ToetsSpecial

# Toetsconstructieproces in 8 stappen

Een handleiding voor toetsconstructeurs met informatie over de acht stappen van het toetsconstructieproces

Henk Moelands | Cito

ToetsWijzer

# Toetsconstructieproces in 8 stappen

## Een handleiding voor toetsconstructeurs met informatie over de acht stappen van het toetsconstructieproces

Auteur: Henk Moelands (Cito)

september 2017 (herziene versie), oorspronkelijke uitgave 2002


info@toetswijzer.nl
www.toetswijzer.nl

Bij samenstelling van dit dossier is gebruik gemaakt van het boek Psychometrie in de Praktijk onder redactie van Eggen en Sanders (Cito, 1993).

# Inhoud

# Inleiding

**Toetsen spelen een belangrijke rol in het onderwijs. Op basis van de resultaten op toetsen worden allerlei beslissingen genomen. Deze beslissingen kunnen betrekking hebben op leerlingen (micro-niveau), op de school (meso-niveau) en op het onderwijs op landelijk niveau (macro-niveau).**

Toetsen hebben gemeenschappelijk dat zij informatie aandragen voor het nemen van beslissingen, maar de aard van de informatie en de soort beslissingen die zij ondersteunen zijn verschillend. Gegevens op toetsen voor de ene beslissing kunnen ongeschikt zijn voor het nemen van andere beslissingen. De functie van een toets heeft consequenties voor het toetsconstructieproces en voor de beslissingen die op basis van de resultaten op de betreffende toets genomen mogen worden. Drie belangrijke vragen die ten grondslag liggen aan het toetsconstructieproces zijn:

- Op welk niveau heeft de te nemen beslissing betrekking?
- Wat is het gebruiksdoel van de te ontwikkelen toets?
- Wat zijn de consequenties van de resultaten op de toets en voor wie?

Het toetsconstructieproces kan in het algemeen uiteengelegd worden in een achttal stappen. De uitwerking van de stappen is afhankelijk van het doel van de toets en de randvoorwaarden waarbinnen de constructie dient plaats te vinden. De acht stappen zijn:

1. Doelspecificatie
2. Toetspecificatie
3. Itemconstructie
4. Toetsafname
5. Itemevaluatie
6. Toetssamenstelling
7. Referentiekader
8. Handleiding en verantwoording

# 1. Doelspecificatie

**De eerste fase van het constructieproces bestaat uit het operationaliseren van de vaardigheid die de toets moet meten en het vaststellen van het gebruiksdoel van de toets. Op het tweede genoemde punt gaat deze site verder in. In eerste instantie komt een aantal functies dat aan toetsen gesteld wordt aan bod, met daaraan gekoppeld drie van elkaar te onderscheiden categorieën meetinstrumenten. Vervolgens wordt nader ingegaan op vier belangrijke kwaliteitseisen die in het algemeen aan toetsen gesteld kunnen worden. Tot slot komt de relatie aan de orde tussen de onderscheiden functies van meetinstrumenten en de daaraan te stellen eisen.**

## 1.1 Functies van toetsen

De functies die men aan toetsen toekent, zijn van invloed op de eisen die aan toetsen worden gesteld. Dat men aan toetsen vele (bedoelde en onbedoelde) functies toekent, laat de onderstaande (tentatieve) inventarisatie zien:

- ***Selectiefunctie:*** afhankelijk van de resultaten op een toets wordt beslist een leerling al of niet toe te laten tot een vervolgtraject.
- ***Allocatiefunctie:*** voorbereiden voor en verwijzen naar bepaalde positie in de maatschappij.
- ***Kwaliteitsbewakingsfunctie:*** voldoen aan bepaalde standaarden. Geeft rechten en/of erkenning. Voor de overheid zijn toetsen in het algemeen en met name examens belangrijke instrumenten voor de kwaliteitsbewaking van het onderwijs.
- ***Prognostische functie:*** doen van een uitspraak over de kans om een bepaald vervolgtraject met succes te volgen. Als een leerling met goed gevolg een opleiding heeft afgesloten, verwacht men dat hij of zij over kennis en vaardigheden beschikt die rechtvaardigt dat de leerling een bepaalde vervolgstudie aanvangt.
- ***Communicatiefunctie:*** middel om met betrokkenen te praten over wat de opleiding inhoudt. Programma's, syllabi en toetsen - en de daaraan verbonden resultaten - verschaffen buitenschoolse personen informatie over wat de opleiding inhoudt en wat de individuele leerling kent en kan.
- ***Operationalisatiefunctie:*** toetsen zijn de meest concrete uitwerking van nagestreefde onderwijsdoelen. Toetsen zijn het middel om aan te geven wat als belangrijk gezien wordt. Toetsen zijn richtinggevend voor het onderwijs.
- ***Didactische functie:*** voorziet belanghebbenden van informatie over het onderwijsleerproces. Stuurt het onderwijsleerproces. Scholen kunnen met toetsen nagaan in hoeverre nagestreefde doelen gehaald zijn en of bijstelling van het onderwijsprogramma noodzakelijk is.
- ***Kwalificerende functie:*** geeft inzicht in de standaarden waaraan voldaan wordt. Over welke kennis en vaardigheden bezit een afgestudeerde aan een bepaalde opleiding.
- ***Evaluatiefunctie:*** geeft een bijdrage aan de beoordeling van de kwaliteit van het onderwijs.

### Drie functies van toetsen

Het voorgaande laat zien dat toetsen voor vele gebruiksdoelen ingezet (kunnen) worden. Uitgaande van het gebruiksdoel van een toets (in het algemeen: onderwijskundig meetinstrument) kunnen drie categorieën onderscheiden worden (Moelands en Sanders, 1996):

1. ***Meetinstrumenten die tot doel hebben de kwaliteit van een onderwijssysteem te beoordelen***
   Met een onderwijssysteem wordt hier zowel een individuele school (meso-niveau) als een verzameling scholen (macro-niveau), bijvoorbeeld een onderwijstype, bedoeld. Een voorbeeld van deze categorie is

het periodiek peilingsonderzoek voor basis- en speciaal onderwijs (PPON), waarmee een grondslag wordt geboden voor een brede discussie over de inhoud en het niveau van het onderwijs op landelijk niveau.

2. ***Meetinstrumenten die tot doel hebben de kwaliteit van de leerling te beoordelen***
   Binnen deze categorie wordt een onderscheid gemaakt tussen meetinstrumenten die bedoeld zijn voor selectie, classificatie, plaatsing of certificering (beheersing) (zie kader voor toelichting).

**Selectie**
Van selectie is sprake als afhankelijk van de resultaten op een meetinstrument beslist wordt een leerling al of niet toe te laten tot een onderwijstraject.

**Classificatie**
Bij classificatie wordt afhankelijk van de resultaten op een meetinstrument beslist welk onderwijstraject een leerling moet volgen. De te onderscheiden trajecten zijn kwalitatief verschillend van aard, bijvoorbeeld wiskunde en natuurkunde. Voor beide vakken worden dan ook verschillende criteria gehanteerd om vast te stellen of een leerling succes heeft gehad in het desbetreffend vak.

**Plaatsing**
Van plaatsing is sprake in het geval leerlingen binnen eenzelfde onderwijstraject op basis van de prestaties op een meetinstrument bepaalde leerroutes kunnen of dienen te volgen. Bij plaatsing gaat het dus om kwalitatief verschillende behandelingen met voor alle behandelingen hetzelfde criterium. Vergelijk: differentiëren binnen klassenverband.

**Certificering**
Van certificering is sprake wanneer afhankelijk van de prestaties op een meetinstrument beslist wordt of een leerling een onderwijstraject met succes doorlopen heeft.

3. ***Meetinstrumenten die tot doel hebben het onderwijsleerproces te beoordelen***
   Dit betreft meetinstrumenten op basis waarvan een docent kan besluiten het onderwijs aan een (groep) leerlingen anders in te richten. Als voorbeeld kunnen de toetsen genoemd worden die ontwikkeld zijn in het kader van het leerlingvolgsysteem voor het basisonderwijs, zoals het Volgsysteem primair en speciaal onderwijs van Cito. Deze instrumenten hebben primair tot doel het onderwijs optimaal te laten aansluiten bij de (cognitieve) ontwikkeling van de leerling.

## 1.2 Kwaliteitseisen

De functies die aan toetsen worden toegekend, komen alleen dan tot hun recht als de toetsen aan bepaalde kwaliteitseisen voldoen. Vier belangrijke eisen zijn:

**Betrouwbaarheid**
Een toets wordt betrouwbaar genoemd als het bij herhaalde afname onder dezelfde omstandigheden eenzelfde resultaat laat zien. Drie belangrijke factoren bij betrouwbaarheid zijn:

- ***De kwaliteit van de toets zelf***
  De opgaven moeten helder en eenduidig geformuleerd zijn en er mag geen twijfel bestaan over het soort antwoord dat van de kandidaat verwacht wordt. Extreem moeilijke of gemakkelijke opgaven moeten vermeden worden. De toets in z'n geheel moet een onderscheid maken tussen de 'goede' en 'zwakke' kandidaten, hetgeen ook geldt voor de afzonderlijke opgaven. De moeilijke opgaven moeten vooral door de goede kandidaten goed gemaakt worden. De betrouwbaarheid wordt ook beïnvloed door het aantal

opgaven dat een toets bevat. Bij een toets met weinig opgaven is de invloed van elke afzonderlijke opgave veel groter dan bij een toets met veel opgaven.

- ***De omstandigheden waaronder de toets wordt afgenomen***
  Belangrijk in dit kader is standaardisatie en objectiviteit. Omstandigheden kunnen op velerlei zaken betrekking hebben: lokale omstandigheden (het in rust kunnen maken van een toets), gebruik van hulpmiddelen, relatie lengte van de toets met de beschikbare toetstijd. Het is belangrijk dat een toets niet ontaardt in een 'race tegen de klok'.

- ***De wijze waarop de resultaten worden beoordeeld***
  Ook bij de beoordeling van de resultaten spelen standaardisatie en objectiviteit een belangrijke rol. Het resultaat van een kandidaat op een toets kan sterk bepaald worden door de beoordelaar. Zo blijkt bijvoorbeeld uit onderzoek dat een beoordelaar vaak beïnvloed wordt door de vorige beoordelingen. Eenzelfde antwoord krijgt een hogere waardering als de voorgaande kandidaten veel slechte antwoorden geven en een lagere beoordeling als voor de voorgaande kandidaten goede antwoorden geven. Een mogelijke oplossing voor dat probleem is de correctie voor zover mogelijk te standaardiseren en te objectiveren. Bijvoorbeeld door gebruik te maken van gesloten vragen. Bij open vragen zijn zo eenduidig mogelijke correctievoorschriften van belang.

### Validiteit

De validiteit van een toets is de eigenschap dat de toets meet wat de constructeur bedoeld heeft ermee te meten. Aangezien een toets veel en uiteenlopende bedoelingen kan hebben zijn er evenzoveel validiteiten te onderscheiden die een toets in verschillende mate kan bezitten. In de literatuur komen veel verschillende benamingen en definities van validiteiten voor. De methoden voor het bepalen van validiteiten zijn te verdelen in:

- methoden die door middel van correlatieberekening de relatie bepalen tussen de scores en een criterium (o.a. predictieve validiteit);
- methoden die tot een uitspraak leiden over de relatie tussen de toetsvragen en de onderwijsdoelstellingen (inhoudsvaliditeit);
- methoden die het toetsgedrag verklaren op grond van een onderliggende trek (begripsvaliditeit);
- methoden die de geldigheid van gevolgtrekkingen op grond van scores vaststellen door middel van experimenten;
- overige methoden waaronder het op het oog vaststellen van de relatie tussen de toets en hetgeen de toets pretendeert te meten (indruksvaliditeit).
  In onderwijssituaties wordt een belangrijke rol toegedeeld aan de inhoudsvaliditeit van een toets.

