Beoordelen van werkstukken

Een voorbeeld uit het tandheelkundig onderwijs

G. J. J. M. STRAETMANS
Katholieke Universiteit Nijmegen

Samenvatting

In een tandheelkundig practicum worden docenten getraind in het beoordelen van werkstukken. Ten behoeve van de training is een werkstukkenverzameling aangelegd, waarin voor ieder werkstuk de kwaliteit op de beoordelingsaspecten is vastgelegd in een referentie-oordeel. De beoordelingsaspecten zijn geordend in twee niveaus: kenmerken en subkenmerken. Tijdens de training worden werkstukken eerst op kenmerkniveau beoordeeld. Voor kenmerken waarop geen overeenstemming wordt bereikt met het referentie-oordeel, moeten de subkenmerken beoordeeld worden. Daarvoor wordt een beoordelingsprotocol gehanteerd. De beoordelingsresultaten laten zien, dat bij hantering van de protocollen de overeenstemming met het referentie-oordeel aanzienlijk hoger is dan op kenmerkniveau.

1 *Inleiding*

Bijna de helft van de beschikbare tijd in de opleiding voor tandarts wordt besteed aan het aanleren van praktische vaardigheden. De basis voor het leren beheersen van deze vaardigheden wordt gelegd in het preklinisch motorisch onderwijs. Evaluatie vindt in deze leerfase plaats door middel van zogenaamde 'worksample tests'. Bij dergelijke tests moeten de geëxamineerden de verworven vaardigheden aanwenden onder condities die de daadwerkelijke werksituatie simuleren. Het gebruik van de 'fantoomkop' staat hierbij centraal. Dit is een metalen of plastic model van een menselijk hoofd, waarin natuurlijke of kunststof elementen geplaatst kunnen worden. De gehanteerde technieken vertonen zoveel mogelijk overeenkomst met die welke gebruikt worden bij echte patiëntbehandeling. Het grote voordeel van 'work-sample tests' is dat ze direct de te evalueren vaardigheden kunnen meten. Deze grote potentiële validiteit staat in schril contrast met de constatering van vele onderzoekers in binnen- en buitenland, dat de kwaliteit van de evaluatie van het preklinisch en klinisch motorisch onderwijs te wensen overlaat. Met name wordt getwijfeld aan de betrouwbaarheid van de evaluatie, gedefinieerd als de mate van overeenstemming binnen een beoordelaar of tussen een aantal beoordelaars. Voor het onderwijs heeft dit ernstige gevolgen:

- onbetrouwbare zak/slaagbeslissingen;
  Als beoordelaars verschillende cijfers toekennen voor hetzelfde werkstuk, dan komen hun oordelen tot stand op basis van niet dezelfde of zelfs van irrelevante factoren.
- inefficiënt leerproces;
  Vaardigheden kunnen niet efficiënt aangeleerd worden als de terugkoppeling naar de lerende inconsistent is.
- frustratie van studenten.
  Het zal duidelijk zijn dat inconsistente evaluatie ongewenste emoties kan oproepen bij studenten, waardoor het leren negatief beïnvloed wordt.

De oorzaken van de onbetrouwbare beoordelingen kunnen gevonden worden in het gebrek aan goed gedefinieerde prestatiecriteria en dito beoordelingsschalen. De oplossing lijkt voor de hand te liggen: de ontwikkeling van beoordelingssystemen waarmee prestaties op zo objectief mogelijke wijze kunnen worden beoordeeld.

In een artikel waarin een overzicht wordt gegeven van het Amerikaanse onderzoek op het gebied van de tandheelkundige klinische evaluatie, concluderen Patridge & Mast (1978) dat '. . . onderzoek slechts heeft kunnen aantonen dat meetschalen weinig schaalpunten moeten bevatten, dat gebruik moet worden gemaakt van prestatiecriteria en dat die criteria in zo objectief mogelijke termen gesteld moeten worden.'

Ook in Nederland is onderzoek verricht naar het beoordelen van tandheelkundige werkstukken. Zonder volledig te willen zijn wordt bij enkele ervan stilgestaan.

