

ToetsSpecial

Serie over Praktijktoetsing, deel 3

Beoordelingsschalen in praktijktoetsen: hoe ontwikkel en gebruik je ze?

Met een checklist voor de evaluatie van de kwaliteit van een praktijktoets

Auteur: [Hans Kuhlemeier](#), onderwijskundige bij [Cito](#)
Datum: september 2017 (herziene versie); oorspronkelijke versie 2002
Internet: <http://www.toetswijzer.nl/berichten/toetsspecial-praktijktoetsen-en-praktische-opdrachten>

Toetswijzer

Toetswijzer is de gids op internet voor iedereen in het onderwijs die betrokken is bij toetsing, examinering en evaluatie. Toetswijzer bevat nieuws, internetlinks en achtergrondinformatie over toetsing in de breedste zin van het woord. Meer weten? Kijk dan op: www.toetswijzer.nl.

Copyright © Stichting Cito

Het auteursrecht op het artikel 'Beoordelingsschalen in praktijktoetsen: hoe ontwikkel en gebruik je ze?', geschreven door Hans Kuhlemeier, berust bij Cito. Overname van het artikel, of onderdelen uit het artikel, is uitsluitend toegestaan na toestemming van de redactie van Toetswijzer. Verzoeken tot overname van kopij dienen schriftelijk te worden gericht aan de [redactie](#) onder vermelding van het medium, de oplage en de doelgroep.

Inhoudsopgave

Inleiding	3
1 Hoe ontwikkel ik een beoordelingsschaal?	4
1.1 Uit welke ontwerpopties kan ik kiezen?	5
1.2 Hoe ontwikkel ik een productschaal?	12
2 Hoe gebruik ik een beoordelingsschaal?	18
2.1 Waarom zou ik beoordelingscriteria onderwijzen?	18
2.2 Hoe onderwijs ik beoordelingscriteria?.....	18
2.3 Hoe voorkom ik storende beoordelaarsfouten?	19
2.4 Hoe train ik beoordelaars?	22
3 Hoe bepaal ik de kwaliteit van een zelfgemaakte praktijktoets?	23
Literatuuroverzicht	27

Inleiding

In deze bijdrage geeft Hans Kuhlemeier achtergronden, vuistregels en tips voor de ontwikkeling en het gebruik van beoordelingsschalen. Dit artikel is het laatste in een serie van drie over praktijktoetsen op Toetswijzer. In de eerste bijdrage is het begrip praktijktoets verhelderd. Het tweede artikel schetst een stapsgewijze procedure aan de hand waarvan docenten zelf een praktijktoets en een bijbehorend beoordelingsmiddel kunnen ontwikkelen. Dit derde artikel gaat dieper in op het maken en gebruiken van beoordelingsschalen. Ter afsluiting wordt een checklist gepresenteerd waarmee docenten de kwaliteit van een zelfgemaakte praktijktoets kunnen beoordelen.

De drie artikelen over praktijktoetsing zijn te vinden op Toetswijzer:
<<http://www.toetswijzer.nl/berichten/toetspecial-praktijktoetsen-en-praktische-opdrachten>>.

Inhoud

Een beoordelingsschaal is meer dan alleen een middel om de prestaties van de leerlingen te beoordelen. Een goede beoordelingsschaal schept heldere verwachtingen en een gemeenschappelijk kader voor kwaliteitsbeoordeling en -verbetering. Dit artikel schetst een stapsgewijze procedure voor het ontwikkelen van beoordelingsschalen. Het maken van een goede beoordelingsschaal is geen sinecure. Ook het gebruik ervan stelt hoge eisen aan de professionaliteit van de docent. Wij hopen dat deze bijdrage de docent een eind op weg kan helpen. De inhoud van deze bijdrage is als volgt:

- 1 Hoe ontwikkel ik een beoordelingsschaal?
 - 1.1 Uit welke ontwerpopties kan ik kiezen?
 - 1.2 Hoe ontwikkel ik een beoordelingsschaal?
- 2 Hoe gebruik ik een beoordelingsschaal?
 - 2.1 Waarom zou ik beoordelingscriteria onderwijzen?
 - 2.2 Hoe onderwijs ik beoordelingscriteria?
 - 2.3 Hoe voorkom ik storende beoordelaarsfouten?
 - 2.4 Hoe train ik beoordelaars?
- 3 Hoe bepaal ik de kwaliteit van een zelfgemaakte praktijktoets?

In hoofdstuk 3 wordt een checklist gepresenteerd aan de hand waarvan de docent de kwaliteit van een zelfgemaakte praktijktoets kan beoordelen.

Aan het eind van dit artikel is een overzicht van geraadpleegde literatuur opgenomen.

1 Hoe ontwikkel ik een beoordelingsschaal?

Bij praktijktoetsen met een open karakter zijn de reacties van de leerlingen vaak zeer gevarieerd. De beoordeling is dan uiterst lastig. Dan kan een beoordelingsschaal uitkomst bieden. Een beoordelingsschaal beschrijft een werkwijze of een product van een leerling op verschillende kwaliteitsniveaus. Met een beoordelingsschaal kan men op een glijdende schaal aangeven in welke mate kennis, vaardigheden of houdingen aanwezig zijn. De glijdende schaal bestaat uit meerdere schaalpunten die een bepaalde positie of rangorde aangeven. Een voorbeeld van een beoordelingsschaal voor een eenvoudige beroepspraktijkvaardigheid is: 1 = beginner; 2 = gevorderd en 3 = expert.

Mits goed geconstrueerd en verstandig gebruikt, bieden beoordelingsschalen verschillende voordelen:

- Ze maken de beoordeling objectiever en consistent.
- Ze helpen de docent bij het verduidelijken van de beoordelingscriteria.
- Ze geven de leerlingen informatie over hoe hun werk beoordeeld wordt en wat er van hen verwacht wordt.
- Ze dragen ertoe bij dat leerlingen zich bewust worden van de beoordelingscriteria en hoe zij de criteria op het werk van hun klasgenoten kunnen toepassen.
- Ze leveren bruikbare feedback over de effectiviteit van het onderwijzen en leren.
- Ze bieden een inhoudelijk kader om de progressie van het leren vast te stellen (lange-termijn doelen).

Het ontwikkelen van een beoordelingsmiddel is de vierde stap in het ontwikkelen van een praktijktoets. Het volgt na de toetsspecificatie (stap 1), het ontwikkelen van de opdrachten (stap 2) en het bepalen van de beoordelingscriteria (stap 3) en gaat vooraf aan het uitproberen en bijstellen van de praktijktoets (stap 5). Het stappenplan voor het ontwikkelen van een praktijktoets is terug te vinden in het tweede artikel over praktijktoetsen.

Om een beoordelingsschaal te kunnen ontwerpen moet men eerst hebben nagedacht over de criteria, dat wil zeggen de eisen waaraan de prestaties van de leerlingen moeten voldoen. Beoordelingscriteria geven een antwoord op de vraag: 'Hoe goed is goed genoeg?' Ze definiëren beheersingsniveaus en maken duidelijk wat we verstaan onder een slechte, voldoende of uitmuntende prestatie. In het tweede artikel over praktijktoetsen hebben wij enkele suggesties gedaan voor het ontwikkelen van bruikbare beoordelingscriteria.

Er bestaat een grote verscheidenheid aan beoordelingsschalen. Een onderscheid wordt wel gemaakt in opdrachtspecifieke, opdracht overstijgende, holistische, analytische, kwantitatieve, kwalitatieve, numerieke, descriptieve en productschalen. Elk type schaal heeft zijn eigen gebruiksmogelijkheden. De docent staat voor de moeilijke taak steeds die beoordelingsschaal te kiezen die gezien de specifieke omstandigheden het beste past. De beste beoordelingsschalen zijn door docenten zelf ontworpen en sluiten daardoor naadloos aan op het gegeven onderwijs. In dit hoofdstuk behandelen we eerst een aantal keuzen waarmee de docent bij het ontwerpen van een beoordelingsschaal wordt geconfronteerd. Daarna schetsen we een stapsgewijze procedure voor het ontwikkelen van een zogeheten productschaal. De indeling is hiermee als volgt:

- Uit welke ontwerptenties kan ik kiezen?
- Hoe ontwikkel ik een productschaal?

1.1 Uit welke ontwerpopties kan ik kiezen?

Voor het ontwerpen van een nieuwe beoordelingsschaal zijn onder meer de volgende ontwerp vragen van belang:

- Opdracht specifiek of opdracht overstijgend?
- Holistisch of analytisch?
- Hoeveel schaalpunten?
- Hoe gedetailleerd omschrijf ik de schaalpunten?
- Kwantitatief of descriptief?
- Illustreer ik de schaalpunten met ankers?

Opdracht specifiek of opdracht overstijgend?

We spreken van een *opdracht specifieke beoordelingsschaal* als de criteria alleen toepasbaar zijn op één specifieke opdracht. Bij een *opdracht overstijgende beoordelingsschaal* zijn de criteria van toepassing op meer gelijksoortige opdrachten, problemen of contexten.

Opdracht specifieke beoordelingsschalen

Opdracht specifieke beoordelingsschalen zijn geschikt als het gaat om duidelijk afgebakende kenniselementen of om procedures die in een vaste volgorde moeten worden doorlopen. Het gebruik ervan is zinvol als een hoge mate van betrouwbaarheid noodzakelijk is, bijvoorbeeld als er belangrijke beslissingen in het geding zijn. Onderstaand beoordelingsmiddel is een voorbeeld van een set opdracht specifieke schalen voor de beoordeling van het opmaken van een bed voor een zieke. De beoordelaar beoordeelt de kwaliteit van elke handeling door een plusje, een plusminnetje of een minnetje te omcirkelen. Bron:

<http://examengids.slo.nl/SloCito/upload/619_06b%20PO%20Bed%20opmaken%20Gredanus%20College.doc>.

Voorbeeld beoordelingsschaal 'Het opmaken van een bed'

Handelingen	Oordeel
Handen wassen	+ ± -
Kiezen voor het juiste materiaal	+ ± -
Werken op de juiste hoogte	+ ± -
Twee stoelen klaarzetten voor afhalen bed	+ ± -
Wasmand klaarzetten voor vuil beddengoed	+ ± -
Molton op matras	+ ± -
Daarover het onderlaken	+ ± -
Onderlaken aan het hoofdeinde goed instoppen	+ ± -
Onderlaken goed strak trekken	+ ± -
Bedzeil aanbrengen	+ ± -
Steeklaken aanbrengen	+ ± -
Bovenlaken en deken aanbrengen	+ ± -
Bovenlaken en deken worden los van elkaar ingestopt (niet te strak)	+ ± -
Slopen kussens vervangen kussens als "huisje" stapelen	
Sprei plaatsen en aan het voeteneinde instoppen	+ ± -
Totaal	

Oprichtoverstijgende beoordelingsschalen

Een opdrachtoverstijgende beoordelingsschaal kan een krachtig onderwijsleermiddel zijn. Dit type schaal is uitdrukkelijk bedoeld voor herhaalde toepassing gedurende een langere periode. De leerlingen passen dezelfde criteria toe op verschillende opdrachten, problemen en contexten. Daardoor raken de kwaliteitseisen op den duur geïnternaliseerd. De leerling leert zo zijn eigen werk te beoordelen, ook al zijn de opdrachten anders en ook al hebben ze de beoordelingsschaal niet op papier voor zich. Kortom, opdrachtoverstijgende beoordelingsschalen kunnen leerlingen helpen bij de transfer van kennis en vaardigheden naar nieuwe situaties met nieuwe problemen. Onderstaand beoordelingsmiddel is een eenvoudig voorbeeld van een opdrachtoverstijgende schaal. Het betreft de beoordeling van de vaardigheid in het presenteren van metingen in grafieken, bron: <<http://www.ascd.org/safeschools/el9710/pophamrubric.html>>. Een grafiek moet niet alleen goed bij de taak en de gegevens passen; de omschrijvingen van de titel, assen en intervallen moeten ook correct zijn. In de beoordelingsschaal worden drie kwaliteitsniveaus onderscheiden: zeer vaardig (score 3), vaardig (score 2) en nog niet vaardig (score 1).