### Aanvaardbaarheid

Het aanvaardbaar maken van beslissingen op grond van een toets is geen eenvoudige zaak en hangt mede af van het belang van de betreffende toetsing. Naast eisen ten aanzien van betrouwbaarheid en validiteit, dient een toets ook voor betrokkenen 'doorzichtig' te zijn ten aanzien van: het kiezen van de te volgen strategie tijdens de voorbereiding, beoordeling en waardering van de toetsresultaten (correctievoorschriften en de procedure voor de bepaling van het cijfer).

### Transparantie

Transparantie betreft vooral de inhoud van de vragen en de berekening van het cijfer. En wat dat laatste betreft, met als voorbeeld een examen, vooral de vraag hoeveel punten nodig zijn voor een voldoende. Bij de inhoud is het van belang bij de formulering van vragen erop te letten dat deze duidelijk verwijzen naar herkenbare begrippen, situaties of vaardigheden zoals die beschreven staan in examenprogramma's. Het

gebruik van standaardformuleringen bevordert de duidelijkheid in dit opzicht. Ook syllabi kunnen bijdragen aan de duidelijkheid van de inhoud van examens. Hetzelfde geldt voor zogeheten voorbereidingstoetsen.

## 1.3 Kenmerk en doel van meetinstrumenten

Uitgaande van het gebruiksdoel van een meetinstrument zijn drie algemene functies van meetinstrumenten te onderscheiden. Meetinstrumenten die tot doel hebben:

| Gebruiksdoel | Functie | Voorbeelden |
|---|---|---|
| ***De kwaliteit van een onderwijssysteem te beoordelen*** | ➨ Monitoren en/of het beoordelen van de kwaliteit van een onderwijssysteem op basis van gegevens van (geaggregeerde) groepen van leerlingen | - *Nemen van een besluit over onderwijsprogramma's en veranderingen van curricula*<br>- *Evalueren van experimentele en innovatieve programma's*<br>- *Rapportage aan onderwijsbetrokkenen (inspectie, overheid, ouders, besturen) over de kwaliteit van het onderwijs* |
| ***De kwaliteit van de leerling te beoordelen*** | ➨ Beoordelen van de kwaliteit van individuele leerlingen | - *Selectie: toelatingsexamen voor een bepaalde school (conservatorium)*<br>- *Plaatsing: differentiatie binnen klassenverband (remedial teaching, hoogbegaafden)*<br>- *Classificatie: streaming (mavo of havo, havo of vwo)*<br>- *Certificering: toekennen van een diploma of een deelcertificaat* |
| ***Het onderwijsleerproces te beoordelen*** | ➨ Voor het monitoren van de voortgang van individuele leerlingen met als doel het onderwijs zo optimaal mogelijk aan te passen aan de leerlingen | - *Vaststellen sterke en zwakke kanten van leerlingen*<br>- *Aanpassen curriculum aan voortgang van leerlingen*<br>- *Nagaan in welke mate leerlingen onderwezen lesstof beheersen* |

Het gebruiksdoel is mede bepalend voor de aan de instrumenten te stellen eisen. Instrumenten met grote consequenties voor leerlingen (zak-/slaagbeslissingen) dienen - in het algemeen - aan strengere eisen te voldoen dan meetinstrumenten die bedoeld zijn voor het managen van het onderwijs in de dagelijkse klassenpraktijk.

Onderstaand overzicht geeft de relatie weer tussen de onderscheiden functies van meetinstrumenten en de daaraan te stellen eisen. Door in het overzicht op de vetgedrukte kenmerken te klikken, komt u bij een nadere toelichting.

| KENMERK INSTRUMENT | Beoordeling van ONDERWIJSLEERPROCES | Beoordeling van ONDERWIJSSYSTEEM | Beoordeling van LEERLING |
|---|---|---|---|
| **A. Rapportage-eenheid** | Individuele leerlingen | Groepen leerlingen | Individuele leerlingen |
| **B. Consequentie beslissing** | Laag | Laag of hoog | Hoog |
| **C. Toetskarakteristieken:** | | | |
| - vergelijkbaarheid van informatie | Laag | Hoog | Hoog |
| - objectieve scoring | Nee | Ja | Ja |
| - gestandaardiseerde afname | Nee | Ja | Ja |
| **D. Aard informatie:** | | | |
| - gedetailleerd/algemeen | Gedetailleerd | Gedetailleerd | Afhankelijk van specifieke functie |
| - frequentie | Hoog | Laag | Laag |
| - resultaten vlug nodig | Ja | Nee | Afhankelijk van specifieke functie |
| **E. Technische eisen:** | | | |
| - hoge betrouwbaarheidseisen | Afhankelijk van specifieke functie | Afhankelijk van steekproefgrootte | Ja |
| - validiteit | Inhoudsvaliditeit | Met name construct | Plaatsing, classificatie en selectie (predictieve validiteit)<br>Certificering (inhoudsvaliditeit) |

### A. Rapportage-eenheid

Zowel bij het verkrijgen van informatie over het onderwijsleerproces als bij het beoordelen van individuele leerlingen dienen gegevens over elke afzonderlijke leerling verzameld te worden. De leerling is ook de eenheid waarop gerapporteerd wordt. Bij het doen van een uitspraak op landelijk niveau is het niet nodig om gegevens te verzamelen bij elke afzonderlijke leerling, maar kan volstaan worden met een steekproef. Door gerichte steekproeftrekking is het mogelijk uitspraken te generaliseren naar landelijk niveau. Voor het verzamelen van gegevens is dit uit kostenoverweging erg belangrijk.

### B. Consequentie beslissing

De consequentie van de te nemen beslissing is bij het beoordelen van het onderwijsleerproces niet hoog. Een lage beoordeling heeft geen desastreuze gevolgen voor de leerling. De beoordeling maakt vaak deel uit van een veelheid aan beoordelingen. Vaak zijn dit korte toetsjes om na te gaan of een leerling een bepaalde hoeveelheid lesstof voldoende beheerst. Een foutieve beoordeling kan een volgende maal weer hersteld worden.

In het algemeen is de consequentie van de beslissing op landelijke niveau laag. Een uitspraak als 'het onderwijsniveau rekenen in Nederland neemt af' zal geen consequentie hebben voor individuele leerlingen. Zij zullen daarop niet afgerekend worden. Een dergelijke uitspraak is meer een signaal voor het beleidsniveau om gerichte acties te ondernemen. Het beoordelen of een actie effect heeft, vraagt in de regel veel tijd.
Als de school als rapportage-eenheid gezien wordt bij het rapporteren op systeemniveau, dan kunnen de consequenties hoog zijn. Mocht een school afgerekend worden (accountability) op zijn prestaties dan kunnen lage prestaties enorme consequenties hebben. Vooral als bijvoorbeeld de overheidsfinanciering gekoppeld is aan bepaalde prestaties.

De consequenties bij het beoordelen van individuele leerlingen zijn hoog. Een foutieve beslissing kan voor een leerling desastreuze gevolgen hebben. Zo kunnen bepaalde vervolgopleidingen afgesneden worden (ingeval van een selectietoets) of kan een maatschappelijk erkend diploma niet toegekend worden.

## C. Toetskarakteristieken

*Vergelijkbaarheid van informatie*

Voor de dagelijkse voortgangscontrole is het niet van belang dat parallelle groepen dezelfde toets voorgelegd krijgen. Het gaat bij deze controle om de interactie tussen leerkracht, leerlingen en onderwezen lesstof. Op basis van de resultaten kan een leerkracht besluiten zijn of haar onderwijs voor zijn of haar leerlingen anders in te richten. Voor het doen van uitspraken op systeemniveau ligt dit anders. Wanneer vastgesteld moet worden of het onderwijsniveau in Nederland in de tijd verandert, zal ervoor gezorgd moeten worden dat de uitspraak op de zelfde lesstof betrekking heeft.
Dat de vergelijkbaarheid van informatie bij beslissingen over leerlingen van belang is spreekt voor zich. Het mag niet zo zijn dat aan examenkandidaten van een bepaalde opleiding in een bepaald jaar andere eisen gesteld worden dan aan de kandidaten van andere jaren.

*Objectieve scoring*

Bij het doen van een uitspraak over het onderwijsleerproces in een klas gaat het om de voortgang van het onderwijs. De gewenste informatie is met name bedoeld om de leerkracht (en de leerling) te informeren over het proces om vast te stellen of (kleine) aanpassingen in het programma gewenst zijn. Objectiviteit is geen harde eis.
Bij een beoordeling van een onderwijssysteem is objectieve beoordeling wel een vereiste. De uitspraken kunnen leiden tot aanpassingen van het onderwijs op grote schaal. Deze uitspraken mogen derhalve geen subjectief karakter hebben. Betrokkenen moeten het onafhankelijk van elkaar eens zijn over bijvoorbeeld de kwaliteit van het onderwijs. Dat de beoordeling van de resultaten over leerlingen waaraan grote consequenties vastzitten objectief moeten zijn, is evident. Het wel of niet slagen van een leerling mag niet afhankelijk zijn van het (subjectieve) oordeel van een leerkracht. Andere leerkrachten zouden tot een zelfde beoordeling moeten komen.

*Gestandaardiseerde afname*

Voor de voortgang van het onderwijsproces in de klas is standaardisatie in afname niet relevant. Het is aan de leerkracht vast te stellen wat het meest geschikte tijdstip is om een toets af te nemen. Dit tijdstip mag tussen scholen verschillen. Wil men een uitspraak doen over de kwaliteit van het onderwijs op een bepaald tijdstip dan dienen de te verzamelen gegevens wel onder gestandaardiseerde condities afgenomen te worden. Storende factoren die de resultaten van leerlingen op toetsen kunnen beïnvloeden dienen vermeden te worden. Met name als de resultaten verzameld worden bij meerdere groepen. Alleen al uit oogpunt van vergelijkbaarheid en gelijkwaardigheid dienen toetsen waaraan consequenties voor leerlingen vastzitten gestandaardiseerd te zijn. Leerlingen mogen niet bevoordeeld of benadeeld worden door verschillen in afnamecondities.

## D. Aard informatie

*Gedetailleerd/algemeen*

Voor het optimaliseren van het onderwijsleerproces is het voor een leerkracht van belang dat hij/zij over gedetailleerde informatie beschikt. De leerkracht dient inzicht te hebben in wat de leerling nu wel of niet beheerst. Op basis van deze informatie kan hij/zij besluiten (gerichte) extra hulp te geven, het onderwerp nogmaals aan bod te laten komen of voort te gaan met het curriculum.

Ook voor het beoordelen van het onderwijssysteem is gedetailleerde informatie gewenst. Om een uitspraak te doen over het onderwijsniveau van een bepaald vak, zal aangegeven moeten worden over welke onderdeel het nu gaat en wat de leerlingen nu nog wel en wat zij nog niet kennen of beheersen.

Ingeval er sprake is van het beoordelen van een leerling dan is de mate van gedetailleerdheid afhankelijk van de specifieke functie van de toets. Zo is bij certificering algemene informatie gewenst. Het gaat erom om aan te geven of een leerling de lesstof voldoende beheerst om in aanmerking te komen voor een certificaat. Het gaat bij deze vorm van toetsen niet meer om het vaststellen van eventuele hiaten. Anders ligt het bij plaatsingstoetsen. Op basis van specifieke kenmerken van leerlingen wordt bepaald wat het beste vervolgtraject voor een leerling is. Het gaat dan niet meer om een algemeen oordeel, maar om een oordeel die ingaat op die onderdelen die relevant zijn voor de te nemen beslissing.

