijn te onderscheiden. Kenmerkende vragen: hoe zit dat, welke delen zijn van groter/kleiner belang, hoe is het te ordenen?
Kenmerkende *werkwoorden* zijn: selecteren, vergelijken, contrasteren, onderzoeken, categoriseren, classificeren, onderscheiden etc.

Evaluatie

Zich een oordeel vormen gebaseerd op criteria and standaarden/ methoden, middels het controleren van feiten en het onderzoeken en kritiseren van vooronderstellingen.
Kenmerkende *werkwoorden* zijn: beoordelen, toetsen, kritiseren, ondersteunen, verdedigen, onderbouwen.

Synthese, iets nieuws bedenken/ ontwikkelen, kritisch beschouwen, ontwikkelen. Het meedenken met auteurs, docenten en medestudenten, een eigen inbreng hebben en niet zomaar alles accepteren wat geschreven staat of gezegd wordt.
Kenmerkende *werkwoorden* zijn: combineren, herformuleren, integraal samenvatten, argumenteren, afleiden, generaliseren, concluderen, bekritiseren, probleem oplossen, innoveren, beslissen, adviseren etc.

Bijlage II: Berekeningswijzen: 'Absolute beoordeling' en 'absolute beoordeling met relatieve component'

Absolute beoordeling

De student moet, om een voldoende te halen, meer dan de helft van de vragen goed beantwoorden, nadat gecorrigeerd is voor de gokkans. De formule voor het absoluut normeren is dus:

((5,6 – 1) * (M – T) /9) + T = aantal vragen goed voor een voldoende

Waarbij:
5,6 – 1 = Voldoende min 1 punt, aangezien er niet lager dan een 1 kan worden gehaald.
M = Maxima haalbare score
T = Totaal aantal vragen / aantal antwoordalternatieven (= verwacht aantal vragen goed obv gokkans)
9 = Dit is de range waarbinnen je punten wilt toekennen (we rekenen namelijk van 1-10).

Rekenvoorbeeld:
Stel dat een tentamen uit 60 vierkeuze-vragen bestaat. De gokkans is 0.25, dus op basis van gokken zullen gemiddeld 15 vragen goed beantwoord worden[21]. De volgende score wordt dan gewaardeerd met een 5,6 en na afronding een 6:

((5,6 - 1) * (60-15)/9) + 15 = 38

Absolute beoordeling met relatieve referentie

Bij een absolute cesuur met relatieve referentie wordt niet uitgegaan van een theoretische maximumscore (alle vragen goed), maar van de in de praktijk 'haalbaar' gebleken maximumscore. In de literatuur wordt dat daarvoor als definitie aangehouden: de gemiddelde score van de beste 5% op de betreffende toets. De volgende formule geeft aan welke score tot een voldoende (5.6) leidt:
In bovenstaande formule wordt dan de in de praktijk haalbare maximumscore ingevuld op de plaats van de theoretische score (M).

Rekenvoorbeeld:
Stel dat een tentamen uit 60 vierkeuze-vragen bestaat. Aan de frequentieverdeling van de toetsscores kun je de gemiddelde score van de beste 5% van de studenten zien, bijvoorbeeld 58. De gokkans is 0.25, dus op basis van gokken zullen gemiddeld 15 vragen goed beantwoord worden. De volgende score wordt dan gewaardeerd met een 6:

((5,6 - 1) * (58-15)/9) + 15 = 37

[21] Van Berkel et all., 2013.

Cijfers toekennen

Met gebruik van dezelfde formule is het ook mogelijk om aan alle ruwe scores een cijfer tussen 1 en 10 toe te kennen. De aangepaste formule die hiervoor gebruikt kan worden is als volgt:

$$\text{Cijfer} = \frac{(X-A)}{(T-A)/9} + 1$$

Waarbij:
T = Gemiddelde score beste 5% (bij gebruik van relatieve referentie) OF Theoretische maximaal haalbare score (alle vragen goed) (bij gebruik van absolute beoordeling)
A = Totaal aantal vragen / aantal antwoordalternatieven
X = Aantal goed beantwoord (ruwe score).

Rekenvoorbeeld:
Het gaat om een tentamen met 60 vierkeuzevragen. Een student heeft 37 vragen goed.
T = 58
A is 60/4 = 15
X = 37

Het cijfer is dan:

$$\frac{(37 - 15)}{(58 - 15)/9} + 1 = 5.6$$

Door deze formule in SPSS of Excel te zetten, worden alle ruwe scores van studenten omgezet in cijfers tussen 1 en 10. Een Excel-programma hiervoor is beschikbaar bij het Kwaliteitszorgteam.