Voorbeeld beoordelingsschaal 'Grafieken'

3	Zeer vaardig <i>De grafiek past perfect bij de taak en de gegevens (bijvoorbeeld staafgrafiek, taartdiagram of lijngrafiek) en de labels bij de titel, assen en intervallen zijn alle correct.</i>
2	Vaardig <i>De grafiek past goed bij de taak en de gegevens (bijvoorbeeld staafgrafiek, taartdiagram of lijngrafiek) en de labels bij de titel, assen en intervallen zijn vrijwel alle correct.</i>
1	Nog niet vaardig <i>De grafiek past nog niet goed bij de taak en de gegevens (bijvoorbeeld staafgrafiek, taartdiagram of lijngrafiek) en minder dan de helft van de labels bij de titel, assen en intervallen zijn correct.</i>

Voors en tegens

Meer informatie over de argumenten voor en tegen opdrachtspecifieke en opdrachtoverstijgende schalen is hieronder weergegeven in een schematisch overzicht.

Argumenten voor opdrachtspecifieke schalen

- De kwaliteit van de leerlingenprestatie is gemakkelijk vast te stellen. Het gaat immers om het registreren van de aan- of afwezigheid van concrete, meestal direct waarneembare kenmerken of inhoudselementen.
- De leerlingen krijgen gedetailleerde feedback op hun prestaties.
- Geschikt voor duidelijk afgebakende kennis.
- Geschikt voor procedures die in een vaste volgorde moeten worden toegepast.
- Een hoge objectiviteit en betrouwbaarheid zijn haalbaar.

Argumenten tegen opdrachtspecifieke schalen

- Het is arbeidsintensief om voor elke opdracht telkens weer een nieuwe schaal te moeten ontwikkelen.
- Door het opdrachtspecifieke karakter is de schaal uitsluitend geschikt voor eenmalig gebruik.
- Het heeft vaak geen zin om de leerlingen van tevoren met de criteria vertrouwd te maken omdat de docent hiermee het antwoord of de oplossing al verklapt.
- Ze zijn niet erg geschikt voor open vraag- en probleemstellingen. De criteria zetten de leerlingen niet aan het denken. Er is weinig transfer van het geleerde naar andere soortgelijke opdrachten, problemen en contexten.
- Bij het nakijken ziet men gemakkelijk a-typische antwoorden of oplossingen over het hoofd.
- Vaak minder geschikt als middel om leerlingen waardevolle leerervaringen op te laten doen.

Argumenten voor opdracht overstijgende schalen

- Ze zijn bij uitstek geschikt voor de ontwikkeling van 'grote', universele, breed toepasbare en permanente vaardigheden zoals kritisch denken, problemen oplossen, communiceren en samenwerken.
- Men kan dezelfde beoordelingsschaal gebruiken bij verschillende opdrachten, problemen en contexten.
- Door de herhaaldelijke toepassing op nieuwe vergelijkbare opdrachten, nemen de leerlingen de criteria eerder op in hun standaardrepertoire (internalisatie); opdracht overstijgende criteria dragen meer bij aan de transfer van het geleerde naar nieuwe, soortgelijke opdrachten, problemen, vaardigheden en contexten.
- Ze bieden meer mogelijkheden om voort te bouwen op hetgeen de leerlingen al kennen en kunnen.
- Ze bieden meer mogelijkheden om de leerlingen bij de kwaliteitseisen te betrekken en deze samen met hen verder te ontwikkelen.
- A-typische antwoorden en oplossingen kunnen gemakkelijk in de beoordeling betrokken worden.

Argumenten tegen opdracht overstijgende schalen

- Ze trekken een zware wissel op de vakinhoudelijke en vakdidactische deskundigheid van de docent.
- Ze veronderstellen kwalitatief hoogwaardig onderwijs waarin vaardigheden systematisch over een langere periode worden aangeleerd.
- Door het vakoverstijgend karakter doen ze een sterk beroep op afstemming en samenwerking binnen en tussen vaksecties.
- Het leren beoordelen kost in het begin vaak veel tijd (omdat docenten en leerlingen moeten leren dezelfde criteria op een grote verscheidenheid aan opdrachten toe te passen).
- Ze doen een sterk beroep op het beoordelingsvermogen van de docent en zijn daarmee meer vatbaar voor subjectiviteit.
- Ze kunnen te algemeen zijn; de feedback voor de docent en leerlingen is dan niet specifiek genoeg.

Holistisch of analytisch?

Voor algemene vaardigheden is een opdracht overstijgende beoordelingsschaal meestal de beste keuze. Wil men zelf zo'n schaal maken, dan zijn er verschillende ontwerpopties. Een daarvan is de keuze tussen een holistische en een analytische schaal.

Met een *holistische beoordelingsschaal* geeft de docent één algemeen oordeel over de kwaliteit van een werkwijze of een product. Bij *analytische beoordeling* worden daarentegen verschillende oordelen op verschillende kwaliteitsdimensies gegeven. Elke dimensie krijgt een afzonderlijke score, waardoor een profiel ontstaat van de sterke en zwakke punten van de leerling. Het geven van één allesomvattend cijfer voor de kwaliteit van een gedocumenteerde schrijfopdracht is een voorbeeld van holistische beoordeling. Geeft de docent daarentegen (ook) cijfers voor afzonderlijke kwaliteitsdimensies – zoals doel- en publiekgerichtheid, inhoud, stijl, organisatie en technische conventies – dan is er sprake van analytische beoordeling.

De voordelen van een holistische beoordelingsschaal zijn evident: het is eenvoudig, kost weinig tijd en geeft een duidelijk overzicht van de prestaties van de leerlingen. Holistische beoordeling is met name bruikbaar als men zich op snelle wijze een beeld wil vormen van de prestaties en men de kwaliteit van de werkwijze of het product op één algemene dimensie kan uitdrukken. Mits het aantal beoordelingsaspecten niet te groot is, heeft analytische beoordeling echter in potentie een hogere onderwijskundige waarde. Een analytische beoordelingsschaal:

- Past bij uitstek bij het beoordelen van complexe vaardigheden.
- Is vooral geschikt als men de leerlingen van tevoren een idee wil geven van de eisen waaraan hun prestatie moet voldoen.
- Is vooral geschikt als men de leerlingen wil leren hoe zij de kwaliteit van hun werk zelf kunnen beoordelen.
- Geeft de leerlingen gedetailleerde feedback over hetgeen zij al kunnen en hetgeen nog nadere uitleg en oefening behoeft.
- Past bij een onderwijsaanpak waarin complexe vaardigheden stapsgewijs worden aangeleerd. Leerlingen kunnen een nieuwe vaardigheid vaak niet in een keer onder de knie krijgen. Een stapsgewijze aanpak verdient dan de voorkeur.

De keuze voor een holistische of een analytische schaal hangt af van het aantal te beoordelen aspecten. Als stelregel kan men daarbij hanteren: hoe complexer de

vaardigheid, hoe meer aspecten men moet beoordelen en hoe meer beoordelingsschalen noodzakelijk zijn voor een goede dekking.

Bij een eenvoudige vaardigheid kan een holistische beoordelingsschaal de beste keuze zijn. Is de vaardigheid complexer, dan ligt een analytische beoordelingsschaal het meest voor de hand.

Voor welk type men kiest, hangt ook af van de toetsfunctie. Wil men de schaal uitsluitend gebruiken voor het geven van een cijfer (resultaatbepaling), dan ligt een holistische beoordelingsschaal het meest voor de hand. Wil men de leerlingen reeds van tevoren vertrouwd maken met de kwaliteitseisen, hen gedetailleerde feedback geven over hun sterke en zwakke punten en hen handvaten bieden voor verbetering (voortgangscntrole en remediëring), dan is een analytische beoordelingsschaal het meest geschikt. De rapportage vindt dan plaats in de vorm van een scoreprofiel, aan de hand waarvan de leerlingen kunnen zien waar zij sterk in zijn en op welke terreinen zij zich nog moeten verbeteren.

Hoeveel schaalpunten?

Heeft men eenmaal voor een beoordelingsschaal gekozen, dan moet men een beslissing nemen over het aantal kwaliteitsniveaus. Het aantal kwaliteitsniveaus bepaalt het aantal punten op de beoordelingsschaal. Een voorbeeld van een eenvoudige beoordelingsschaal voor een beroepspraktijkvaardigheid met drie kwaliteitsniveaus is: 1 = beginner; 2 = gevorderd en 3 = expert. De in Nederland gebruikelijke rapportcijferschaal kent zoals bekend tien schaalpunten (1 = zeer slecht, 2 = slecht, 3 = matig, 4 = onvoldoende, 5 = bijna voldoende, 6 = voldoende, 7 = ruim voldoende, 8 = goed, 9 = zeer goed en 10 = uitmuntend). Dit aantal is voldoende om leerlingen op grond van hun prestaties van elkaar te onderscheiden. Praktijkttoetsen hebben echter vaak een formatief doel. Ze willen laten zien waar de leerling goed in is en waar verbetering nodig is. Voor dat doel is tien schaalpunten onnodig veel. De ervaring leert dat meer dan zeven schaalpunten in de praktijk niet goed werkt. Het lukt de beoordelaars dan niet meer om een betrouwbaar onderscheid tussen de kwaliteitsniveaus te maken. Drie of minder schaalpunten is vaak weer te weinig om de vooruitgang van de leerlingen goed te kunnen rapporteren. Wij adviseren dan ook: minimaal vier en maximaal zeven schaalpunten. Onze persoonlijke voorkeur gaat in veel gevallen uit naar vijfpuntsschalen. Hiermee kan men de progressie van leerlingen nog net goed in kaart brengen, zonder dat het beoordelen voor de docent al te lastig wordt.

Tip

Bij schalen met een oneven aantal schaalpunten neigen beoordelaars nogal eens tot een overmatig gebruik van de middencategorie. Probeer deze tendens naar het midden zoveel mogelijk te vermijden. Breng een spreiding in oordelen aan die in overeenstemming is met de variatie in leerlingprestaties.

Hoe gedetailleerd omschrijf ik de schaalpunten?