*Frequentie*

Toetsen die tot doel hebben het onderwijsleerproces zo optimaal mogelijk af te stemmen op de leerlingen dienen uit dat oogpunt frequent te worden afgenomen. Door continue informatie over het verloop van het proces, kan een leerkracht het onderwijs bijstellen.

Uitspraken op het niveau van een onderwijssysteem doorlopen vaak een lange cyclus voordat een beslissing genomen wordt. Bovendien zijn effecten van doorgevoerde beleidswijzigingen niet direct zichtbaar. Voor de onderzoeken die afgenomen worden in het kader van PPON is gekozen voor een cyclus van vijf jaren. Beslissingen die genomen worden voor certificering, plaatsing, classificatie en selectie vinden niet zo vaak plaats. De frequentie van deze toetsen is dan ook laag.

*Resultaten vlug nodig*

Om het onderwijsleerproces in de klas zo optimaal mogelijk aan te laten sluiten bij het niveau van de leerlingen is het van belang dat de resultaten op toetsen zo vlug mogelijk bekend zijn. Met name ook omdat deze toetsen met een relatief hoge frequentie worden afgenomen. Resultaten op toetsen voor het beoordelen van de kwaliteit van het onderwijs op systeemniveau, zijn niet vlug nodig. Vaak kan dit ook niet vanwege de complexiteit waarmee deze vorm van toetsen gepaard gaat. Of de resultaten voor het beoordelen van individuele leerlingen vlug nodig zijn, is afhankelijk van de specifieke doelstelling van de toets. Als er sprake is van modulair opgezet onderwijs dan dienen de resultaten in principe vlug beschikbaar te zijn. Bij andere vormen is wellicht wat meer tijd beschikbaar.

## E. Technische eisen

*Betrouwbaarheidseisen*

Toetsen voor de voortgang van het onderwijsproces in de klas zijn vaak bedoeld om een indruk te krijgen hoe verder te gaan. De leerkracht krijgt informatie of hij/zij verder kan met de lessen of dat bepaalde onderdelen nog extra aandacht behoeven. In de regel behoeft de betrouwbaarheid van dergelijke toetsen niet hoog te zijn. De beslissingen hebben voor de leerlingen geen grote consequenties. Als het om toetsen gaat die tot doel hebben de leerlingen in de tijd te volgen dan dienen de toetsen wel aan hoge toetstechnische eisen te voldoen.

Bij beslissingen op landelijk niveau (onderwijssysteem) is het van belang om een betrouwbare uitspraak te krijgen. De betrouwbaarheid wordt mede bepaald door de grootte van de steekproef. Ingeval er sprake is van grote steekproeven dan zijn de eisen ten aanzien van de betrouwbaarheid 'geringer' dan bij kleine steekproeven. Bij kleinere steekproeven dienen de toetsen aan hogere betrouwbaarheidseisen te voldoen om 'staat te kunnen maken' op de toetsresultaten.

Bij uitspraken over individuele leerlingen is het belangrijk het aantal misclassificaties op basis van de resultaten op toetsen zoveel mogelijk te beperken. Voor een leerling is er veel aan gelegen dat op basis van zijn resultaten betrouwbare uitspraken gedaan worden. Een hoge betrouwbaarheid van de toetsen is een *must*.

*Validiteit*

Bij de voortgang van leerlingen gaat het om de vraag in welke mate zij een bepaalde hoeveelheid onderwezen lesstof beheersen. De toets dient dan ook inhoudsvalide te zijn ten opzichte van het lesprogramma. Bij uitspraken op systeemniveau gaat het om uitspraken die verwijzen naar (specifieke) onderdelen van een vak en/of vaardigheden. Het betreft uitspraken hoe leerlingen het doen in een bepaald vak of op onderdelen van een vak. Het is van belang dat de toetsen een uitspraak over die onderdelen rechtvaardigen. Van belang in dat kader is met name de constructvaliditeit.

Bij een oordeel over leerlingen wordt de aard van de beslissing bepaald door de specifieke functie van een toets. Bij certificering gaat het om de vraag in hoeverre een leerling bepaalde eindtermen (bijvoorbeeld kerndoelen basisonderwijs en examenprogramma's vo) beheersen. In dat geval gaat het om inhoudsvaliditeit. Bij plaatsing, classificatie en selectie staat de vraag centraal in hoeverre een leerling in staat is een bepaald vervolgtraject met succes af te ronden. Het betreft hier een vraag naar een toekomstig onderwijsprogramma. In dat geval ligt het accent op de predictieve validiteit.

# 2. Toetsspecificatie

**Nadat helder is wat de functie van de te ontwikkelen toets dient te zijn, moet aangegeven worden op welk domein de vragen uit de toets betrekking hebben en wat eventuele randvoorwaarden voor toetsconstructie en -afname zijn.**

Relevante vragen voor toetsconstructeurs in dit kader zijn (Eggen en Sanders, 1993):

- **Bij *wie* wordt de toets afgenomen?**
  Belangrijk is te weten bij welke personen de toets met welk doel wordt afgenomen. Het toetsconstructieproces zal anders verlopen wanneer het een toets betreft voor een heterogene groep personen voor een diploma of certificaat, dan wanneer het een toets betreft voor een homogene groep personen met het doel om de meest vaardige personen te selecteren.
- **Hoeveel *toetstijd* is er beschikbaar?**
  Leerlingen moeten ruim de tijd krijgen voor het beantwoorden van de toets. Te weinig beschikbare toetstijd kan betekenen dat te weinig vragen afgenomen kunnen worden om de gewenste vaardigheid te meten. Bovendien kan dit tot gevolg hebben dat andere vaardigheden dan bedoeld een rol gaan spelen bij het beantwoorden van de toetsvragen, zoals bijvoorbeeld de snelheid van uitvoering.
- ***Hoe* wordt de toets afgenomen?**
  Kiest men voor een individuele benadering, of wordt het een groepsafname? In het laatste geval ligt een keuze voor een schriftelijke toetsing waarschijnlijk meer voor de hand. Ontwikkelingen op het gebied van ict maken het mogelijk om vragen via het beeldscherm te presenteren, de antwoorden in de computer in te voeren en te laten scoren. Door deze mogelijkheid wordt individuele toetsafname niet alleen minder bezwaarlijk maar kan voor bepaalde toepassingen zelfs grote voordelen hebben.
- **Wat is de *inhoud* van de toets?**
  Het vaststellen van de inhoud van de toets is de belangrijkste toetsspecificatie. Voor deze specificatie wordt bij studietoetsen gebruik gemaakt van een toetsmatrijs. Aan de hand van een toetsmatrijs wordt vastgesteld hoe de vragen uit de toets verdeeld zullen worden over de inhouds- en gedragscategorieën. Ingeval de toets een uitspraak dient te doen over de mate van beheersing van een hoeveelheid onderwezen stof, dan zal de inhoud van de toets een goede representatie van deze stof dienen te zijn. Ingeval het referentiekader extern is, bijvoorbeeld bij selectie[1] en classificatie[2], dan zullen de kennis en vaardigheden in de toets aan bod dienen te komen die van belang zijn voor het toekomstig onderwijstraject.
- **In welke *vorm* wordt de toets afgenomen?**
  Wanneer de vaardigheid met een schriftelijke toets gemeten kan worden, zullen meestal gesloten vragen of open vragen gebruikt worden. Een gesloten vraag is een vraagtype waarbij een persoon uit twee of meer alternatieven of antwoordmogelijkheden het goede antwoord moet kiezen. De open vraag is een vraagtype waarbij een leerling het antwoord zelf moet formuleren. Voor het meten van psychomotorische vaardigheden zoals autorijden, typen en timmeren, kan de motorische component niet met een schriftelijke toets gemeten worden. Bij deze zogenaamde 'performance tests' zal de opdracht of toetsvorm veelal gelijk zijn aan de situatie waarin het geleerd moet worden toegepast.

---

1 Van selectie is sprake als afhankelijk van de resultaten op een meetinstrument beslist wordt een leerling al of niet toe te laten tot een onderwijstraject.

2 Bij classificatie wordt afhankelijk van de resultaten op een meetinstrument beslist welk onderwijstraject een leerling moet volgen. De te onderscheiden trajecten zijn kwalitatief verschillend van aard, bijvoorbeeld wiskunde en natuurkunde. Voor beide vakken worden dan ook verschillende criteria gehanteerd om vast te stellen of een leerling succes heeft gehad in het desbetreffend vak.

- **Hoe worden de vragen of opdrachten *gescoord*?**
  Bij het scoren van vragen is een onderscheid te maken tussen dichotome en polytome scoring. Bij dichotome scoring wordt uitsluitend aan het goed antwoord een puntenaantal, meestal één scorepunt, toegekend. Bij polytome scoring wordt ook aan een antwoord dat gedeeltelijk goed is een puntenaantal toegekend.
  Bij de beoordeling van de antwoorden op open vragen en opdrachten wordt veelal gebruik gemaakt van een antwoordmodel dat de antwoorden en de bij de verschillende antwoorden behorende aantal scorepunten bevat.
  Een antwoordmodel is bedoeld om tot een objectieve beoordeling te komen, dat wil zeggen een beoordeling waarbij het aantal toegekende scorepunten onafhankelijk is van de persoon die beoordeelt.
- ***Hoeveel* items moeten geconstrueerd worden?**
  Het antwoord op deze vraag is afhankelijk van de vraag naar de dekking per leerstofonderdeel, naar het aantal toetsversies dat geconstrueerd dient te worden en naar het aantal items dat afvalt na een proeftoetsing.
- **Wat zijn de gewenste *psychometrische kenmerken* van de items en de toets?**
  Het antwoord op deze vraag is sterk afhankelijk van het doel van de toets. Aan toetsen die bedoeld zijn om de leerkracht te informeren over de voortgang van zijn leerlingen, zullen andere eisen gesteld worden dan aan toetsen die gebruikt worden voor selectiedoeleinden. Ook het niveau van de te toetsen groep speelt bij deze vraag een grote rol: is de toets bedoeld voor zwakke of voor goede leerlingen?

## 2.1 Leerdoelen formuleren

### Algemeen

Het vaststellen van de inhoud van de toets is de belangrijkste toetsspecificatie. Door het ontwikkelen van een toetsmatrijs ontstaat een overzicht van de gewenste samenstelling van de toets. Natuurlijk zijn randvoorwaarden als bijvoorbeeld beschikbare toetstijd, aantal te stellen vragen en samenstelling van de doelgroep mede van belang. De toetsmatrijs moet gezien worden als een blauwdruk; de matrijs geeft richting aan het constructieproces. De toets kan daardoor een afspiegeling zijn en tevens een operationalisatie worden van de beoogde leerdoelen. Deze leerdoelen dienen dan wel duidelijk en concreet geformuleerd te zijn.

### Formulering leerdoelen

Uit de leerdoelen moet duidelijk worden wat er onderwezen en getoetst dient te worden. De basis hiervoor zijn vaak de eindtermen, die gezien kunnen worden als een beschrijving van de kwaliteiten van leerlingen op het gebied van kennis, inzicht, vaardigheden en beroepshoudingen. Over het algemeen zijn deze globaal geformuleerd, waardoor deze vaak vertaald moeten worden naar doelstellingen die aangeven welk gedrag van leerlingen verwacht mag worden. Op deze wijze ontstaat ook informatie over de wijze waarop de studenten verwacht worden te leren, in de zin van hoe zij de doelen moeten bereiken. In wezen ontstaat hierdoor informatie over de te hanteren didactische werkvormen. In dit kader wellicht de belangrijkste doelstelling betreft de informatie die het leerdoel geeft over de richting waarop getoetst moet worden. Dat wil zeggen dat als de doelstelling het werkwoord 'opschrijven' bevat er schriftelijk getoetst dient te worden. Zijn doelstellingen op handelingsniveau geformuleerd dan is wellicht eerder te denken aan een praktische toetsing.

Voor een adequate toetsing dienen leerdoelen zo concreet mogelijk geformuleerd te zijn: ze moeten toetsbaar zijn. Dat wil zeggen dat uit de doelstellingen duidelijk moet worden welk gedrag de leerling moet vertonen.

Voor het formuleren van doelstellingen zijn in het algemeen vier criteria van belang:
1. gedrag
2. inhoud
3. voorwaarden
4. vorm

**Gedrag** kan gevat worden in werkwoorden waarmee aangegeven wordt wat we nastreven. Woorden als 'inzicht hebben in', 'weten dat' zijn niet toetsbaar. Het is beter gebruik te maken van werkwoorden als: opschrijven, herkennen, berekenen, formuleren, enz. Dit zijn toetsbare werkwoorden die gemeten kunnen worden.

De **inhoud** waarop de leerling zijn gedrag moet toepassen, dient nader gespecificeerd te worden: op welke context, in welke situatie. Voorbeeld: 'de leerling noemt drie situaties waar links inhalen verboden is'. Onder het kopje middelen of **voorwaarden** staat informatie over formules, boeken, tabellen en dergelijke hulpmiddelen die een leerling als gereedschap moet gebruiken bij het bereiken van het doel. Deze hulpmiddelen dienen vermeld te worden.

Met **norm** wordt de minimale prestatie of wel de norm bedoeld waaraan de prestatie als succesvol kan worden beschouwd. Voorbeeld: 'in twee van de drie gevallen', of 'binnen 10 minuten'.

## 2.2 Toetsmatrijs

### Wat is een toetsmatrijs?

Een toetsmatrijs is een tabel waarin aangegeven wordt hoe de opgaven, behorende bij bepaalde doelstellingen, worden verdeeld over tenminste twee dimensies: inhoudscategorieën en gedragscategorieën. De toetsmatrijs is een blauwdruk, een uitgewerkt plan, dat een systematische constructie van een toets wil garanderen. Een toets die is geconstrueerd op basis van een systematisch plan zal eerder bruikbare en betekenisvolle scores opleveren dan een toets waarvan de vragen op niet systematische wijze bij elkaar zijn gehaald.

**Enkele functies van een toetsmatrijs**

- U vermijdt dat teveel opdrachten worden gemaakt die gericht zijn op dezelfde leerstof dan wel op dezelfde vaardigheid. Wanneer de toetsmatrijs een juiste verhouding weergeeft van het aantal vragen in vergelijking met de verschillende onderdelen van het leer- of examenprogramma, is de kans dat de toets een representatieve steekproef vormt van de te toetsen doelstellingen groter.
- U kunt als u twee toetsen over dezelfde leerstof wilt maken de gelijkwaardigheid tussen die toetsen vergroten door ze beide op te stellen aan de hand van één toetsmatrijs .
- De toetsmatrijs kan dienen als een verantwoording van de inhoud van de toets naar anderen, zoals collega-vakdocenten en inspectie.

Een toetsmatrijs is een hulpmiddel om weer te geven op welke onderdelen van de leerstof en op welke gedragscategorieën de vragen uit één of meer toetsen gericht zijn. Een toets is immers méér dan een wat toevallige verzameling goed geformuleerde vragen. Het dient een betrouwbaar beeld te geven van datgene waarin de student onderwezen is. Een goede toets moet representatief zijn voor het betreffende gedeelte van het leerstofgebied. Nu is het natuurlijk niet mogelijk zoveel vragen te stellen dat de student van elk deel van de leerstof kan demonstreren wat hij ermee kan. Dat zou dagenlange zittingen vergen. Er wordt dus een keuze gemaakt: over onmisbare onderdelen worden altijd één of meer vragen gesteld en van de andere onderdelen worden sommige wél en andere niet getoetst. Het spreekt vanzelf dat deze keuze niet elke keer precies dezelfde moet zijn, omdat men dan het risico loopt dat het onderwijs zich zal beperken tot een vaste

keuze uit de leerstof. Ook mag de toets niet op te weinig onderdelen van de leerstof gericht zijn. Dan zou het een te eenzijdige steekproef worden. Bij de constructie van een toets is het nuttig van te voren een overzicht te maken van de gewenste samenstelling van de toets. Op die manier voorkomt men dat na het construeren van de afzonderlijke vragen blijkt dat er te veel op dezelfde vakinhoud of hetzelfde gedrag zijn gericht en dat andere onderdelen of gedragscategorieën minder aandacht zouden krijgen dan men zou wensen.
Welke van de onderwerpen het meest essentieel en relevant zijn, kan niet altijd eenvoudig bepaald worden. Een bruikbaar criterium is wel eens de tijd die in het onderwijs aan het betreffende onderwerp wordt besteed. Hoe meer tijd voor een onderwerp wordt uitgetrokken, hoe belangrijker het onderwerp waarschijnlijk is en hoe meer vragen over het onderwerp in de toets worden opgenomen.

### Inhouds- en gedragscategorieën

De toetsmatrijs geeft een bepaling van de inhoud van de toets in termen van het gedragsaspect en het inhoudelijk aspect van de te toetsen vaardigheid. De inhoudscategorieën zijn vanzelfsprekend voor ieder vak verschillend. Dat is niet het geval met de gedragscomponent. Bij de gedragscomponent gaat men er vanuit dat algemene psychologische processen ten grondslag liggen aan de vaardigheden die in doelstellingen van de verschillende vakken genoemd worden. Een veel gebruikte indeling van gedragsaspecten van het cognitieve domein is de hiërarchische ordening van hoofdvaardigheden van Bloom c.s. (1956):

1. Kennis
2. Begrip
3. Toepassing
4. Analyse
5. Synthese
6. Beoordeling

Meestal wordt deze zesdeling, terwille van de hanteerbaarheid, terug gebracht tot een tweedeling: kennis en toepassing, of te wel: reproductie en productie. De verschillen zijn dan dat bij kennis (reproductie) de nadruk ligt op het toetsen van zaken die als zodanig geleerd en/of onderwezen zijn. Bij het toetsen van toepassing (productie) wordt van de leerling gevraagd zijn kennis toe te passen in een andere context dan hem is onderwezen.
Naast de taxonomie van Bloom zijn nog andere taxonomieën bekend.
Ook het inhoudelijke aspect (de leerstof) kan in subcategorieën worden verdeeld (m.b.v. docenten en/of andere deskundigen uit het betreffende vakgebied, onderwijstypen, vervolgonderwijs en mogelijk uit het bedrijfsleven). Deze subcategorieën concentreren zich rond bepaalde onderwerpen uit het totale leerstofgebied. Op deze wijze verkrijgt men een redelijke afbakening van de vakinhoudelijke kant van de doelstellingen.

### Voorbeelden van een toetsmatrijs

De toetsmatrijs vormt nu de grondslag voor de toetsconstructie. Hij geeft richting aan het constructieproces, in die zin dat de toets nu een afspiegeling en tevens een operationalisatie kan worden van de beoogde doelstellingen. De percentages van het totaal aantal items in de toets per kolom (gedragsaspect), per rij (inhoudsaspect) en ten slotte per cel (de items als operationalisatie van de concrete doelstellingen) geven als het ware gewichten aan die de belangrijkheid van de te toetsen doelstellingen weerspiegelen. Overigens hoeven niet alle cellen gevuld te worden. Onderstaande tabel geeft aan hoe in algemene zin een toetsmatrijs eruit ziet.

| LEERSTOF | GEDRAG | | Totaal per leerstofonderdeel |
|---|---|---|---|
| | Reproductie | Productie | |
| **A** | | | 4 |
| **B** | | | 3 |
| **C** | | | 4 |
| **D** | | | 1 |
| **Totaal** | 4 | 8 | In totaal 12 vragen of een veelvoud ervan |

Uit de matrijs is af te lezen dat in de toets over leerstof A vier keer zoveel vragen opgenomen (dienen) te worden dan over leerstof D. Die verhouding (4:1) kan veroorzaakt worden door het examenprogramma. Maar het kan ook zijn dat een toetsconstructeur het ene onderdeel belangrijker vindt dan het andere en op basis daarvan de verhouding bepaalt. In het voorgaande voorbeeld zijn als gedragsaspecten reproductie en productie gehanteerd, een in de praktijk veel voorkomend onderscheid. Ook andere indelingen van gedragscategorieën zijn mogelijk, zoals de volgende tabel laat zien.

**Toetsmatrijs module 1 biologie**

| Onderwerp: | Weten | Inzien | Toepassen | Integreren | Weging |
|---|---|---|---|---|---|
| Bouw van de cel | 4 (14%) | 2 (7%) | | | 6 (20%) |
| Fotosynthese | 4 (13%) | 2 (7%) | 2 (7%) | | 8 (27%) |
| Verbranding | 4 (13%) | 2 (7%) | 2 (7%) | | 8 (27%) |
| Onderscheid plant/dieren | 2 (6%) | 1 (3%) | | | 3 (10%) |
| Practicum biologisch onderzoek | | | 3 (10%) | 2 (7%) | 5 (17%) |
| **Totaal (aantal vragen en %)** | 14 (46%) | 7 (23%) | 7 (23%) | 2 (7%) | 30 (100%) |

# 3. Itemconstructie

***Als er duidelijkheid is over de functie van de toets en de specificaties waaraan de te construeren toets moet voldoen, dan is de volgende stap het construeren van items.***

Een toets is een verzameling items. Deze verzameling is echter niet een willekeurige. De in de toets op te nemen items moeten aansluiten bij de doelstelling van de toets. De te kiezen toetsvorm is nauw verbonden met de geformuleerde eindtermen of kerndoelen. Zo leggen de kerndoelen voor de basisvorming meer nadruk op (toegepaste) kennis in contexten en vaardigheden. De te ontwikkelen toetsen dienen daarbij aan te sluiten.
Bij examens in de tweede fase van het voortgezet onderwijs zien we een scala aan toetsvormen, variërend van de bekende toetsen met open en gesloten vragen tot praktische opdrachten, verwerkingsopdrachten en het profielwerkstuk. Elke toetsvorm heeft zijn voor- en nadelen. Niet iedere toetsvorm is geschikt voor elk type doelstelling, voor elk type leerstof of voor elk type leerlingen. Het is zaak de meest efficiënte toetsvorm te kiezen die het beste past bij de doelen van het onderwijs.