De omschrijving van de schaalpunten (kwaliteitsniveaus) van een beoordelingsschaal kan meer of minder gedetailleerd zijn. Over het algemeen geldt: hoe gedetailleerder, hoe meer steun de docent heeft bij het beoordelen en hoe minder de oordelen van verschillende docenten van elkaar zullen verschillen. Een betrouwbare beoordeling is bijvoorbeeld belangrijk als verschillende docenten van een vaksectie een gemeenschappelijke praktijktoets moeten beoordelen. Te veel detail maakt de beoordelaarstaak echter onnodig complex. Nog belangrijker is echter dat de docent de mate van detaillering afstemt op wat de leerlingen aankunnen. Een goede beoordelingsschaal is meer dan alleen een middel om de prestaties van de leerlingen te beoordelen. Een goede schaal schept heldere verwachtingen en een gemeenschappelijk kader voor kwaliteitsbeoordeling en -verbetering.

Zoals gezegd kunnen de omschrijvingen van de schaalpunten meer of minder uitgebreid zijn. Bij een *zuiver numerieke beoordelingsschaal* zijn de schaalpunten niet inhoudelijk gedefinieerd. Het grote voordeel is de eenvoudige en snelle beoordeling, maar de betrouwbaarheid zal over het algemeen niet hoog zijn. Bovendien is een zuiver numerieke beoordelingsschaal voor leerlingen weinig informatief.

Een voorbeeld van een numerieke beoordelingsschaal met minimale verduidelijking van de schaalpunten is de in Nederland gangbare (rapport)cijferschaal. De schaalpunten 1 t/m 10 zijn gedefinieerd door middel van de labels zeer slecht tot en met uitmuntend. Met dit schaaltype kan men leerlingen op basis van hun prestaties met elkaar vergelijken (resultaatbepaling), maar voor doeleinden van voortgangscntrole en remediëring is het nauwelijks geschikt. Daarvoor zeggen de omschrijvingen van de schaalpunten te weinig over de bereikte kwaliteitsniveaus, over hoe goed de leerlingen de kennis en vaardigheden daadwerkelijk beheersen en over wat de leerlingen de volgende keer beter moeten doen.

Tip

Gebruik bij het omschrijven van de schaalpunten geen globale adjectieven als altijd, bijna altijd, soms, zelden of nooit. Dit soort termen zijn voor velerlei uitleg vatbaar en geven de leerlingen weinig feedback over wat zij de volgende keer beter moeten doen.

Kwantitatief of descriptief?

Men kan de schaalpunten (kwaliteitsniveaus) op verschillende manieren verduidelijken. We maken hier een onderscheid in kwantitatieve en descriptieve beoordelingsschalen.

Bij een *kwantitatieve beoordelingsschaal* zijn de schaalpunten gedefinieerd in termen van een aantal kenmerken, zoals het aantal inhoudselementen, begrippen, argumenten, conclusies, literatuurverwijzingen, fouten of misconcepties. De docent moet simpelweg tellen hoe vaak het kenmerk voorkomt. Hoe meer wenselijke kenmerken of hoe minder fouten, des te hoger de score op de schaal. Kwantitatieve beoordelingsschalen zijn vooral geschikt als men de leerlingen zeer gedetailleerde feedback op een specifieke taak wil geven.

Voor algemene vaardigheidsdoelen zijn kwantitatieve schalen minder geschikt. Ze geven al snel een verkeerd signaal ten aanzien van wat belangrijk is om te leren. Excessief gebruik van kwantitatieve schalen kan er bijvoorbeeld toe leiden dat leerlingen spelfouten gaan vermijden door alleen eenvoudige woorden te gebruiken, een verslag overladen met onzinnige literatuurverwijzingen of veel meer conclusies trekken dan de gegevens toelaten. Bovendien geven kwantitatieve schalen vaak een beperkte en sterk vereenvoudigde beschrijving van wat een goede of slechte prestatie nu precies is. In een kwantitatieve schaal voor de beoordeling van argumenteervaardigheid zou men bijvoorbeeld drie goede argumenten van de leerling kunnen verlangen. Echter, wat een standpunt overtuigend maakt, is niet zozeer het aantal argumenten als wel de kwaliteit van de argumentatie. En deze kwaliteit valt moeilijk met een kwantitatieve beoordelingsschaal vast te stellen.

In een *descriptieve beoordelingsschaal* wordt de kwaliteit van de werkwijze of het product verduidelijkt door de schaalpunten van inhoudelijke omschrijvingen te voorzien. Deze omschrijvingen maken duidelijk hoe leerlingwerken van verschillende kwaliteit zich van elkaar onderscheiden. De beoordeling vindt plaats door het werk van een leerling te vergelijken met de omschreven kwaliteitsniveaus. Hiermee krijgt de leerling tegelijkertijd wat meer aanwijzingen voor verbetering.

Voorbeeld van een descriptieve beoordelingsschaal

Onderstaande schaal is een voorbeeld van een descriptieve schaal met een uitgebreide toelichting van de schaalpunten. Het betreft de vaardigheid in het verzamelen van

informatie voor een onderzoeksopdracht bij het vak geschiedenis. Het is een voorbeeld van een opdracht overstijgende beoordelingsschaal die toepasbaar is op allerlei soorten geschiedkundig onderzoek over allerlei onderwerpen. Het is overigens niet altijd nodig om alle schaalpunten uitgebreid te omschrijven. Zonder al te veel verlies aan informatie en functionaliteit had men in dit voorbeeld alleen het middelste schaalpunt en de beide extremen van een nadere toelichting kunnen voorzien.

Voorbeeld beoordelingsschaal 'Verzamelen van informatie'

SCHAAL	OMSCHRIJVING VAN DE SCHAALPUNTEN
1. Slecht	<p>De informatie is afkomstig uit een beperkt aantal bronnen die gemakkelijk toegankelijk zijn en erg voor de hand liggen; het onderzoek is oppervlakkig; belangrijke bronnen zijn over het hoofd gezien en blijven onbenut.</p> <p>De leerling geeft geen informatie over de geloofwaardigheid en potentiële relevantie van de informatiebronnen; neemt vrijwel alle informatie over het onderwerp klakkeloos over.</p> <p>De pogingen om informatie te documenteren zijn inconsistent; de leerling maakt vele fouten in de wijze van documenteren; op grond van veelvuldig voorkomende omissies in de documentatie is er voldoende reden om aan te nemen dat de leerling nog weinig begrip heeft van het doel en de wijze van documenteren.</p>
2. Voor verbetering vatbaar	<p>De informatie is hoofdzakelijk afkomstig uit een beperkt aantal voor de hand liggende bronnen. De werkwijze is niet altijd even goed georganiseerd en er ontbreken enkele voor de hand liggende bronnen.</p> <p>De leerling neemt inconsistente beslissingen over de kwaliteit van de informatiebronnen; maakt ten onrechte geen gebruik van belangrijke bronnen; en gebruikt bronnen die noch bruikbaar noch geloofwaardig zijn.</p> <p>Een deel van de informatie is gedocumenteerd; er zijn echter veel documentatiefouten of doordat de informatie gebrekkig is zijn de conclusies moeilijk te verifiëren.</p>
3. Voldoende	<p>De leerling verzamelt informatie uit een aantal bronnen, met name bronnen die gezien het onderwerp en de situatie voor de hand liggen, en gebruikt een gestructureerde procedure om relevante en potentieel bruikbare bronnen op te sporen.</p> <p>De leerling neemt doorgaans accurate beslissingen over de geloofwaardigheid en bruikbaarheid van de informatiebronnen; ziet echter soms relevante bronnen over het hoofd; gebruikt zo nu en dan slechte bronnen.</p> <p>Vrijwel alle informatiebronnen zijn gedocumenteerd; de manier waarop de bronnen geverifieerd zijn, is acceptabel; wel maakt de leerling kleine verificatiefouten of laat ten onrechte documentatie weg.</p>
4. Goed	<p>De leerling verzamelt informatie uit verschillende bronnen volgens een goed gestructureerd aanpak.</p> <p>De leerling neemt goede beslissingen over de geloofwaardigheid en potentiële bruikbaarheid van de informatiebronnen.</p> <p>De leerling documenteert alle informatie netjes en citeert uit geverifieerde bronnen.</p>
5. Zeer goed	<p>De leerling verzamelt informatie uit een breed scala aan bronnen volgens een gestructureerd onderzoeksplan; er worden zowel gemakkelijk toegankelijke als minder voor de hand liggende bronnen gebruikt; de leerling gebruikt ook bronnen waarvan bekend is dat</p>

deze met zijn of haar interpretatie in tegenspraak zijn.
De leerling neemt volledig accurate beslissingen over de geloofwaardigheid, potentiële relevantie en bruikbaarheid van de gebruikte informatiebronnen.
De leerling besteedt speciale aandacht aan: de reputatie van de uitgever op het onderzoeksgebied; de autoriteit van de auteur en eerdere publicaties; of het al dan niet een primaire bron betreft (bijv. origineel document, onderzoeksverslag, oorspronkelijke taal; publicatiedatum). De leerling documenteert de informatie volledig, citeert uit geverifieerde bronnen; geeft aan waar de bron geverifieerd kan worden.

Illustreer ik de schaalpunten met ankers?

Soms zijn de reacties van de leerlingen zo divers en ongrijpbaar dat zelfs een descriptieve beoordelingsschaal nog te weinig sturing biedt, hoe goed men de schaalpunten ook toelicht. Het is dan raadzaam de schaalpunten te voorzien van zogeheten ankers. Een anker is een voorbeeld van een concreet leerlingenproduct dat het 'niveau' van het schaalpunt illustreert. Zo'n van ankers voorziene schaal wordt wel een *productschaal* genoemd. Het anker kan de vorm hebben van een tastbaar product, zoals een brief, opstel, tekening, foto of bouwconstructie. Bij vluchtige presentaties, zoals een mondelinge presentatie, een dans of een muziekstuk, kan het anker uit een audio- of video-opname bestaan.

Het beoordelen met een productschaal kan men het beste zien als een sorteertaak. De docent beoordeelt elk product door dat te vergelijken met de voorbeeldproducten van de beoordelingsschaal. De docent kent een score toe door het te beoordelen product als het ware op de schaal te plaatsen.

Een productschaal biedt de beoordelende docent een concreet referentiekader waarin de kwaliteit van het leerlingenproduct zichtbaar wordt. Productschalen zijn bij uitstek geschikt om de beoordeling te objectiveren en de leerlingen vertrouwd te maken met de beoordelingscriteria. Docenten en leerlingen ontwikkelen hiermee een gemeenschappelijke definitie van kwaliteit. Een productschaal helpt leerlingen bij het beoordelen en reviseren van hun werkwijzen en producten. Een nadeel is echter dat het maken van een goede productschaal geen sinecure is. In de volgende paragraaf geven wij enkele aanwijzingen voor het maken van een productschaal.

Tips voor het ontwikkelen van beoordelingsschalen

Wellicht ten overvloede presenteren we hier enkele samenvattende tips voor het ontwikkelen van beoordelingsschalen:

- Zorg ervoor dat de beoordelingscriteria volledig duidelijk zijn ten aanzien van wat er van de leerlingen verwacht wordt.
- Maak een keuze tussen een opdrachtverstijgende en een opdrachtspecifieke beoordelingsschaal.
- Indien een opdrachtverstijgende schaal, kies dan voor een holistische of analytische beoordelingsschaal. Splits de beoordeling indien nodig op in verschillende beoordelingsaspecten.
- Laat het aantal beoordelingsschalen afhangen van de complexiteit van de te meten vaardigheid en het doel van de beoordeling, maar maak het aantal schalen niet zo groot dat het gebruik ervan teveel tijd vergt.
- Zorg ervoor dat elke beoordelingsschaal betrekking heeft op een relevant aspect van de te meten vaardigheid.
- Neem een beslissing over het aantal schaalpunten (kwaliteitsniveaus) per schaal. Beperk het aantal schaalpunten tot hooguit vijf à zeven.
- Definieer elk beoordelingsaspect (schaal) zo duidelijk mogelijk.
- Geef een duidelijke omschrijving van de schaalpunten (zodat duidelijk wordt welke verschillen in prestaties corresponderen met de verschillende punten op de schaal).