## 3.1 Vraagvormen

De meest 'vertrouwde' toetsvormen zijn de open en gesloten vragen. Maar ook andere toetsvormen zoals praktische opdrachten, beroepspraktijk vorming (BPV), stageverslagen, performance tests, en scripties zijn niet meer weg te denken. Vooralsnog beperken we ons tot de open en gesloten vragen. Aan elke vraagvorm zijn voor- en nadelen te onderscheiden (ontleend aan: Het maken van toetsen bij methoden : handreikingen voor auteurs basisvorming.  Arnhem: Citogroep, 1999):

| | Voordelen | Nadelen |
|---|---|---|
| **Gesloten vragen** | ▪ De correctie kost weinig tijd; antwoorden op geprecodeerde vragen kan de docent zelf nakijken met een mal.<br>▪ De beoordeling is volledig objectief. Afgezien van slordigheidsfouten, zal elke docent tot dezelfde score komen.<br>▪ De formuleervaardigheid van de leerling speelt geen rol.<br>▪ Er is veel kennis en inzicht afvraagbaar in korte tijd. | ▪ Niet alle soorten doelen en inhouden kunnen met gesloten vragen worden getoetst.<br>▪ Gesloten vragen doen vaak een sterker beroep op leesvaardigheid en kennis van de wereld dan open vragen; deze 'taalbarrière' kan bijvoorbeeld leerlingen met een anderstalige achtergrond benadelen.<br>▪ De leerlingen kunnen 'gokgdrag' vertonen als zij naar het antwoord raden. De raadkans wordt groter naarmate de vraag minder alternatieven kent. Er moeten dan veel vragen gemaakt worden en dat is kostbaar en tijdrovend.<br>▪ De constructie is wat moeilijker en tijdrovender, zeker als men niet alleen feitenkennis wil toetsen, maar ook inzicht en vaardigheden. |
| **Open vragen** | ▪ Open vragen zijn meer geschikt als er veel goede antwoorden mogelijk zijn, als men wil nagaan in hoeverre de leerling iets kan uitleggen of samenvatten en als men de redactionele vaardigheden van de leerling wil toetsen.<br>▪ De leerling heeft grotere vrijheid in het beantwoorden van de vraag.<br>▪ De toetsontwikkelaar kan aan de antwoorden van de leerlingen zien of de vraag al dan niet duidelijk gesteld is en de vraag desgewenst herformuleren.<br>▪ De antwoorden op open vragen geven de docent meer informatie over de mate waarin de leerlingen de stof beheersen en waar zich de problemen precies voordoen.<br>▪ De kans dat de leerling het goede antwoord raadt, is (bijkans) nul. | ▪ De beantwoording van een open vraag kost relatief veel afnametijd. Daardoor kan men in dezelfde tijd minder vragen stellen. Dat kan ten koste gaan van de betrouwbaarheid van de toets.<br>▪ Open vragen zijn minder objectief. Ze moeten door de docent worden nagekeken. Hierbij kunnen allerlei storende beoordelaarseffecten optreden.<br>▪ Het is de kunst open vragen zo te formuleren dat het voor alle leerlingen duidelijk is wat voor antwoord van hen verlangd wordt. Is dat niet volledig gelukt, dan zijn leerlingen die de vraag niet helemaal begrepen hebben in het nadeel.<br>▪ Het corrigeren kost de docent doorgaans meer tijd en moeite. Dit geldt uiteraard meer voor lang-antwoordvragen dan voor kort-antwoordvragen. |

## 3.2 Scoring en/of beoordeling

### Gesloten vragen

De beoordeling van een gesloten vraag, ook wel meerkeuzevraag genoemd, is objectief. Alle goed te rekenen antwoorden zijn van tevoren vastgelegd. Gesloten vragen worden gescoord aan hand van een correctiesleutel: een lijst van de goed te rekenen antwoorden.

### Open vragen

Bij het nakijken van open vragen, en bijvoorbeeld ook praktijktoetsen/praktische opdrachten, speelt de subjectiviteit van de beoordelaar altijd een rol. Deze subjectiviteit dient zoveel mogelijk gereduceerd te worden. Dit kan met behulp van een:

1. ***Correctievoorschrift***
   Dit is een lijst met richtlijnen voor de docent en bestaat doorgaans uit drie onderdelen:
   - antwoordmodel
   - scoringsvoorschrift
   - beoordelaarsinstructie

2. ***Beoordelingsschema***

   Dit is een instructie, soms in de vorm van een schema, dat dient als richtlijn bij de beoordeling van uitwerkingen bij opgaven waarbij geen eenduidig antwoordmodel op te stellen is. Dit kan zich voordoen bij vrije opdrachten en bij opstelvragen. In het beoordelingsmodel staan de criteria aan de hand waarvan de docent de uitwerking of het antwoord beoordeelt.

   **Voorbeeld beoordelingsschema**

   *Een antwoord moet worden goed gerekend, wanneer:*

   1. de student zowel met betrekking tot de eerste als met betrekking tot de tweede vraag nagaat welke consequenties een (positief) antwoord op die vraag volgens hem/haar heeft of zou kunnen hebben voor de kijk van de mens op zichzelf en op diens relatie tot de wereld om hem heen;
   2. de student uiteenzet welke van de bij (1) bedoelde consequenties voor hem/haar het zwaarste wegen, het meest ingrijpend zijn;
   3. de student op basis van (2) de conclusie trekt dat hij/zij het met de auteur eens is, respectievelijk het niet met de auteur eens is.

3. ***Beoordelingsschaal***

   Bij een beoordelingsschaal kunnen de docent en/of de leerlingen op een glijdende schaal aangeven in welke mate kennis, vaardigheden of houdingen bij een leerling aanwezig zijn. De glijdende schaal bestaat uit meerdere punten die een bepaalde positie of rangorde aangeven, bijvoorbeeld lopende van 'zwak' naar 'uitstekend' of van 'zelden of nooit' naar 'vrijwel altijd'. Bij een beoordelingsschaal moet de toetsontwikkelaar bepalen welke aspecten van de prestatie beoordeeld worden.

   **Voorbeeld van een schriftelijke (zelf)beoordeling**

| | | | | | | |
|---|---|---|---|---|---|---|
| - In hoeverre is aan de eisen voldaan | 0 | 1 | 2 | 3 | 4 | 5 |
| - Herkenbaarheid van de voorstelling | 0 | 1 | 2 | 3 | 4 | 5 |
| - Compositie | 0 | 1 | 2 | 3 | 4 | 5 |
| - Vaardigheid in het materiaalgebruik | 0 | 1 | 2 | 3 | 4 | 5 |
| - Presentatie | 0 | 1 | 2 | 3 | 4 | 5 |
| - Eigen beoordeling | 0 | 1 | 2 | 3 | 4 | 5 |

## 3.3 Construeren van gesloten vragen

**Richtlijnen voor het construeren van gesloten vragen (ontleend aan: Wijnen en Alberts, 1993):**

- Toets in principe in één vraag één leerdoel.
- De vraag moet niet anders meten dan het bij de vraag behorende leerdoel.
- De alternatieven moeten helder en eenduidig zijn.
- De afleiders moeten geloofwaardig zijn voor degen die de stof niet goed hebben bestudeerd.
- De antwoordalternatieven mogen niet te veel verschillen in woordgebruik, omvang, lengte, e.d.
- Geef de bedoeling van de vraag expliciet in de stam aan.
- Natuurlijk mag er slechts één alternatief juist zijn.
- Stel altijd eerst het juiste alternatief op en daarna pas de afleiders.
- Formuleer zowel de stam als de alternatieven zo kort mogelijk.
- Verberg het leerdoel dat getoetst wordt niet onder een hoeveelheid irrelevante informatie in de stam.
- Figuren moeten functioneel zijn; als ze niet tot het verhelderen van het probleem bijdragen, laat ze dan weg.
- Vermijd, indien mogelijk, een ontkennende zin in de stam.
- De alternatieven moeten grammaticaal in overeenstemming zijn met elkaar en met de stam.

- Wanneer de leerling uitspraken op hun juistheid moet beoordelen, moeten ze ondubbelzinnig juist of onjuist zijn.

**Verder lezen**

- Een handzame brochure, getiteld Toetsen met gesloten vragen, over het construeren van gesloten vragen en voorzien van tips, kunt u als pdf-file downloaden of als document inzien.

**Richtlijnen voor het construeren van open vragen (ontleend aan: Wijnen en Alberts, 1993):**

- Zorg ervoor dat een opgave te herleiden is tot een omschreven leerdoel en zorg ervoor dat het onderwerp van de vraag duidelijk is afgebakend.
- Wees duidelijk in wat de leerling precies moet doen: formuleer de opgave zo zakelijk mogelijk, maar geef wel voldoende informatie voor een correcte beantwoording.
- Zorg ervoor dat het gesteld probleem oplosbaar is met het in de doelstelling beoogde vaardigheidsniveau.
- Formuleer de vraag taalkundig correct en stem taalgebruik en stijl af op het niveau van het lesmateriaal en de leerlingen. Geef de leerling voldoende informatie over de lengte van het antwoord, de gewenste vorm van het antwoord en de elementen die het antwoord moet bevatten.
- Maak een correctievoorschrift bestaande uit een antwoordmodel, een scoringsvoorschrift en eventueel een beoordelaarsinstructie.
- Geef in het antwoordmodel een opsomming van goede, gedeeltelijk goed en indien nodig ook onjuiste antwoorden. En wees zo duidelijk mogelijk over wat er in de antwoorden van leerlingen als 'niet juist' of 'niet geheel juist' beoordeeld moet worden. Maak indien mogelijk voor elke vraag een antwoordmodel. En neem geen antwoorden op die leerlingen niet of nauwelijks zullen geven.
- Vermeld in het scoringsvoorschrift de maximaal haalbare toetsscore, het aantal scorepunten per vraag en de scoring van goede, gedeeltelijk goede en geheel goede antwoorden. Omschrijf indien nodig wanneer bonuspunten worden toegekend c.q. aftrekpunten in mindering worden gebracht.
- Ken aan elk antwoordelement één scorepunt toe (en geen twee of meer). Als een goed antwoordelement namelijk twee punten 'krijgt', bestaat het gevaar dat docenten ten onrechte één punt toekennen als het antwoord in hun ogen gedeeltelijk goed is.
- Maak het correctievoorschrift niet te algemeen zodat van een uniforme beoordeling weinig terechtkomt. Maar maak het voorschrift ook weer niet te gedetailleerd in de zin dat het voor de beoordelende docent door de omvang moeilijk hanteerbaar is.
- Geef het correctievoorschrift zodanig vorm dat de beoordelende docent snel inzicht krijgt in de beoordelingstaak.
- Indien één of meer vragen niet in het antwoordmodel ondergebracht kunnen worden, maak dan gebruik van beoordelingsschema's of beoordelingsschalen.

**Verder lezen:**

- Voor een handzame brochure over het construeren van open vragen, voorzien van tips, wordt verwezen naar Toetsen met open vragen. Een handleiding voor het construeren van toetsen met open vragen, een uitgave van Cito uit 1991/1992.
- Verder noemen we Toetsen met open vragen, geschreven door Tom Erkens van Cito. Het betreft een hoofdstuk uit het boek 'Toetsen in het hoger onderwijs' door Henk van Berkel en Anneke Bax (2002), een uitgave van Bohn Stafleu Van Loghum (ISBN 9031336394).

## 3.5 Richtlijnen voor het ontwikkelen van onpartijdige toetsen

In Nederland is onderzoek uitgevoerd naar de vraag of de geringe toetsprestaties van allochtone en vrouwelijke leerlingen wellicht het gevolg zijn van onbedoelde kenmerken van de opgaven. Uit dat onderzoek bleek dat sommige opgaven moeilijker waren voor allochtone leerlingen dan voor autochtone leerlingen als gevolg van hun culturele en linguïstische achtergrond. Ook bleek dat sommige opgaven moeilijker waren 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 leerlingen veroorzaakt wordt door aspecten van de opgave die niet relevant zijn voor dat wat de opgave beoogt te meten, wordt in de literatuur 'itembias' genoemd.

De Richtlijnen hebben in de eerste plaats tot doel er voor te zorgen dat de toetsen die door Cito ontwikkeld worden geen opgaven met itembias bevatten. Daarnaast beogen de Richtlijnen te voorkomen dat de toetsen opgaven met kwetsende inhoud bevatten. De Richtijnen kunnen zowel gebruikt worden bij de constructie van nieuwe toetsen als voor beoordeling van bestaande toetsen.
Bekijk de informatie en download de Richtlijnen voor het ontwikkelen van onpartijdige toetsen, geschreven door K. Bügel en P.F. Sanders van Cito in Arnhem (1998).