- Illustreer, indien nodig, de schaalpunten nog verder door middel van voorbeelden van concrete leerlingwerken (ankers).
- Weeg, indien nodig, de beoordelingsaspecten. Maak een eenduidig voorschrift dat aangeeft hoe uit de scores op de beoordelingsschalen een totaalscore valt af te leiden.
- Tracht, indien zinvol, leerlingen bij de ontwikkeling van de beoordelingsschaal te betrekken.
- Zorg ervoor dat het de beoordelaar volledig duidelijk is wat er van hem of haar verwacht wordt.
- Geef aan, indien nodig, door hoeveel docenten het werk van de leerlingen beoordeeld moet worden teneinde een betrouwbaar oordeel te verkrijgen.
- Zorg ervoor dat de beoordelingsschaal hanteerbaar is voor de docent en - indien van toepassing - ook voor de leerlingen.

1.2 Hoe ontwikkel ik een productschaal?

In deze paragraaf laten we aan de hand van een voorbeeld zien hoe docenten een productschaal volgens een stapsgewijze procedure kunnen ontwikkelen. De hier beschreven procedure illustreert tevens hoe men via het ontwikkelen van criteria een bijdrage kan leveren aan het verhelderen van belangrijke maar ongrijpbaar lijkende onderwijsdoelen. We illustreren de verschillende stappen aan de hand van het voorbeeld van de algemene vaardigheid 'zelfreflectie en -evaluatie'.

Stap 1 Verzamel werk van zwakke, middelmatige en goede leerlingen

Verzamel concrete voorbeelden van leerlingproducten die de vaardigheid zo goed mogelijk illustreren. Zorg ervoor dat de hele range aan kwaliteitsniveaus in de verzameling vertegenwoordigd is. Kies met andere woorden producten van zwakke, middelmatige en goede leerlingen. In ons voorbeeld van de algemene vaardigheid 'zelfreflectie en -evaluatie' zou men deze producten kunnen verzamelen door de leerlingen de volgende vragen te stellen:

- Kies een _____ waarover je het meest tevreden bent en waarvoor je het meest je best hebt gedaan.
Op de plaats van de onderstreping noteert de docent een product dat karakteristiek is voor het vak of het onderwijs dat men geeft. Bijvoorbeeld een tekening, compositie, brief, werkstuk, presentatie of onderzoeksverslag.
- Waarom koos je juist dit _____?
- Wat zegt dit _____ over jou als iemand die goed kan _____.
Op de plaats van de tweede onderstreping noteert de docent de vaardigheid waarover gereflecteerd moet worden, zoals tekenen, schrijven, ontwerpen, onderzoeken, problemen oplossen, kritisch denken, informatie opzoeken of conclusies trekken.

Andere vragen die uitnodigen tot zelfreflectie en -evaluatie zouden kunnen zijn:

- Waarom heb je juist deze vragen gesteld of voor deze invalshoek gekozen?
- Ben je tevreden over de manier waarop je deze opdracht hebt uitgevoerd?
- Wat heb je deze keer goed gedaan? Wat heb je van deze opdracht precies geleerd?
- Wat ging nog niet zo goed? Wat zou je de volgende keer anders doen?
- Wat is het belangrijkste wat je dit jaar in de lessen geleerd hebt? Wat zou je er dit jaar nog bij willen leren?

De eerste fase resulteert in een verzameling producten van zwakke, middelmatige en goede leerlingen die karakteristiek is voor de te toetsen vaardigheid.

Stap 2 Leg het werk op stapels en noteer de redenen

Sorteer het werk van de leerlingen in drie stapels: slecht, middelmatig en goed. Toewijzingscriterium is de mate waarin de vaardigheid – in ons voorbeeld 'zelfreflectie en -evaluatie' – bij de leerling ontwikkeld is.

Maak tijdens het sorteren een lijst met kenmerken van zwakke, middelmatige en goede prestaties. Noteer bij elk leerlingwerk de redenen waarom u vindt dat dit werk zwak, middelmatig of goed is. Anders gezegd: schrijf op waarom dit werk nu juist in deze stapel thuis hoort en niet in een andere. Het beste kunt u te werk gaan volgens de hardop denkmethode. Ook kunt u denken aan wat u tegen een collega zou zeggen als u de beslissing om een bepaalde prestatie als zwak, middelmatig of sterk te karakteriseren, zou moeten verantwoorden. Samen vormen de redenen een verdere concretisering van de eisen waaraan het werk van de leerlingen moet voldoen. Deze kenmerken zullen we later gebruiken ter verduidelijking van de schaalpunten (zie stap 4).

Het zoeken naar kenmerken van zwakke prestaties leidt soms tot sterk negatief geformuleerde omschrijvingen, zoals 'helemaal fout', 'hopeloos' of 'domme denkfout'. U kunt dit vermijden door te denken aan het commentaar dat u een zwakke leerling zou geven als u zijn of haar werk zou teruggeven.

Hoe lang moet men doorgaan met het sorteren en beargumenteren van het werk van de leerlingen? Men gaat net zolang door totdat men geen nieuwe kenmerken van zwakke, middelmatige of goede prestaties meer tegenkomt.

Probeer de lijst met redenen in eerste instantie zo lang en divers mogelijk te maken. De opbrengst van deze fase is een groslijst met potentiële kenmerken van zwakke, middelmatige en goede prestaties. Het maakt in dit stadium van de ontwikkeling nog niet uit of de redenen rijp en groen zijn.

Merk op dat het sorteren op zich niet zo belangrijk is. Het is helemaal niet erg als er eens een leerlingwerk op de verkeerde stapel belandt. Het doel is immers te komen tot een volledig en gedetailleerd overzicht van potentiële kenmerken van zwakke, middelmatige en goede prestaties.

Hieronder staan twee voorbeelden van een dergelijke groslijst. Het eerste voorbeeld heeft betrekking op de vaardigheid 'zelfreflectie en -evaluatie' en het tweede op wiskundige communicatievaardigheden.

Voorbeeld 1

Een groslijst met potentiële kenmerken van zwakke en goede prestaties voor de vaardigheid 'zelfreflectie en -evaluatie'.

Zwak

Onnadenkend, geen reflectie op eigen functioneren, geen evaluatie van eerder werk, geen oriëntatie op toekomstige verbetering, vaag, eenvoudige uitspraken, mechanisch, oppervlakkig, te voor de hand liggend, geen voorbeelden, zonder duidelijk doel, ééndimensionaal, overbodig, komt niet verder dan leuk versus niet leuk, neemt geen enkel risico, niet oprecht, niet betrokken, inaccuraat, slordig, onduidelijk, moeilijk leesbaar, slecht georganiseerd et cetera.

Goed

Gedetailleerd, betrokken, origineel, dekkend, geeft blijk van inzicht, verrassend, voorzien van voorbeelden, gemotiveerd, stelt zichzelf doelen, kijkt naar meer dan een ding tegelijkertijd, let zowel op inhoud als werkwijze, neemt risico's, accuraat, goed georganiseerd, eerlijk, oprecht, bevat vergelijking over tijd, kijkt zowel naar sterke als zwakke punten, diepgang, goede argumenten en verklaringen, onthullend, goed leesbaar, gericht op verbetering, persoonlijke reflectie, to the point, vooruitziende blik, grondig, synthese van ideeën, netjes et cetera.

Voorbeeld 2

Een groslijst met potentiële kenmerken van zwakke, middelmatige en goede prestaties voor wiskundige communicatievaardigheid.

Zwak

Ik vind de verschillende stappen in het oplossingsproces niet terug. Het antwoord lijkt niet goed beredeneerd. Gebruikt de verkeerde getallen. Gebruikt het verkeerde oplossingsproces. De leerling lijkt maar wat te doen. Onlogisch. Onduidelijk. Geen uitleg van het antwoord. Ik moet vaak raden wat de leerling precies gedaan heeft. De oplossing bracht mij in verwarring.

Middelmatig

Het idee is goed, maar er zitten fouten in de berekening. Een stap vergeten, maar de rest is in orde. Gaat niet ver genoeg. Geen verduidelijking. Men moet soms maar raden wat de leerling precies gedaan heeft. Laat niet alle stappen zien. Gebruikt correcte gegevens, maar fouten in het oplossingsproces. Het oplossingsproces is goed, maar gebruikt verkeerde gegevens.

Goed

Ik kan precies vertellen wat de leerling bij elke stap gedaan heeft. Beknopt en to the point. Correct antwoord. Logisch en sequentieel. Goede uitleg. Precies. Correcte labels. Verklaart via proces. Geeft argumenten voor redenering. Passend gebruik van figuren, diagrammen of symbolen. Gebruikt de juiste getallen in de juiste volgorde. Stappen in de goede volgorde. Het antwoord is geverifieerd. Ik kan precies zien waarom de leerling een bepaalde stap gezet heeft.

Stap 3 Extraheer één of hooguit enkele belangrijke aspecten

De groslijst zal doorgaans een groot aantal redenen bevatten die sterk op elkaar lijken. Dit is uiteraard niet werkbaar. Voeg soortgelijke redenen daarom samen tot afzonderlijke aspecten (ook wel dimensies genoemd). Dit clusteren van redenen is de derde stap in het maken van een productschaal.

Bij het indikken van de redenen voor het categoriseren van wiskundeprestaties blijken docenten vaak tot drie beoordelingsaspecten te komen: 1) correcte berekening, 2) correcte oplossing, redenering of uitleg, en 3) heldere communicatie. Het beste kan men deze aspecten een nummer geven (1, 2 en 3).

Loop vervolgens de groslijst met redenen nogmaals door en wijs ze toe aan de genummerde aspecten. Plaats het nummer van het aspect achter de reden. Een van de redenen voor het kwaliteitsoordeel 'goed' van het aspect 'heldere communicatie' zou kunnen zijn 'Het is volstrekt duidelijk wat de leerling bij iedere stap in het oplossingsproces gedaan heeft'.

Maak nu ook een onderscheid tussen meer en minder belangrijke redenen. Neem alleen de allerbelangrijkste redenen op.

Probeer bij het clusteren van redenen op zo weinig mogelijk beoordelingsaspecten uit te komen. Vaak zijn er echter meer dan één of twee aspecten nodig. Hoe weet men nu hoeveel aspecten men moet onderscheiden? Een goede graadmeter is het optreden van sorteerproblemen. Als men regelmatig leerlingwerk tegenkomt dat op één aspect van hoge kwaliteit is en op een ander aspect van lage kwaliteit, dan kan men overwegen de kwaliteit op twee aspecten te gaan beoordelen. Een presentatie kan bijvoorbeeld goed georganiseerd zijn, terwijl het taalgebruik zwak is. In ons voorbeeld van zelfreflectie en -evaluatie zou men zelfs vier kwaliteitsaspecten kunnen onderscheiden:

- Analyse en evaluatie van de eigen vaardigheid - wat ging goed en wat ging slecht en waarom is dat zo?
- Oriëntatie op de toekomst - wat moet de volgende keer beter?
- Oprechtheid - hoe serieus, eerlijk en oprecht heeft de leerling het eigen functioneren geanalyseerd en geëvalueerd?
- Presentatie - hoe goed heeft de leerling zijn of haar reflectie en evaluatie gepresenteerd?

Stap 4 Omschrijf de beoordelingsaspecten

De vierde stap is het omschrijven van de beoordelingsaspecten. Het beoordelingsaspect 'oprechtheid' van de vaardigheid 'zelfreflectie en -evaluatie' zou men bijvoorbeeld als volgt kunnen omschrijven:

Oprechtheid is de eerlijkheid en betrokkenheid bij het proces van zelfanalyse en -evaluatie en de oriëntatie op toekomstige taken.

Omschrijf het beoordelingsaspect zo neutraal mogelijk. Definieer het aspect zelf, en niet wat een goede prestatie is. Dit laatste is voorbehouden aan de kenmerken, dit wil zeggen de omschrijvingen van de schaalpunten.

Stap 5 Omschrijf de schaalpunten

Verduidelijk nu elk schaalpunt aan de hand van een beperkt aantal kenmerken (redenen, omschrijvingen). Het schaalpunt 'goed' van het beoordelingsaspect 'analyse en evaluatie van de eigen vaardigheid' zou men bijvoorbeeld van de volgende omschrijving kunnen voorzien:

- kijkt zowel naar sterke als zwakke punten;
- let zowel op de werkwijze als op het product;
- uitgebreid, gedetailleerd, van voorbeelden voorzien;
- geeft oorzaken van eigen functioneren en disfunctioneren;
- geeft een samenvattende conclusie.

Het maken van een goede beoordelingsschaal kost tijd en moeite. Houd er rekening mee dat het nooit in één keer af is. Ideeën hebben tijd nodig om te rijpen. Wees erop voorbereid dat men de set kenmerken bij een volledig nieuw beoordelingsaspect wellicht enkele keren zal moeten herzien.

Stap 6 Illustreer de schaalpunten met concrete leerlingproducten

In veel gevallen geven de omschrijvingen van zwakke, middelmatige en goede prestaties de beoordelende docent en de leerlingen nog te weinig houvast. Het is dan zinvol dat men de schaalpunten nog verder verduidelijkt aan de hand van concrete voorbeelden van leerlingproducten (ankers). Deze voorbeelden maken niet alleen een betrouwbaarder beoordeling mogelijk. Ze geven de leerlingen ook extra aanwijzingen over wat zij moeten doen en nalaten om de opdracht met succes uit te voeren.

Zoek per beoordelingsaspect ten minste drie voorbeelden: één van een zwakke prestatie, één van een middelmatige prestatie en één van een goede prestatie. Zorg ervoor dat elk voorbeeld het desbetreffende schaalpunt zo zuiver mogelijk illustreert. Een voorbeeld bij het schaalpunt 'goed' moet met andere woorden aan alle kenmerken (redenen, omschrijvingen) van een goede prestatie voldoen.

Het zoeken van goede voorbeelden is veel werk. En het vereist een grote mate van nauwkeurigheid. Bij vijf of meer schaalpunten is het doorgaans niet nodig dat men alle schaalpunten van een voorbeeld voorziet. Bij een vijfpuntsschaal kan men bijvoorbeeld vaak volstaan met voorbeelden bij alleen het middelste schaalpunt en de beide uiterste schaalpunten.

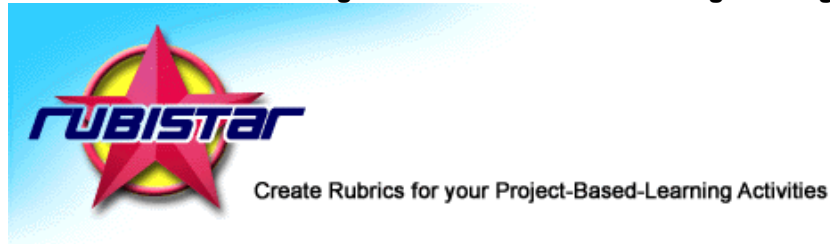
Stap 7 Maak de beoordelingsschaal nog beter

Een goede beoordelingsschaal heeft een lange levensduur. Hij gaat namelijk net zo lang mee als de onderwijsdoelstelling belangrijk is en onderwezen wordt. Criteria evolueren met de toepassing ervan. Heeft men eenmaal een beoordelingsschaal ontwikkeld, dan moet men deze nog uitproberen en op grond van de resultaten bijstellen. Men kan dan ontdekken dat de schaaldefinitie nog niet helemaal juist is, dat men bepaalde omschrijvingen toch maar beter kan verwijderen of dat men een belangrijk kenmerk over het hoofd heeft gezien. Ook kan men ontdekken dat er nog betere voorbeelden te vinden zijn ter illustratie van de schaalpunten. Voorbeelden waarin de kenmerken nog zuiverder vertegenwoordigd zijn en die een nog betere communicatie met de leerlingen mogelijk maken. Een beoordelingsschaal is nooit helemaal af. Het is de moeite waard erin te investeren. Het is immers niet alleen een hulpmiddel bij de beoordeling, maar het kan ook een krachtig onderwijsleermiddel zijn.

Tip

Het is niet altijd noodzakelijk een beoordelingsschaal helemaal zelf te ontwikkelen. Een handig hulpmiddel is het Engelstalige RubiStar, zie: <<http://rubistar.4teachers.org/index.php>>. Deze website bevat een omvangrijke verzameling beoordelingsschalen voor allerlei vakoverstijgende vaardigheden, zoals wiskundige problemen oplossen, onderzoek doen, een planning maken en een website ontwerpen. Voor elke vaardigheid kunnen docenten een keuze maken voor één of meer schalen. Ook kan men de definitie van de schalen en de omschrijvingen van de indicatoren bewerken en de gekozen schalen opslaan en uitprinten.

Voorbeeld van beoordelingsschalen voor samenwerkingsvaardigheid ontleend aan Rubistar



Collaborative Work Skills

Teacher name: _____

Student name: _____

CATEGORY	Excellent	Good	Satisfactory	Needs improvement
Contributions	Routinely provides useful ideas when participating in the group and in classroom discussion. A definite leader who contributes a lot of effort.	Usually provides useful ideas when participating in the group and in classroom discussion. A strong group member who tries hard!	Sometimes provides useful ideas when participating in the group and in classroom discussion. A satisfactory group member who does what is required.	Rarely provides useful ideas when participating in the group and in classroom discussion. May refuse to participate.
Quality of work	Provides work of the	Provides high quality	Provides work that	Provides work that

	highest quality.	work.	occasionally needs to be checked/redone by other group members to ensure quality.	usually needs to be checked/redone by others to ensure quality.
Time management	Routinely uses time well throughout the project to ensure things get done on time. Group does not have to adjust deadlines or work responsibilities because of this person's procrastination.	Usually uses time well throughout the project, but may have procrastinated on one thing. Group does not have to adjust deadlines or work responsibilities because of this person's procrastination.	Tends to procrastinate, but always gets things done by the deadlines. Group does not have to adjust deadlines or work responsibilities because of this person's procrastination.	Rarely gets things done by the deadlines AND group has to adjust deadlines or work responsibilities because of this person's inadequate time management.
Working with others	Almost always listens to, shares with, and supports the efforts of others. Tries to keep people working well together.	Usually listens to, shares with, and supports the efforts of others. Does not cause "waves" in the group.	Often listens to, shares with, and supports the efforts of others, but sometimes is not a good team member.	Rarely listens to, shares with, and supports the efforts of others. Often is not a good team player.
Problem-solving	Actively looks for and suggests solutions to problems.	Refines solutions suggested by others.	Does not suggest or refine solutions, but is willing to try out solutions suggested by others.	Does not try to solve problems or help others solve problems. Lets others do the work.
Attitude	Never is publicly critical of the project or the work of others. Always has a positive attitude about the task(s).	Rarely is publicly critical of the project or the work of others. Often has a positive attitude about the task(s).	Occasionally is publicly critical of the project or the work of other members of the group. Usually has a positive attitude about the task(s).	Often is publicly critical of the project or the work of other members of the group. Often has a positive attitude about the task(s).
Focus on the task	Consistently stays focused on the task and what needs to be done. Very self-directed.	Focuses on the task and what needs to be done most of the time. Other group members can count on this person.	Focuses on the task and what needs to be done some of the time. Other group members must sometimes nag, prod, and remind to keep this person on-task.	Rarely focuses on the task and what needs to be done. Lets others do the work.
Preparedness	Brings needed materials to class and is always ready to work.	Almost always brings needed materials to class and is ready to work.	Almost always brings needed materials but sometimes needs to settle down and get to work.	Often forgets needed materials or is rarely ready to get to work.
Pride	Work reflects this student's best efforts.	Work reflects a strong effort from this student.	Work reflects some effort from this student.	Work reflects very little effort on the part of this student.
Monitors group effectiveness	Routinely monitors the effectiveness of the group and makes suggestions to make it more effective.	Routinely monitors the effectiveness of the group and works to make the group more effective.	Occasionally monitors the effectiveness of the group and works to make the group more effective.	Rarely monitors the effectiveness of the group and does not work to make it more effective.

2 Hoe gebruik ik een beoordelingsschaal?

Beoordelingscriteria kan men niet alleen toepassen op het werk van de leerlingen, men kan ze ook onderwijzen. In dit hoofdstuk geven we een antwoord op de volgende vragen:

- Waarom zou ik beoordelingscriteria onderwijzen?
- Hoe onderwijs ik beoordelingscriteria?
- Hoe voorkom ik storende beoordelaarsfouten?
- Hoe train ik beoordelaars?

2.1 Waarom zou ik beoordelingscriteria onderwijzen?

Goede beoordelingscriteria en -schalen kunnen krachtige onderwijsleermiddelen zijn. Docenten zijn van oudsher geneigd de beoordelingscriteria voor zichzelf te houden. De antwoorden op een meerkeuzetoets of een toets met open vragen geeft de docent natuurlijk niet van tevoren prijs. Geheimhouding valt te verantwoorden als men ervan uit mag gaan dat leerlingen precies weten wat zij moeten doen om een voldoende beoordeling te krijgen. De vraag- of probleemstelling van een praktijktoets is vaak ongestructureerd; er zijn veel goede oplossingen mogelijk. De leerlingen weten dan natuurlijk niet precies wat er van hen verwacht wordt. Door de criteria expliciet te onderwijzen en de beoordelingsschaal van tevoren te bespreken kan men de leerlingen helpen bij de oriëntatie op de taak.

Daarnaast kan een beoordelingsschaal de leerlingen zowel tussentijds als achteraf informatieve feedback geven over hun sterke en zwakke punten. Stel dat uw directeur uw prestaties als docent wil beoordelen. Waaraan zou u dan de voorkeur geven: aan een traditioneel rapportcijfer of aan een gedetailleerd oordeel op een beoordelingsschaal met een uitgebreide omschrijving van de schaalpunten? Een goede beoordelingsschaal beschrijft zowel de fouten die de leerlingen maken als de aspecten waarin zij excelleren. Leerlingen kunnen er meer van opsteken dan van een 'kaal' cijfer.