## 3.6 Eenvoudig taalgebruik in toetsen

Het is van belang tijdens de toetsconstructie erop te letten dat de leerlingen bij het maken van een toets of een examen geen fouten zullen maken als gevolg van een voor hen te moeilijk of te vaag taalgebruik. Kwalitatief verantwoorde toetsen en examens vereisen het voorkomen van een onbedoelde benadeling van allochtone en minder taalvaardige leerlingen.

# 4. Toetsafname

***Bij toetsafname dient een onderscheid gemaakt te worden tussen een try-out of proefafname en de definitieve toetsafname. Een proefafname is bedoeld om een indruk te krijgen van hoe de items inhoudelijk en psychometrisch functioneren bij de leerlingen waarvoor de definitieve toets bedoeld is. Op basis van de resultaten van de proefafname zullen sommige items verwijderd of gereviseerd worden. Zo mogelijk vindt na de revisie een nieuwe proefafname plaats. In de praktijk gebeurt dit laatste niet altijd. Het aantal leerlingen waaraan de toets wordt voorgelegd, is bij de proefafname kleiner dan bij een definitieve afname.***

Het is van belang dat de definitieve afname onder gestandaardiseerde condities wordt afgenomen. Standaardisatie houdt in dat de toets door alle leerlingen onder gelijke omstandigheden wordt uitgevoerd. Alleen dan is het mogelijk de toetsprestaties van leerlingen met elkaar te vergelijken.

## 4.1 Proefafname

Hoe vaak items ook besproken en bijgesteld zijn, het blijft moeilijk de reacties van leerlingen op de items te voorspellen. Vandaar dat de items worden gepretest. Pretesten is een noodzakelijke stap om defecte items te detecteren en na te gaan of de moeilijkheidsgraad van de items geschikt is voor of aansluit bij de leerlingen waarvoor de uiteindelijke toets bedoeld is. Screeners van concept-items voorzien toetsconstructeurs van kwalitatieve informatie. Pretestresultaten voorzien toetsconstructeurs van kwantitatieve en statistische informatie, waarbij er altijd rekening mee gehouden moet worden dat de resultaten uit een pretest geen absolute indicatoren zijn over de kwaliteit van de items. Ze geven alleen informatie over de groep die deelnam aan de pretest. Indien deze groep groot genoeg is, dan voorspellen ze nauwkeurig genoeg hoe de items zich in de totale populatie zullen gedragen.

## 4.2 Try-out

Een try-out is een situatie waarin toetsen of bepaalde opgaven aan een oppervlakkig vooronderzoek worden onderworpen. Bij een try-out gelden minder strenge onderzoeksvoorwaarden dan bij pretesten. Vaak wordt een try-out gehouden met een vrij klein aantal proefpersonen (steekproef) en worden aan de representativiteit van de steekproef geen hoge eisen gesteld. De bij een try-out verkregen gegevens bevatten alleen aanwijzingen voor het verbeteren van de opgaven, maar kunnen niet worden gebruikt voor het doen van uitspraken over de groep personen waarvoor de opgaven zijn bedoeld.

Een voorbeeld waarbij een try-out prima tot zijn recht komt, is het nagaan of een antwoordmodel bij een open vraag aanpassing behoeft. Door een beperkt aantal leerlingwerken voor te leggen aan een aantal beoordelaars, kan op relatief eenvoudige wijze nagegaan worden of het antwoordmodel door alle beoordelaars op dezelfde wijze gehanteerd (geïnterpreteerd) wordt, of dat aanpassing gewenst is.

Aan de hand van de leerlingantwoorden kan men nagaan of de vraagstelling eenduidig is of aanpassing behoeft. Als vuistregel geldt dat het beoordelen van 25-50 leerlingwerken door vijf à tien leerkrachten voldoende informatie oplevert om te komen tot gewenste bijstellingen.

In onderstaand kader geven we een voorbeeld waaruit het belang van een try-out naar voren komt.

**Voorbeeld resultaten try-out naar beoordelaarsovereenstemming**

| Kandidaten | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 |
| 2 | 2 | 2 | 2 | 2 |
| 3 | 2 | 3 | 1 | 2 |
| 4 | 3 | 4 | 3 | 3 |
| 5 | 3 | 4 | 4 | 4 |
| 6 | 4 | 4 | 3 | 4 |
| 7 | 2 | 3 | 2 | 1 |
| **Totaal** | **16** | **21** | **15** | **16** |

Uit de totaalscores blijkt dat beoordelaar 2 op basis van het antwoordmodel de hoogste scores geeft. Belangrijk is de vraag op basis  waarvan. Met name omdat de andere beoordelaars maximaal 1 punt van elkaar afwijken. Interpreteert beoordelaar 2 het antwoordmodel anders? Behoeft het correctievoorschrift aanpassing?

Als een vraag verschillend beoordeeld wordt door beoordelaars dan kan het nodig zijn de vraag te herformuleren. Soms echter volstaat het aanpassen van het antwoordmodel.

# 5. Itemevaluatie

***Twee belangrijke statistische indices voor het beoordelen van de kwaliteit van items, zijn:***

- ***moeilijkheidsgraad (p-waarde)***
- ***discriminatie-index ($r_{it}$ of $r_{ir}$)***

## 5.1 Moeilijkheidsgraad

De p-waarde van een meerkeuzevraag is de proportie kandidaten die het goede antwoord heeft gekozen. Met dit getal wordt de moeilijkheidsgraad van een item weergegeven. Voor open vragen gebruikt men de p'-waarde: de p'-waarde is een getal tussen 0 en 1 dat de moeilijkheidsgraad van een item weergeeft. De p'-waarde wordt berekend door de gemiddelde score op een opgave te delen door de maximaal haalbare score op die opgave. Een opgave met een p'-waarde van .10 is erg moeilijk; een opgave met een p'-waarde van .90 is erg gemakkelijk.

Opgaven met te hoge of te lage p-waarden dienen vermeden te worden. Deze opgaven zijn weinig informatief. Extreem gemakkelijke opgaven maken geen onderscheid tussen goede en minder goede leerlingen. Bovendien zetten te gemakkelijke opgaven de betere leerling soms op het verkeerde been. Zij denken dat er een addertje onder het gras zit. In de regel streeft men naar p-waarden tussen .27 en .79 (zie kader).

**Normen voor p- en p'-waarden**

| Aantal alternatieven | Optimale p waarde (p=0.5+0.5/m) | optimale p-waarde (Lord) |
|---|---|---|
| 2 | 0.75 | 0.85 |
| 3 | 0.67 | 0.77 |
| 4 | 0.63 | 0.74 |
| 5 | 0.60 | 0.70 |

**Toelichting**
In de literatuur vinden we verschillende opvattingen over de optimale p-waarde van een item. Crocker en Algina (1986) stellen dat de optimale p-waarde halverwege de raadkans en 1.0 moet liggen. De veronderstelling hierbij is dat er geraden wordt als men niet weet wat het goede antwoord op een meerkeuze-item is. In formulevorm uitgedrukt: p = 0.5 + 0.5/m, waarin m het aantal alternatieven is en p de gewenste p-waarde. Naar aanleiding van een simulatie-onderzoek komt Lord (1952) tot een andere conclusie. De conclusie van een onderzoek van Feldt (1993) is, dat de optimale p-waarde tussen 0.57 en 0.67 moet liggen wanneer er geraden kan worden. Indien er geen reden is om aan te nemen dat er geraden wordt, of als er niet geraden kan worden zoals bij open vragen, is de optimale p-waarde gelijk aan 0.50. Het effect van de moeilijkheid van een item op de betrouwbaarheid blijkt echter verbazingwekkend klein te zijn, zelfs als de p-waarden variëren van 0.27 tot 0.79.

Behalve naar het percentage correcte antwoorden, kan ook gekeken worden naar het percentage kandidaten dat een afleider bij een meerkeuzevraag kiest of slechts een deel van de maximale score krijgt. Samen met de p-waarde is deze informatie belangrijk bij het beoordelen van de kwaliteit van een item.

## Voorbeeld van het gebruik van p-waarden bij het beoordelen van de kwaliteit van open vragen

| Opgave | maximum | gemiddelde score | p'-waarde | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 1.50 | 0.75 | 10 | 30 | 60 | | |
| 2 | 2 | 0.50 | 0.25 | 60 | 30 | 10 | | |
| 3 | 2 | 1.00 | 0.50 | 10 | 80 | 10 | | |
| 4 | 4 | 3.00 | 0.75 | 30 | 0 | 0 | 0 | 70 |

**Toelichting**
De tabel laat zien dat opgave 1 een maximum score heeft van 2 punten. De gemiddelde score van deze opgave is 1,50 en de p'-waarde 0.75 (de gemiddelde score is 75% van de maximale score). Uit de rechterkant van de tabel blijkt dat 10% van de kandidaten op deze opgave een score 0 had, 30% een score 1 en 60% de maximale score. De opgave was niet moeilijk, hoewel toch nog een aantal kandidaten een gedeeltelijk goed of een fout antwoord gaf.
Opgave 2 daarentegen was duidelijk moeilijker. De meeste kandidaten (60%) behaalden geen scorepunten. Of omdat ze helemaal geen antwoord gaven of een verkeerd antwoord. 30% kreeg slechts één punt. Slechts 10% van de kandidaten kreeg het maximaal aantal punten. Merk op dat in beide besproken vragen een substantieel deel van de kandidaten een deel van de maximum score kreeg. Blijkbaar werkte de differentiatie in het antwoordmodel prima.
Opgave 3 was redelijk moeilijk. Opvallend is de verdeling van de scorepunten. Slechts 10% van de leerlingen behaalde de maximale score. Daarentegen behaalde 80% één scorepunt. Blijkbaar was het niet realistisch een volledig goed antwoord te verwachten of was de vraagstelling niet geheel helder. Een mogelijke oplossing is de vraag opnieuw te formuleren en duidelijker te maken wat precies bedoeld wordt. Ook het aanpassen van het antwoordmodel behoort tot de mogelijkheden, waarbij eventueel gebruik gemaakt kan worden van de leerlingantwoorden.
Bij opgave 4 zijn de deelscores niet gebruikt. Wellicht dat het beter is om het antwoordmodel aan te passen aan twee antwoordmogelijkheden: goed/fout.