2.2 Hoe onderwijs ik beoordelingscriteria?

Een van de doelen van vaardigheidgericht onderwijs is dat leerlingen zo vertrouwd raken met de beoordelingscriteria dat zij deze haast als vanzelf toepassen op nieuwe opdrachten, problemen en contexten. Er zijn tenminste drie manieren die ertoe bijdragen dat leerlingen zich de beoordelingscriteria eigen maken.

1. Onderwijsleergesprek

De docent kan de desbetreffende beoordelingsschaal van tevoren in een onderwijsleergesprek aan de orde stellen. Stel iedere leerling een kopie van de schaal ter beschikking. Begin het gesprek met de kenmerken van het hoogste prestatieniveau. Probeer samen met de klas te komen tot overeenstemming over wat een zwakke, middelmatige of goede prestatie is.

2. Leerlingen beoordelen eigen of andermans werk

De docent kan de leerlingen naderhand laten oefenen met het toepassen van de criteria. Dit kan op eigen werk, maar ook op andermans werk. Het is belangrijk dat iedere leerling een eigen exemplaar van de beoordelingsschaal heeft. Voordat de leerlingen beginnen kan de docent de toepassing klassikaal demonstreren en de uitkomsten ervan met de klas bespreken. Als het beoordelingsinstrument uit meer schalen bestaat, oefen het beoordelen dan aspect voor aspect.

3. Criteria samen ontwikkelen

De docent kan de criteria ook samen met de klas ontwikkelen. Als de leerlingen bijvoorbeeld een mondelinge presentatie moeten geven, laat hen dan eerst een video-opname van een excellente professionele presentatie zien en daarna een opname van een slechte presentatie van een leerling van vorig jaar (uiteraard na toestemming van de desbetreffende leerling te hebben verkregen). Focus het onderwijsleergesprek op de overeenkomsten en verschillen tussen beide presentaties en stel samen met de klas een lijst met kenmerken van goede en zwakke prestaties op. Als de leerlingen belangrijke criteria over het hoofd zien, breng deze dan zelf in en leg uit waarom ze belangrijk zijn.

Werk in het gesprek met de klas toe naar drie, vier of hooguit vijf prestatieniveaus. U kunt daarbij gebruik maken van de techniek 'ja, ja maar, nee maar en nee'. Als u de leerlingen bijvoorbeeld wilt leren hoe zij de plot van een verhaal beknopt kunnen samenvatten, beschrijf de vier prestatieniveaus dan als volgt:

- Niveau 1: Ja, ik gaf een korte samenvatting van de plot.
- Niveau 2: Ja, ik gaf een samenvatting, maar ik gaf ook enkele onnodige details of zag belangrijke informatie over het hoofd.
- Niveau 3: Nee, ik gaf geen korte samenvatting, maar ik gaf wel enkele relevante details over het verhaal.
- Niveau 4: Nee, ik gaf geen korte samenvatting.

Wees niet bezorgd of de beoordelingsschaal wel helemaal juist is. Zorg echter wel voor duidelijke beschrijvingen van goede prestaties en van de fouten die leerlingen regelmatig maken. Voor dit laatste kunt u de leerlingen simpelweg vragen naar de fouten die zij in het verleden maakten of naar de fouten die zij denken te gaan maken. Een goede beoordelingsschaal is vrijwel nooit in één keer helemaal perfect. Leg de leerlingen in een volgende les het concept voor en vraag hen om commentaar. Wellicht dat zij u vragen de schaal te herzien.

De heersende toetscultuur in het voortgezet onderwijs brengt met zich mee dat leerlingen hun inzet graag met een cijfer beloond zien. Als u dat nodig vindt of als leerlingen daarom vragen, wijs dan cijfers toe aan schaalpunten. Zo zou u in het voorbeeld hierboven een onvoldoende kunnen toekennen voor niveau 4, een magere zesje voor niveau 3, een zeven of acht voor niveau 2 en een 9 of 10 voor niveau 1.

2.3 Hoe voorkom ik storende beoordelaarsfouten?

Antwoorden op gesloten en open vragen zijn relatief objectief te beoordelen met behulp van eenduidige antwoordmodellen en scoringsvoorschriften. Bij praktijktoetsen is de diversiteit van de reacties van de leerlingen vaak veel groter. Verschillende docenten kennen aan een zelfde leerlingproduct dan verschillende scores (cijfers) toe. Hiermee doet men de leerlingen die ten onrechte een te laag oordeel krijgen onrecht aan. Ter verdediging kunnen docenten aanvoeren dat beoordelingsfouten meestal optreden zonder dat zij dat willen en zonder dat zij zich ervan bewust zijn. Met goede beoordelingsinstructies, -criteria en -instrumenten en een serieuze opvatting van de beoordelingstaak kan men beoordelingsfouten echter een heel eind terugdringen. Op welke bronnen van subjectiviteit moeten docenten dan vooral bedacht zijn? Van belang voor praktijktoetsen zijn met name de volgende storende beoordelaarsfouten:

- halo-effect;
- signifisch effect;
- contaminatie-effect;
- volgorde-effect;
- persoonlijke vergelijkingseffect;
- normverschuivingseffect;
- vermoeidheidseffect.

Halo-effect

Het halo-effect heet in goed Nederlands stralenkrans-effect. Het oordeel over een voor de beoordeling irrelevant aspect 'straalt door' naar het aspect dat men eigenlijk wil beoordelen. Het voor de beoordeling irrelevante aspect kan een algemeen kenmerk zijn dat doorwerkt in de beoordeling van een deelaspect. Ook het omgekeerde is mogelijk, als het oordeel over een deelaspect doorwerkt in het algemene oordeel.

Voorbeeld

Een docent geeft het werk van vlijtige en oplettende leerlingen onbewust een hogere beoordeling dan het even goede werk van een minder vlijtige en oplettende leerling.

Signifisch effect

Verschillende docenten letten bij het beoordelen op verschillende aspecten of wegen deze aspecten verschillend.

Voorbeeld

Bij de beoordeling van een gemeenschappelijke schrijftoets let de ene docent vooral op spelling, interpunctie en grammatica, terwijl de andere docent meer naar de inhoud en de structuur van het schrijfproduct kijkt.

Contaminatie-effect (in engere zin)

Het oordeel is gekleurd doordat de beoordelaar met de toetsing (ook) andere doeleinden dan toetsing nastreeft.

Voorbeeld

Ondanks een matig werkstuk geeft een docent een leerling een voldoende beoordeling omdat deze zo zijn best heeft gedaan en de docent hem of haar niet wil ontmoedigen.
Een docent geeft extra lage cijfers om leerlingen te disciplineren of hen aan het werk te krijgen.

Volgorde-effect

De nawerking van voorafgaande beoordelingen van het werk van andere leerlingen bij het beoordelen.

Voorbeeld

Na een reeks zwakke opstellen wordt een middelmatig opstel ten onrechte overgewaardeerd.

Persoonlijke vergelijkingseffect

Elke beoordelaar heeft zijn eigen karakteristieke wijze van beoordelen. De een is strenger of juist milder dan de ander of brengt meer of juist minder spreiding aan.

Voorbeeld

Een gelijke prestatie beoordeelt de ene docent strenger of juist milder dan de ander.
De ene docent gebruikt bij het beoordelen alle schaalpunten, terwijl de ander alleen de middencategorieën gebruikt.

Normverschuivingseffect

De neiging om zich bij het beoordelen aan te passen aan het 'gemiddeld' niveau van de klas. Ook bij zeer goede beoordelingsschalen is dit effect moeilijk volledig uit te schakelen.

Voorbeeld

De ene klas is veel beter dan de andere, maar toch is het gemiddelde oordeel in de ene klas niet hoger dan in de andere. De docent beoordeelt de leerlingen uit een goede klas derhalve strenger dan die uit een minder goede klas.

Vermoeidheidseffect

De hoogte van het oordeel is afhankelijk van de fysieke en psychische gesteldheid van de beoordelaar.

Voorbeeld

Na het zoveelste opstel nemen vermoeidheid en irritatie toe en wordt het oordeel lager.

Tip 1

Wat valt er te doen als twee docenten sterk verschillen in hun beoordeling? Als de een bijvoorbeeld veel strenger is dan de ander? Er zijn in beginsel vier mogelijkheden om docenten meer op één lijn te krijgen:

- Scherp de richtlijnen voor het beoordelen aan.
- Betrek een derde persoon bij de beoordeling, bij voorkeur iemand met een hoge vakinhoudelijke status.
- Organiseer een trainingssessie met behulp van concrete producten van leerlingen.
- Neem het gemiddelde over de oordelen van de verschillende docenten.

De laatste mogelijkheid zien wij als een noodoplossing. Grote beoordelingsverschillen tussen docenten kunnen wijzen op verschillen in visie binnen de groep docenten. Die kan men beter eerst uitdiscussiëren.

Tip 2

Binnen vaksecties worden vaak gemeenschappelijke praktijktoetsen afgenomen. Hoe kan een sectie er nu toe bijdragen dat de docenten bij het beoordelen dezelfde maatstaven aanleggen? In deze situatie kan een trainingsbijeenkomst nuttig zijn.

Voor de training is het volgende materiaal benodigd:

- Een beoordelingsschaal.
- Richtlijnen voor het beoordelen.
- Een setje ankers met voor elk schaalpunt een concreet voorbeeld van een leerlingproduct. Gebruik alleen voorbeelden die de desbetreffende schaalpunten zo zuiver mogelijk illustreren.

Hanteer daarbij de volgende strategie:

- Leg de docenten een concreet leerlingproduct ter beoordeling voor.
- Vergelijk de oordelen van de docenten onderling en met uw eigen oordeel.
- Vergelijk het te beoordelen leerlingproduct met de ankers.
- Bespreek de discrepanties met het doel tot meer overeenstemming te komen.
- Ga na in hoeverre de training effect gesorteerd heeft. Laat de docenten daartoe enkele leerlingproducten onafhankelijk van elkaar beoordelen en bepaal de overeenstemming.
- Pas zo nodig de richtlijnen voor het beoordelen aan, de definitie van de beoordelingsaspecten en/of de kenmerken (omschrijvingen, voorbeelden) van de schaalpunten.

2.4 Hoe train ik beoordelaars?

Met goede beoordelingsinstructies, -criteria en -instrumenten en een oplettende houding ten opzichte van mogelijke fouten kunnen docenten beoordelingsfouten een heel eind terugdringen. Daarnaast kan een beoordelaarstraining nuttig zijn. Een training kan in sectieverband worden gegeven, bijvoorbeeld aan de hand van leerlingproducten die via een gemeenschappelijke praktijktoets zijn verkregen. De leiding van de training kan in handen zijn van de sectieleider, een vakdocent of een externe trainer (bij voorbeeld van de Citogroep). Hieronder laten we zien hoe men zo'n training kan opzetten.

Stap 1 Oriëntatie op de beoordelingstaak

Geef de deelnemers informatie over het doel van de training, de verschillende stappen, de opdrachten voor de leerlingen, de criteria, de beoordelingsschalen en de relatie met de onderwijsdoelstelling waarvan men de beheersing wil vaststellen.

Stap 2 Verduidelijk de beoordelingscriteria

Stel het beoordelingsinstrument in een groepsdiscussie aan de orde. Bespreek de beoordelaarsinstructies, de beoordelingsaspecten, de schaalpunten en de omschrijvingen van de schaalpunten. Verhelder de schaalpunten zo nodig nog verder aan de hand van voorbeelden van concreet leerlingwerk (ankers). Zorg ervoor dat de belangrijkste interpretatieverschillen uitgediscussieerd zijn voordat de docenten gaan oefenen met het toepassen van de criteria op concreet werk van leerlingen.

Stap 3 Oefen het beoordelen

De kern van iedere beoordelaarstraining is het oefenen met 'echt' werk van leerlingen. Begin eenvoudig. Neem eerst producten waarin de onderscheiden kenmerken van een zwakke, middelmatige of goede prestatie zo zuiver mogelijk vertegenwoordigd zijn. Stel a-typische of anderszins moeilijk te beoordelen leerlingwerken pas aan de orde als het eenvoudige werk geen problemen meer oplevert.

Stap 4 Reviseer zo nodig het instrument

Tijdens het oefenen moet men soms vaststellen dat de beoordelingsregels niet helemaal sluitend zijn. Een deel van de leerlingen heeft de opdracht bijvoorbeeld op een andere manier geïnterpreteerd dan de bedoeling was, zonder dat hun interpretatie als helemaal fout gerekend moet worden. Men kan dan beslissen het beoordelingsinstrument aan te passen.

Stap 5 Bepaal de overeenstemming

Bepaal regelmatig de overeenstemming tussen de beoordelaars. Vergelijk de oordelen van de verschillende docenten en bespreek eventuele discrepanties. Stop met het oefenen van het desbetreffende beoordelingsaspect als er een aanvaardbaar niveau van overeenstemming bereikt is. In het geval van een vijfpuntsschaal kan men bijvoorbeeld afspreken te stoppen op het moment dat de oordelen nooit meer dan één punt uit elkaar liggen.

3 Hoe bepaal ik de kwaliteit van een zelfgemaakte praktijktoets?

Aan het einde van deze serie van drie artikelen over praktijktoetsing op [Toetswijzer](#) presenteren wij een checklist aan de hand waarvan de docent de kwaliteit van een zelfgemaakte praktijktoets kan beoordelen. Deze checklist is nog in ontwikkeling. Uw praktijkervaringen en eventuele verbeteringssuggesties zijn van harte welkom. U kunt daartoe via e-mail contact opnemen met Hans Kuhlemeier van de Citogroep: <hans.kuhlemeier@cito.nl>.

Checklist voor de evaluatie van de kwaliteit van een praktijktoets

Toetsspecificatie	
Is de functie van de praktijktoets duidelijk (bijv. instructie en oefening, voortgangscontrole en/of resultaatbepaling)?	<input type="checkbox"/>
Is een praktijktoets een goede keuze gezien de functie van de toets?	<input type="checkbox"/>
Is er een duidelijk omschreven doelstelling die men wil onderwijzen en waarvan men de beheersing wil vaststellen?	<input type="checkbox"/>
Is een praktijktoets het meest geschikte middel gezien het leerdoel, de leerstof en het gegeven onderwijs?	<input type="checkbox"/>
Is duidelijk wat de leerlingen na afloop moeten kennen en kunnen?	<input type="checkbox"/>
Is de vaardigheid die men wil onderwijzen en toetsen van belang voor het dagelijks leven en/of de latere beroepspraktijk?	<input type="checkbox"/>
Heeft de praktijktoets betrekking op duurzame kennis? Gaat het om een universele, breed toepasbare vaardigheid?	<input type="checkbox"/>
Wordt alleen het belangrijkste gemeten (geen triviale details)?	<input type="checkbox"/>
Worden er geen belangrijke aspecten van wat men wil onderwijzen en toetsen over het hoofd gezien?	<input type="checkbox"/>

De opdrachten van de praktijktoets	
Bestaat de praktijktoets uit meer opdrachten?	<input type="checkbox"/>
Zijn de opdrachten tot op zekere hoogte vergelijkbaar in de zin dat ze alle een beroep doen op dezelfde overkoepelende vaardigheid?	<input type="checkbox"/>
Passen de opdrachten bij de doelstelling en de beoogde kennis, vaardigheden en/of houdingen?	<input type="checkbox"/>
Is de vraag- of probleemstelling open in de zin dat er meer goede antwoorden of oplossingen mogelijk zijn?	<input type="checkbox"/>
Zijn de leerlingen voldoende vrij in het bepalen van de werkwijze, de benodigde materialen en gereedschappen, het uiteindelijke product en de verwerkings- of presentatievorm?	<input type="checkbox"/>
Zijn de opdrachten realistisch en authentiek?	<input type="checkbox"/>
Zijn de opdrachten uitdagend en motiverend?	<input type="checkbox"/>
Zijn de opdrachten noch te moeilijk noch te gemakkelijk?	<input type="checkbox"/>
Geven de opdrachten de leerlingen voldoende gelegenheid hun toegenomen kennis en vaardigheden te demonstreren?	<input type="checkbox"/>

Zijn de opdrachten voorbeelden van goed onderwijs?	
Sluiten de opdrachten goed aan bij eerder opgedane kennis en vaardigheden?	
Zijn er mogelijkheden tot zelfreflectie en -evaluatie?	
Is de instructie voor de leerlingen voldoende duidelijk? Is het taalgebruik adequaat en afgestemd op de groep leerlingen?	
Is duidelijk wat de verwerkings- of presentatievorm is (indien van toepassing)?	
Zijn de opdrachten beoordeeld (door leerlingen, collega's, toetsdeskundigen), uitgetoetst en op grond van de resultaten bijgesteld?	
Is duidelijk welke (hulp)middelen (materiaal, gereedschap, apparatuur) de leerlingen mogen/moeten hanteren?	

Beoordelingscriteria en -middelen	
Passen de criteria en het beoordelingsmiddel bij het doel van de toets, de leerstof en het gegeven onderwijs?	
Geven de beoordelingscriteria een duidelijk antwoord op de vraag 'Hoe goed is goed genoeg?' Krijgen de leerlingen een reële indruk van wat er van hen verwacht wordt?	
Is alles wat van belang is in de criteria vertegenwoordigd?	
Bevatten de criteria geen irrelevante eisen?	
Zijn de criteria voldoende duidelijk? Is duidelijk wat de kenmerken zijn van zwakke, middelmatige en goede prestaties?	
Zijn de criteria voldoende algemeen (opdracht overstijgend in plaats van opdrachtspecifiek)? Gelden ze met andere woorden voor een brede range aan opdrachten, problemen en/of contexten?	
Past de beoordelingsmethode bij de aard van de te nemen beslissingen (bijv. holistische beoordeling voor resultaatbepaling en analytische beoordeling voor voortgangscontrole en remediëring)?	
Heeft elk beoordelingsaspect betrekking op een belangrijk kwaliteitskenmerk van de gevolgde werkwijze en/of het gemaakte product?	
Zijn de beoordelingsaspecten (schalen) duidelijk omschreven?	
Zijn de schaalpunten voorzien van duidelijke omschrijvingen (kenmerken)?	
Zijn de schaalpunten, indien nodig, voorzien van ankers (dit wil zeggen: leerlingproducten die laten zien hoe onvoldoende, voldoende en uitmuntende prestaties zich van elkaar onderscheiden)?	
Biedt het hoogste schaalpunt de leerlingen werkelijke uitdaging zonder onhaalbaar te zijn? En is het laagste schaalpunt niet triviaal?	
Is de beoordelingsschaal makkelijk te hanteren?	
Zijn de schaalordelen bruikbaar en makkelijk te interpreteren?	
Zijn de oordelen objectief in de zin dat leerlingen met dezelfde prestatie een zelfde beoordeling krijgen, onafhankelijk van wie er beoordeelt en wanneer de beoordeling plaatsvindt?	
Is er een beoordelaarsinstructie?	
Zijn er richtlijnen voor het omzetten van oordelen in cijfers (indien van toepassing)? Is bij groepswork duidelijk hoe het cijfer tot stand komt?	

Randvoorwaarden

Is er een duidelijke afname-instructie voor de docent?	
Zijn er voldoende hulpmiddelen, materiaal en gereedschap voorhanden (bijv. computer met internet) en zijn deze hulpmiddelen voor alle leerlingen makkelijk verkrijgbaar of toegankelijk?	
Zijn de activiteiten voldoende veilig (bijv. geen gevaarlijk gereedschap of gevaarlijke routes)?	
Is duidelijk hoeveel tijd de leerlingen hebben?	
Is duidelijk waar de leerlingen de opdrachten uitvoeren?	
Is duidelijk of de leerlingen de opdrachten van de praktijktoets alleen en/of in groepjes maken?	
Is duidelijk hoe zelfstandig de leerlingen mogen/moeten werken en op welke momenten zij welke vormen van hulp en begeleiding kunnen krijgen (en welke gevolgen dat heeft op de hoogte van de beoordeling)?	

Validiteit

Meet de zelfontwikkelde praktijktoets inderdaad wat de docent ermee wil meten?	
Roepen de opdrachten van de praktijktoets de beoogde leeractiviteiten ook daadwerkelijk op? Passen de leerlingen de beoogde vaardigheden inderdaad toe?	
Doen de leerlingen tijdens het maken van de opdrachten en het hanteren van de beoordelingscriteria waardevolle leerervaringen op?	
Is alles wat van belang is om te toetsen in de toets vertegenwoordigd? Worden er belangrijke kennis en vaardigheden over het hoofd gezien? Worden er niet onbedoeld triviale aspecten in de toetsing betrokken?	
Sluit de inhoud van de praktijktoets goed aan bij wat de leerlingen al wisten en konden?	
Dekken de opdrachten en de beoordelingscriteria het gegeven onderwijs? Zijn de leerlingen op het moment van toetsing voldoende in de gelegenheid (geweest) om zich de getoetste kennis en vaardigheden eigen te maken?	
Kan men op grond van de gemaakte opdrachten een geldige conclusie trekken over de vaardigheid van de leerling? Anders gezegd: vormen de opdrachten een representatieve steekproef uit de verzameling van alle opdrachten die men bij de te toetsen vaardigheid zou kunnen ontwikkelen?	
Heeft het gebruik van de praktijktoets een positieve invloed op onderwijzen en leren?	
Brengt het gebruik van de praktijktoets geen onbedoelde (neven)effecten met zich mee? Een goede praktijktoets maximaliseert de positieve (neven)effecten en minimaliseert de negatieve.	
Geeft de praktijktoets de leerlingen de juiste boodschap over wat belangrijk is om te leren?	
Is de verzamelde informatie relevant voor de te nemen beslissing? Is de 'bewijsvoering' voor de te nemen beslissing voldoende?	