## Voorbeeld van het gebruik van p-waarden bij het beoordelen van de kwaliteit van meerkeuzevragen

| Opgave | maximum-score | gemiddelde score | p-waarde | O/D | Frequentie alternatieven in % | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | A | B | C | D |
| 11 | 1 | 0.75 | 0.75 | 0 | 7 | 10 | 8 | 75 |
| 12 | 1 | 0.25 | 0.26 | 10 | 23 | 27 | 26* | 24 |
| 13 | 1 | 0.27 | 0.27 | 0 | 27* | 0 | 45 | 28 |
| 14 | 1 | 0.11 | 0.11 | 6 | 71 | 11* | 10 | 8 |

**Toelichting**
Bovenstaande tabel laat zien dat opgave 11 een maximum score heeft van 1 punt, een gemiddelde score van 0,75 punten en een p-waarde van 0.75. De rechterkant van de tabel geeft het percentage leerlingen aan dat voor een bepaald alternatief gekozen heeft. Zo koos 7% van de leerlingen voor alternatief A, 10 % voor B, 8 % voor C en 75% voor D. De asterisk geeft aan dat D het goede antwoord was. De indices laten zien dat opgave 11 niet moeilijk was. Bovendien blijkt dat alle alternatieven door een kleine groep leerlingen zijn gekozen, hetgeen betekent dat alle alternatieven plausibel waren.
De p-waarde van opgave 12 is laag (0.27). Alle afleiders hebben dezelfde waarde, hetgeen zou kunnen wijzen op gokgedrag. Deze opvatting wordt versterkt door de kolom O/D, die aangeeft dat 10% van de kandidaten meer dan één alternatief aankruisten of de opgave niet gemaakt hebben. Het item zou te moeilijk geweest kunnen zijn of voor leerlingen verwarrend.
Ook opgave 13 is opvallend. Meer leerlingen kiezen voor het foute antwoord C dan het goede antwoord A. De moeite waard om eens naar de formulering van de alternatieven te kijken. Alternatief B is door niemand gekozen. Wellicht is het het verstandigste dit alternatief weg te laten. Met deze opgave is dus duidelijk iets aan de hand. Men zou ervoor kunnen kiezen de opgave te herformuleren, maar aangezien er meerdere opmerkingen bij geplaatst kunnen worden, is het de overweging waard de opgave weg te doen.
Opgave 14 toont een beeld dat mogelijk gemakkelijk te veranderen is. Opvallend is dat B het goede antwoord is, terwijl afleider A de hoogste frequentie p-waarde heeft. In dit geval is het verstandig na te gaan of er sprake is van een sleutelfout. Een opgave doorloopt een aantal stadia, waarbij vaak sprake is van herordening van alternatieven. Mogelijk dat daarbij een vergissing begaan is. Het blijft natuurlijk altijd mogelijk dat afleider A voor leerlingen attractiever is dan B. Een aspect dat in de beoordeling van het item meegenomen moet worden.

## 5.2 Discriminatie-index

De discriminatie-index (Rit[3]) geeft aan in hoeverre een item onderscheid maakt tussen personen met hoge toetsscores en personen met lage toetsscores

Een hoge Rit betekent dat veel personen met een hoge toetsscore het item goed hebben beantwoord en veel personen met een lage toetsscore het item fout hebben beantwoord. Een hoge Rit betekent ook dat het item relatief veel bijdraagt aan de betrouwbaarheid van de toets.

Voor Rit-waarden vindt men in de literatuur geen absolute normen. De waarden kunnen variëren tussen -1 en +1. Een Rit-waarde van .50 en hoger is echter in de praktijk bij toetsen met meer dan veertig items al erg hoog. Onderstaande tabel geeft een indicatie van geaccepteerde normen.

**Normen voor Rit-waarden**

| Rit-waarde | beoordeling |
|---|---|
| 0.19 en lager | slecht |
| 0.20 – 0.29 | twijfelachtig |
| 0.30 – 0.39 | goed |
| 0.40 en hoger | zeer goed |

De grootte van de Rit is onder andere afhankelijk van het aantal items in een toets. Daarom moet men strikt genomen bovenstaande normen alleen hanteren bij Rit-waarden die gecorrigeerd zijn voor toetslengte. Vanwege het geringe effect kan de correctie achterwege blijven indien de items afkomstig zijn uit toetsen met veertig of meer items.

Naast de Rit is de Rir een veel gebruikte discriminatie-index. De Rir is een soortelijke index als de Rit. Gaat het bij de Rit om de correlatie tussen itemscores en toetsscores, bij de Rit gaat het om de correlatie tussen itemscores en restscores. De restscore van een persoon is gelijk aan zijn toetsscore minus de score op de desbetreffende items. Een persoon heeft dus evenzoveel restscores als er items zijn in de toets.

Zowel aan de Rit als aan de Rir kleven bezwaren. De Rit geeft een geflatteerd beeld van de samenhang tussen de score op een item en de toetsscore, omdat de itemscore onderdeel is van de toetsscore. We correleren dus het item voor een deel met zichzelf. De Rir ondervangt dit bezwaar, maar heeft als bezwaar dat de restscore waarmee een item gecorreleerd wordt met het item varieert. De Rir-waarden van eenzelfde toets zijn daardoor onderling niet te vergelijken. Als echter het aantal items in een toets veertig of meer is, zijn beide bezwaren van geen belang meer.

Bij een toets met meerkeuzevragen is het mogelijk, naast een discriminatie-index voor het goede antwoord discriminatie-indices voor de afleiders (foute antwoorden) te berekenen. Per item zijn er uiteraard evenveel Rar-waarden als er afleiders zijn. De Rar-waarde wordt berekend door personen die het desbetreffende foute antwoord hebben gekozen een itemscore 1 en de anderen een score 0 te geven. Vervolgens wordt de correlatie tussen het foute antwoord en de restscore berekend, waarbij de restscore per definitie dezelfde waarde heeft als bij de berekening van de Rir. Omdat we toetsen met een hoge betrouwbaarheid nastreven, zijn items met positieve Rir- en negatieve Rar-waarden gewenst. Zulke waarden impliceren dat relatief veel personen met een hoge toetsscore het item goed hebben beantwoord en relatief veel personen met een lage toetsscore het item fout hebben beantwoord. Een positieve Rar geeft aan dat

[3] De Rit is de product-moment-correlatie tussen de itemscore en de toetsscore. Deze correlatie wordt bij dichotoom gescoorde items wel puntbiseriële correlatie genoemd: het is de correlatie tussen een dichotome (goed of fout) en een continu geachte variabele. Een product-moment-correlatie neemt de waarden aan tussen +1 en -1. Een correlatie van +1 betekent dat er een perfect positief lineair verband bestaat tussen twee variabelen. In dit geval de itemscore en de toetsscore.

relatief veel goede personen de desbetreffende afleider als het goede antwoord hebben aangemerkt. Soms kan dit een sleutelfout zijn: de verkeerde sleutel is per ongeluk opgegeven of bij nader inzien blijkt dat de afleider met de positieve Rar het goede antwoord is.

De belangrijkste regels voor de interpretatie van Rir- en Rar-waarden zijn:

- de Rar-waarde voor een afleider moet negatief zijn;
- de combinatie van een Rar-waarde tussen .10 en 0 en een Rir-waarde tussen .10 en .20 voor het juiste antwoord (de sleutel) is verdacht;
- ga nooit alleen af op verdachte waarden, maar betrek het item altijd in de beoordeling;
- besteed geen aandacht aan waarden voor afleiders die door minder dan 5% van de kandidaten zijn gekozen.

# 6. Toetssamenstelling

***Voor het kunnen selecteren van vragen is het nodig dat zowel kwalitatieve kenmerken (bijvoorbeeld leerstofcategorieën) als kwantitatieve kenmerken (bijvoorbeeld moeilijkheidsgraad) van de items bekend zijn. De mogelijkheden voor selectie worden uiteraard bepaald door de omvang van de verzameling items. Wanneer de verzameling uit een groot aantal items bestaat die van kwalitatieve en kwantitatieve kenmerken voorzien zijn, spreekt men van een itembank[4].***

Aan bod komen drie soorten eisen die van belang zijn bij het samenstellen van toetsen.

## 6.1 Richtlijnen

Bij het samenstellen van toetsen zijn drie soorten eisen van belang:
1. psychometrische eisen;
2. inhoudelijke eisen;
3. praktische eisen.

- De **psychometrische eisen** zullen veelal betrekking hebben op de gewenste meetnauwkeurigheid van de samen te stellen toetsen.

- Met **inhoudelijke eisen** worden de vakinhoudelijke en onderwijskundige eisen bedoeld: de verdeling van de vragen over de leerstofcategorieën, de gewenste moeilijkheidsgraad van de toets en dergelijke. Ook relaties op itemniveau kunnen een rol spelen bij het samenstellen van toetsen. Als bijvoorbeeld het antwoord op item 4 een aanwijzing bevat voor de antwoorden op item 12, dan kan de toetsconstructeur eisen dat als item 4 in de toets wordt opgenomen, item 12 niet meer wordt opgenomen.

- Onder **praktische eisen** worden die aspecten van toetsconstructie verstaan die psychometrische noch inhoudelijke betekenis hebben, maar bij het samenstellen van toetsen wel degelijk een rol spelen. Een voorbeeld is de tijd die voor het afnemen van een toets beschikbaar is. Aangezien de tijd niet onbeperkt is, zal men hier bij het samenstellen van een toets rekening mee moeten houden. Een ander voorbeeld betreft het budget dat beschikbaar is om een toets te kunnen afnemen. Een bepaald budget kan betekenen dat niet meer dan drie beoordelaars ingeschakeld kunnen worden.

Bij het samenstellen van toetsen houdt men rekening met de specificaties waaraan de toets moet voldoen. Een belangrijke informatiebron daarbij is de toetsmatrijs. De keuzes worden bepaald door de toetsconstructeurs, rekening houdend met gemaakte afspraken zoals deze neergelegd zijn in *toetsplannen* (zie kader).

[4] Itembanken zijn vaak onderdeel van een zogenaamd toetsservicesysteem, een geautomatiseerd stelsel van voorzieningen voor het opslaan, terugzoeken en selecteren van items, het samenstellen van toetsen en het analyseren van toetsresultaten.

**Toetsplannen**

Een uitgewerkt plan van toetsing geeft antwoord op de volgende acht vragen:

- Waarom gaat er getoetst worden (uitgangspunten en functies)?
- Bij welke docenten en leerlingen gaat er getoetst worden (doelgroep)?
- Wanneer en waar gaat er getoetst worden (tijdsplanning en plaats van de toetsen binnen het onderwijsprogramma)?
- Wat gaat er getoetst worden (doelstellingen, kennis, inzicht en vaardigheden)?
- Hoe gaat er getoetst worden (toets- en vraagvormen, toetstijd, afname, scoring, beoordeling en waardering)?
- Aan welke eisen moeten de toetsen voldoen (zoals validiteit, betrouwbaarheid en bruikbaarheid)?
- Hoe wordt er bepaald of de opgaven en toetsen aan de kwaliteitseisen voldoen?
- Hoe kan men het opgave- en toetsmateriaal waar nodig verbeteren?

Deze bevatten een beschrijving van de uitgangspunten, functies, inhouden, vormgeving en planning van de toetsing.

Om toetsen samen te stellen die voldoen aan psychometrische, inhoudelijke en praktische specificaties kan men ook gebruik maken van wiskundige modellen. Deze modellen zijn ontleend aan een tak van de wiskunde, aangeduid met operationele research of mathematische programmering, die als doel heeft het ontwikkelen van modellen ter ondersteuning van besluitingvorming. Voor een bespreking hiervan wordt verwezen naar het hoofdstuk Het samenstellen van toetsen (uit: Psychometrie in de praktijk / Theunissen, T.J.J.M, Sanders, P.F., en Verschoor A.J. - Arnhem : Citogroep, 1993).

Cito heeft een computerprogramma OTD (Optimal Test Design) ontwikkeld dat behulpzaam kan zijn bij het samenstellen van toetsen en gebruik maakt van voornoemde modellen.

# 7. Referentiekader

***Het referentiekader gaat in op de wijze van rapporteren van de op toetsen behaalde scores. De scores hebben op zichzelf geen betekenis. De score die een leerling behaalt, krijgt pas betekenis wanneer die score vergeleken wordt met een bepaalde standaard of met de scores die andere leerlingen behaald hebben.***

## 7.1 Normgroep en normschaal

Door het cijfer voor een toetsprestatie te laten afhangen van een vergelijking van deze prestatie met de prestaties van een belangrijke groep personen kan de relatieve waarde van de prestatie beter worden beoordeeld. De vergelijkingsgroep wordt een normgroep of referentiepopulatie genoemd, en een cijferschaal waarop de prestaties van een normgroep zijn af te lezen heet een normschaal.