Betrouwbaarheid

Geeft de docent verschillende leerlingen voor eenzelfde prestatie eenzelfde beoordeling?	<input type="checkbox"/>
Geeft de docent voor eenzelfde prestatie eenzelfde beoordeling als zijn of haar collega's?	<input type="checkbox"/>
Is de hoogte van de beoordeling afhankelijk van het moment waarop de docent de prestatie beoordeelt? Als de docent dezelfde prestatie enige tijd later nogmaals beoordeelt, is het oordeel dan hetzelfde als bij de eerste beoordeling?	<input type="checkbox"/>
Is alles gedaan wat menselijkerwijs mogelijk is om storende beoordelaarsfouten uit te schakelen?	<input type="checkbox"/>

Bruikbaarheid

Staat de verkregen informatie in verhouding tot de investering (qua tijd, moeite en financiën)?	<input type="checkbox"/>
Is de beoogde vaardigheid onderwijsbaar en toetsbaar gegeven de groep leerlingen en de beschikbare tijd en middelen?	<input type="checkbox"/>
Is de afname van de praktijktoets uitvoerbaar gegeven de groep leerlingen en de beschikbare tijd en middelen?	<input type="checkbox"/>
Is de beoordeling praktisch uitvoerbaar?	<input type="checkbox"/>

Literatuuroverzicht

Bij het schrijven van de drie artikelen over praktijktoetsing is onder meer gebruik gemaakt van de hieronder vermelde literatuur. Voorbeelden van praktijktoetsen en beoordelingsmiddelen zijn onder meer ontleend aan de websites <www.cito.nl>, <www.toetswijzer.nl> en <www.examengids.nl>.

- Arter, J. (1998). *Improving classroom assessment: a toolkit for professional developers (Toolkit98)*. Portland, OR: Northwest Regional Educational Laboratory.
- Baron, J.B. & Wolf, D.P. (1996). *Performance-based student assessment: challenges and possibilities*. Chicago: University of Chicago.
- Bennett, R. & Ward, W. (1993). *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Brown, J.H. & Shavelson, R.J. (1996). *Assessing hands-on science: a teacher's guide to performance assessment*. Thousand Oaks, CA: Corwin Press.
- Bruyne, H.C.D. de (1983). *Evalueren in de klas*. Amsterdam: Van Goor Zonen.
- Bügel, K. & Sanders, P.F. (1998). *Richtlijnen voor de ontwikkeling van onpartijdige toetsen*. Arnhem: Cito. Beschikbaar via internet: <http://www.cito.nl/onderzoek%20en%20wetenschap/psychometrie/richtlijnen_onpartijdige_toetsen>.
- Cito (1998). *Toetstechnische begrippenlijst*. Arnhem: Cito. Beschikbaar via internet: <<http://www.cito.nl/onderzoek%20en%20wetenschap/kenniscentrum/begrippenlijst>>.
- Cohen, E. (1992). *Designing group work (2nd edition)*. New York: Teachers College Press.
- Dousma, T., Horsten, A. & Brants, J. (1997). *Tentamineren*. Maastricht: Vakgroep Onderwijsontwikkeling en Onderwijsresearch.
- Drenth, P.J.D. & Sytsma, K. (1990). *Testtheorie: inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn Stafleu Van Loghum.
- Ebbens, S., Ettekoven, S. & Rooijen, J. van (1997). *Samenwerkend leren: praktijkboek*. Groningen: Wolters-Noordhoff.
- Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs: Prentice Hall.
- Eisner, E.W. (1999). The Uses and Limits of Performance Assessment. *Phi Delta Kappa International*, 5. Beschikbaar via internet: <<http://www.kiva.net/~pdkintl/kappan/keis9905.htm>>.
- Erkens, T.T.M.G. & Moelands, H.A. (1992). *Toetsen met open vragen: een handleiding voor het construeren van toetsen met open vragen*. Arnhem: Cito. Beschikbaar via Toetswijzer op internet: <www.toetswijzer.nl>.
- Foster, M. & Masters, G. (1996). *Performances assessment resource kit: doing outcomes, planning observations, judging & recording, estimating achievement, reporting*. Melbourne: Australian Council for Educational Research.
- Groot, A.D. de & Naerssen, R.F. van (1973). *Studietoetsen: construeren, afnemen, analyseren. Deel I*. Den Haag: Mouton.
- Groot, A.D. de & Naerssen, R.F. van (1975). *Studietoetsen: construeren, afnemen, analyseren. Deel II*. Den Haag: Mouton.
- Groot, A.D. & Wijnen, W.H.F.M. (1983). *Vijven en zessen: cijfers en beslissingen: het selectieproces in ons onderwijs*. Groningen: Wolters-Noordhoff.
- Hermans, P. (1992). *Denken en doen: een oriënterende notitie over het toetsen van praktische vaardigheden bij de afsluiting van de basisvorming*. Arnhem: Cito.
- Hibbard, M., Wagenen, L. van & Lewbel, S. (1996). *Performance-based learning and assessment: a teacher's guide*. Alexandria: Association for Supervision and Curriculum Development.
- Hoogenboom, J. & Ribot, C. (1993). *Praktijktoetsen: constructie en beoordeling van praktische opdrachten*. Arnhem: Cito.

- Houtman, I.L.D. & Brants, J. (1983). *Een literatuuronderzoek naar het samenstellen van toetsen voor praktische vaardigheden*. Amsterdam: Vrije Universiteit.
- Hubert, F. & Pilot, A. (1974). *Specificeren van onderwijsdoelstellingen. Een overzicht van een aantal benaderingen die dienen voor het verhelderen van onderwijsleerdoelen van vooral kleinere onderwijsprogramma's*. Utrecht: Rijksuniversiteit Utrecht.
- Kagan, S. (1990). *Cooperative learning. Resources for teachers*. San Juan Capistrano, CA: Kagan Cooperative Learning.
- Kok, J., Krieken, R. van & Luijten, A.J.M. (1986). *Het construeren van open vragen*. Arnhem: Cito.
- Kroft, M. van der & Sinkeldam, R. (1998). *Handleiding praktische opdrachten tweede fase havo/vwo*. Arnhem: Cito.
- Krogt, M., van der (1998). *Eindverslag netwerk examendossier*. Arnhem: Cito.
- Kuhlemeier, H. (1997). *Toetsing van algemene vaardigheden in de afsluitingstoetsen basisvorming*. Arnhem: Cito.
- Kuhlemeier, H. (1999). Het maken van toetsen bij methoden. In: Instituut voor Toetsontwikkeling (red.), *Het maken van toetsen bij methoden. Handreikingen voor auteurs basisvorming* (pp. 9-64). Enschede: PrintPartners Ipskamp.
- Kuhlemeier, H. & Bergh, H. van den (1998). Relationships between language skills and task effects. *Perceptual and Motor Skills*, 86, 443-463.
- Linn, R.L. (1989). *Educational measurement (3rd ed.)*. Washington: American Council on Education.
- Lindeman, H. (1999). *Het examendossier, het toetsen van vaardigheden*. Utrecht: APS. Beschikbaar via internet: <<http://www.aps.nl/vo/aardr/examendossier%20toetsen%20vaardigheden.doc>>.
- Mager, R.F. (1974). *Leerdoelen formuleren. Hoe doe je dat?* Groningen: Wolters-Noordhoff.
- Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappa International*, 1999, 5. Beschikbaar via internet: <<http://www.kiva.net/~pdkintl/kappan/kmab9905.htm>>.
- Mehrens, W.A. (1988). Consequences of assessment: what is the evidence? *Education Policy Analysis Archives*, 13, 6, 1-39. Beschikbaar via internet: <<http://olam.ed.asu.edu/epaa/v6n13.html>>.
- Meuffels, B. (1993). Cijfers en cijferschalen. *Spiegel*, 11, 1, 49-69.
- Meuffels, B. (1994). *De verguisde beoordelaar*. Amsterdam: Thesis Publishers.
- Ministerie van Onderwijs, Cultuur en Wetenschappen (1998). *Kerndoelen 1998-2003. Relaties in beeld. Over de relaties tussen de algemene doelen en de kerndoelen per vak*. Den Haag: Sdu Servicecentrum.
- Moelands, H.A., Noijons, J. & Rem, J. (1992). *Toetsen met gesloten vragen: een handleiding voor het construeren van toetsen met meerkeuzevragen*. Arnhem: Cito. Beschikbaar via Toetswijzer op internet: <www.toetswijzer.nl>.
- Moerkerke, G. (1996). *Assessment for flexible learning. Performance assessment, prior knowledge state assessment and progress assessment as new tools*. Utrecht: Lemma.
- Molen, H. van der (1998). *Naar een schoolbrede aanpak van de basisvorming. Een handreiking voor de evaluatie en invoering van de algemene onderwijsdoelen*. Enschede: Technimedia.
- Nederlands Instituut voor Psychologen (1988). *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen*. Amsterdam: Nederlands Instituut van Psychologen.
- Popham, W. (1997). What's wrong--and what's right--with rubrics. *Educational Leadership*, 55, 2. Beschikbaar via internet: <<http://www.ascd.org/safeschools/el9710/pophamrubric.html>>.
- Sanders, P.F. (1980a). *Een inleiding tot toetsen praktische vaardigheden (Specialistisch Bulletin Nr. 5)*. Arnhem: Cito.
- Sanders, P.F. (1980b). *Een procedure voor de constructie van Toetsen Praktische Vaardigheden (Specialistisch Bulletin Nr. 9)*. Arnhem: Cito.

- Schotten, J.G.M. & Robroek, W.C.L. (1997). *Verantwoord meten in het gezondheidszorgonderwijs. Handboek voor beoordeling van leerresultaten*. Houten: Bohn Stafleu Van Loghum.
- Sinkeldam, R. (1998). *Handleiding profielwerkstuk*. Arnhem: Cito.
- Slavin, R.E. (1992). When and why does cooperative learning increase achievement? Theoretical and empirical perspectives. In: R. Hertz-Lazarowitz & N. Miller (Eds.). *Interaction in cooperative groups. The theoretical anatomy of group learning* (pp. 145-173). Cambridge: University Press.
- Slavin, R.E. (1995). *Cooperative learning: theory, research and practice (2nd edition)*. Boston: Allyn & Bacon.
- Sluifster, C., Kleintjes, F.G.M., Schalk, H.H., Roosmalen, W.W.M. van, Hermans, P.H.L. & Bogaerts, C.A.M.J. (1996). *De constructie van beoordelingsschalen bij afsluitingstoeten voor de basisvorming. Uitgangspunten en een voorbeeld*. Arnhem: Cito.
- Stiggings, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practices*, 6, 33-41.
- Stokking, K. (1997). *Algemene vaardigheden in het curriculum. Verkenning en programma*. Utrecht: Vakgroep Onderwijskunde/ISOR.
- Straetmans, G. & Sanders, P. (2001). Beoordelen van competenties van docenten. (EPS-brochurereeks nr. 5). Beschikbaar via internet: <<http://www.educatiefpartnerschap.nl/docs/eps/brochure05.pdf>>.
- Thorndike, R.L. (1971). *Educational Measurement (2nd ed.)*. Washington: American Council on Education.
- Traub, R. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In: R. Bennett & W. Ward (eds.). *Construction versus Choice in Cognitive Measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Vermeulen, W. (1993). *Toetsing van communicatieve vaardigheden. Constructie en evaluatie van gedragstoetsen voor professionele gespreksvaardigheden*. Dissertatie. Arnhem: Cito.
- Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs*. Den Haag: Staatsuitgeverij.
- Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation*, 2 (2).