Voor de constructie van een normschaal moet een zogenaamd normeringsonderzoek worden uitgevoerd. Hiertoe moet in de eerste plaats een normgroep ondubbelzinnig worden afgebakend. Een normgroep is bijvoorbeeld 'alle kinderen in Nederland in groep 8 die niet hebben gedoubleerd'. Het is belangrijk dat een normgroep nauwkeurig is omschreven, zodat precies duidelijk is wie er wel en wie er niet toe behoort. Verder moet zij betekenisvol zijn in relatie tot de toetsresultaten. Als de toets bijvoorbeeld is gericht op het meten van rekenvaardigheden in groep 5 van de basisschool voor de kerstvakantie, dan kan de normgroep precies deze groep bevatten. Echter, als de normgroep beter interpreteerbaar zou worden door alleen de leerlingen te nemen die niet zijn blijven zitten, dan verdient dit de voorkeur.

De constructie van een normschaal vereist dat de frequentieverdeling van de cijfers in de normgroep wordt geschat. Hiertoe moet een representatieve steekproef uit de normgroep worden getrokken. De schatting van de frequentieverdeling is het uitgangspunt voor een ruime keuze aan normschalen. H. Verstralen bespreekt in het hoofdstuk Schalen, normen en cijfers (uit: Psychometrie in de praktijk / Theunissen, T.J.J.M, Sanders, P.F., en Verschoor A.J. - Arnhem : Citogroep, 1993) de volgende vier hoofdtypen van normschalen:

1. cumulatieve verdelingen
2. genormeerde lineaire transformaties
3. genormaliseerde schalen
4. ontwikkelingsschalen

# 8. Handleiding en verantwoording

***De laatste fase van het toetsconstructieproces bestaat uit het maken van een handleiding en instructies voor de diverse categorieën personen die bij de toetsing betrokken zijn.***

Ten behoeve van de opdrachtgever en het wetenschappelijk forum dient een verantwoording geschreven te worden. De COTAN (Commissie Testaangelegenheden Nederland van het Nederlands Instituut van Psychologen (NIP) hanteert een beoordelingssysteem voor de kwaliteit van tests. In dit systeem staan de eisen beschreven waarop toetsmateriaal, handleiding en verantwoording beoordeeld worden. De beoordeling van een test (toets) leidt tot een waardering op de volgende aspecten:

**1. Uitgangspunten van de testconstructie**
Deze categorie telt drie vragen, waarvan één basisvraag[5]. Eerst wordt beoordeeld of het gebruiksdoel en de meetpretentie van de test is aangegeven. Voorts wordt de theoretische achtergrond en de operationalisatie daarvan in de testinhoud beoordeeld. De beoordeling van deze categorie is van invloed op de waardering van de andere categorieën, omdat de meetpretentie bepaalt welk type normerings-, betrouwbaarheids- en validiteitsonderzoek moet worden verricht.

**2a. Kwaliteit van het testmateriaal**
Deze categorie telt zes vragen, waarvan drie basisvragen. In deze categorie komt aan de orde of testopgaven, scoring en instructie zijn gestandaardiseerd. Ook wordt een vraag gesteld over de voor specifieke bevolkingsgroepen mogelijk kwetsende inhoud van items.

**2b. Kwaliteit van de handleiding**
Deze categorie telt zes vragen, waarvan één basisvraag. In deze categorie wordt gevraagd naar de informatie die wordt geboden ter ondersteuning van de testgebruiker bij afname en interpretatie van de test.

**3. Normen**
Deze categorie bevat acht vragen waarvan twee basisvragen. Er wordt vastgesteld of er normen worden verstrekt, of de gekozen normgroepen overeenkomen met het aangegeven gebruiksdoel van de test en wat de kwaliteit is van de normen en de erbij verstrekte informatie.

**4. Betrouwbaarheid**
Deze categorie bestaat uit drie vragen, waarvan één basisvraag. De tweede vraag is verdeeld in vier subvragen waarin de uitkomsten van verschillende typen betrouwbaarheidsonderzoek worden beoordeeld. Met behulp van de derde vraag wordt de kwaliteit van het uitgevoerde onderzoek bij het oordeel betrokken.

**5a. Begripsvaliditeit**
Deze categorie telt drie vragen, waarvan één basisvraag. Eerst worden de uitkomsten en vervolgens de kwaliteit van het uitgevoerde onderzoek naar de begripsvaliditeit beoordeeld.

**5b. Criteriumvaliditeit**
Deze categorie telt eveneens drie vragen, waarvan één basisvraag. Net als bij de begripsvaliditeit wordt eerst de hoogte van de uitkomsten beoordeeld en vervolgens worden deze geëvalueerd in het licht van de kwaliteit van de onderzoeksprocedure.

[5] Met behulp van basisvragen wordt vastgesteld of aan bepaalde minimum vereisten is voldaan, zonder welke verdere beoordeling van de betreffende categorie overbodig wordt of niet mogelijk is.

Het oordeel voor elk van deze categorieën kan zijn 'onvoldoende', 'voldoende' of 'goed'. Een negatief oordeel op een basisvraag leidt direct tot het oordeel 'onvoldoende' voor de betreffende categorie.

# Bijlage: Schematisch overzicht beoordelings-categorieën

## 1. Uitgangspunten van de toetsconstructie

*Basisvraag:*

1.1 Is aangegeven wat het gebruiksdoel is van de test? *(Bij negatieve beoordeling van vraag 1.1 kan men direct doorgaan naar categorie 2.)*

1.2 Is de herkomst van het constructie-idee beschreven en/of word(t)(en) het (de) te meten construct(en) gedefinieerd?

1.3 Wordt de relevantie van de testinhoud voor het (de) te meten construct(en) aannemelijk gemaakt?

## 2. De kwaliteit van het testmateriaal en de handleiding

### 2a. Testmateriaal

*Basisvraag:*

2.1 Zijn de testopgaven gestandaardiseerd?

*Basisvraag:*

2.2.a Is er sprake van een objectief scoringssysteem?

of:

2.2b Indien de scoring door de beoordelaars of observatoren gebeurt, is dan het beoordelings- of observatiesysteem volledig en duidelijk?

*Basisvraag:*

2.3 Zijn de items vrij van racistische of voor bepaalde bevolkingsgroepen kwetsende inhoud? *(Bij negatieve beoordeling van één van de basisvragen (2.1, 2.2 en 2.3) kan men direct doorgaan naar categorie 2B.)*

2.4a Zijn items, testboekje, antwoordschalen en/of antwoordformulier zodanig ontworpen dat fouten bij de invulling kunnen worden vermeden?

2.4b Hoe is de kwaliteit van het testmateriaal?

2.5 Is het scoringssysteem zodanig ontworpen en beschreven dat fouten bij de scoring kunnen worden vermeden?

2.6 Is de instructie voor de geteste volledig en duidelijk?

### 2b. Handleiding

*Basisvraag:*

2.7 Is een handleiding beschikbaar? *(Bij negatieve beoordeling van vraag 2.7 kan men direct doorgaan naar categorie 3.)*

2.8 Zijn de aanwijzingen voor de testleider volledig en duidelijk?

2.9 Wordt informatie gegeven over de gebruikersmogelijkheden en beperkingen van de test?

2.10 Wordt in de handleiding een samenvatting van de onderzoeksresultaten gegeven?

2.11 Wordt met behulp van voorbeelden aangegeven hoe toetsscores kunnen worden geïnterpreteerd?

2.12 Wordt gewezen op soorten informatie die bij de interpretatie van belang kunnen zijn?

2.13 Wordt de mate van deskundigheid die vereist is voor afname en interpretatie van de test vermeld?

## 3. Normen

*Basisvraag:*

3.1 Worden normen (waaronder verwachtingstabellen of grensscores) verstrekt?

3.2 Is (zijn) de normgroep(en) zo gekozen dat deze overeenkomt (overeenkomen) met het gebruiksdoel van de test? *(Bij negatieve beoordeling van de vragen 3.1 en/of 3.2 kan men direct doorgaan naar categorie 4.)*

3.3 Wordt aangegeven naar welke andere groepen kan worden gegeneraliseerd en met welke kans op fouten?

3.4 Worden de betekenis en de beperkingen van het gebruikte type normscore duidelijk gemaakt voor de gebruiker en is het type normscore in overeenstemming met het doel van de test?

3.5 Worden gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling vermeld?

3.6 Worden de standaardmeetfout(en) en/of standaardschattingsfout(en) met de daarbij behorende intervallen vermeld?

3.7 Worden gegevens verstrekt over mogelijke verschillen tussen subgroepen (bijvoorbeeld allochtonen-autochtonen, vrouwen-mannen)?

3.8 Wordt voor elke normgroep vermeld in welk(e) jaar (jaren) de normgegevens zijn verzameld?

## 4. Betrouwbaarheid

*Basisvraag:*

4.1 Worden gegevens over de betrouwbaarheid verstrekt? *(Bij negatieve beoordeling van vraag 4.1 kan men direct doorgaan naar categorie 5.)*

4.2 Zijn de resultaten voldoende gelet op het beoogde type beslissingen dat met behulp van de test moet worden genomen?

a. paralleltestbetrouwbaarheid;
b. interne-consistentiebetrouwbaarheid;
c. test-hertestbetrouwbaarheid;
d. interbeoordelaarsbetrouwbaarheid.

4.3 Beoordeel hieronder de kwaliteiten van het onderzoek:

a. Zijn de procedures op grond waarvan de betrouwbaarheidsgegevens zijn berekend correct?
b. Zijn de steekproeven op grond waarvan de betrouwbaarheidsgegevens zijn berekend in overeenstemming met het beoogde testgebruik?
c. Maken de gegevens die worden verstrekt een gefundeerd oordeel over de betrouwbaarheid mogelijk?

## 5. Validiteit

### 5a. Begripsvaliditeit

*Basisvraag:*

5.1 Worden gegevens over de begripsvaliditeit verstrekt? *(Bij negatieve beoordeling van vraag 5.1 kan men direct doorgaan naar categorie 5b.)*

5.2 Maken de resultaten voldoende aannemelijk dat het begrip zoals bedoeld wordt gemeten (of: maken de resultaten voldoende duidelijk wat wordt gemeten)?

5.3 Beoordeel hieronder de kwaliteit van het onderzoek:

a. Zijn de procedures op grond waarvan de begripsvaliditeitgegevens zijn berekend correct?

b. Komen de steekproeven die in het begripsvalideringsonderzoek zijn gebruikt overeen met groepen waarvoor de test is bedoeld?
c. Wat is de kwaliteit van de andere maten die in het begripsvalideringsonderzoek zijn gebruikt?
d. Maken de gegevens die worden verstrekt een gefundeerd oordeel over de begripsvaliditeit mogelijk?

### 5b. Criteriumvaliditeit

*Basisvraag:*

5.4 Worden er gegevens verstrekt over het verband test-criterium *(Bij negatieve beoordeling van vraag 5.4 kan men de rest van de vragen overslaan.)*

5.5 Zijn de resultaten voldoende gelet op het beoogde type beslissingen dat met de test moet worden genomen?

5.6 Beoordeel hieronder de kwaliteit van het onderzoek:
a. Zijn de procedures op grond waarvan de criteriumvaliditeitsgegevens zijn berekend correct?
b. Zijn de steekproeven op grond waarvan de criteriumvaliditeitsgegevens zijn berekend in overeenstemming met het beoogde testgebruik?
c. Wat is de kwaliteit van de criteriummaten?
d. Maken de gegevens die worden verstrekt een gefundeerd oordeel over de criteriumvaliditeit mogelijk?

## Over deze ToetsSpecial

Handleiding voor toetsconstructeurs met informatie over de acht stappen van het toetsconstructieproces. Samengesteld door Henk Moelands van Cito. Bij samenstelling van deze handleiding is gebruik gemaakt van het boek Psychometrie in de Praktijk onder redactie van Eggen en Sanders (Cito, 1993).