Pedagogische Studiën 1984 (61) 218–226

- Steures & Tromp (1978) doen verslag van de vernieuwing van het preklinisch practicum aan de Universiteit van Amsterdam. Op basis van een analyse van te verrichten taken en het gewenste prestatieniveau zijn nieuwe instructiematerialen en objectieve beoordelingscriteria vervaardigd. Er wordt gebruik gemaakt van beoordelingsschema's (goed/fout-scoring) waarop de criteria door trefwoorden zijn aangegeven. Zowel docenten als studenten vullen de schema's in bij het beoordelen van oefen- en toetswerkstukken. Onderzoek wees uit dat de verschillen in beoordelingen tussen docenten en studenten significant waren. Echter ook tussen de docenten onderling bleken de verschillen aanzienlijk. Dit ondanks de ingevoerde trainingen, die tot doel hadden verschillen tussen docenten in de beoordelingen vast te stellen en terug te dringen door bespreking van de mogelijke oorzaken.
- Wiegman (1982) rapporteert over de ontwikkeling van een beoordelingssysteem voor 'opwaswerkstukken'[1] aan de Groningse Subfaculteit Tandheelkunde. Werkstukken worden beoordeeld aan de hand van een lijst met 25 goed/fout-items. Voor alle items zijn criteria omschreven evenals de te gebruiken meetinstrumenten. De intra- en inter-beoordelaarsovereenstemmingen die bereikt werden, varieerden respectievelijk van 70-95% en van 25-75%. Aanbevolen werd, om een trainingsprogramma te ontwikkelen om docenten op een gedegen wijze met het beoordelingssysteem vertrouwd te maken.
- In Nijmegen concludeert de Subcommissie 'Toetsing en Beoordeling Motorische vaardigheden' in haar verslag (Otto, 1981) dat '... zowel bij het toekennen van cijfers voor, als bij het nemen van voldoende/onvoldoende-beslissingen over werkstukken (c.q. handelingen) de tussenbeoordelaarsovereenstemming op totaalniveau matig tot laag is.'

Het onderzoek waarvan in dit artikel verslag wordt uitgebracht heeft, net als de hiervoor besproken onderzoeken, tot doel om de betrouwbaarheid van werkstukbeoordelingen in het preklinisch onderwijs op een aanvaardbaar niveau te brengen. Daartoe werden een nieuwe beoordelingsmethode (het 'beoordelingsprotocol') en een trainingsprogramma ontwikkeld voor één type werkstuk: de klasse II preparatie.[2] Tot op heden wordt de klasse II preparatie beoordeeld op een zestal kenmerken. Deze kenmerken zijn over het algemeen vaag geformuleerd, waardoor de besproken betrouwbaarheidsproblemen kunnen ontstaan. Anders dan bij de hiervoor genoemde onderzoeken wordt geen tweepuntsschaal (goed/fout) gehanteerd maar, in verband met de verbeterde terugkoppeling naar de student, een driepuntsschaal. Een ander aspect waarop het onderhavige onderzoek verschilt van soortgelijk onderzoek, betreft de opzet van de beoordelaarstraining. Door gebruik te maken van een microcomputer en doordat gewerkt wordt met een werkstukkenbestand, kan geïndividualiseerd getraind worden.

2 *Het beoordelingsprotocol*

'Beoordelingsprotocol' is een verzamelnaam voor de omschrijving, de beoordelingsmethode en het scoringsvoorschrift van te onderscheiden aspecten aan tandheelkundige werkstukken. In dit protocol worden de in het preklinisch onderwijs gebruikte kenmerken geoperationaliseerd door middel van subkenmerken. In totaal zijn er 32 subkenmerken verdeeld over 6 kenmerken. Voor elk subkenmerk wordt omschreven:

1. aan welke eisen werkstukken moeten voldoen;
2. hoe vastgesteld dient te worden of ze aan die eisen voldoen;
3. hoe de observatie in een score vertaald moet worden.

ad 1. De eisen voor de subkenmerken werden ontleend aan de in het preklinisch onderwijs gebruikte geïndividualiseerde cursus 'Preparatie en Restauratie I en II' (1981). Om die eisen in zo objectief mogelijke bewoordingen te kunnen formuleren, werden aanwijzingen van Mackenzie (1974) gevolgd:

- Criteria moeten gerelateerd zijn aan de doelstellingen. Dat wil zeggen, dat de nadruk moet liggen op dié aspecten die direct van invloed zijn op de kwaliteit van het werkstuk.
- Criteria dienen operationeel (meetbaar) gedefinieerd te worden. In plaats van: 'De preparatiebreedte is adequaat' formuleert men als volgt: 'Bij premolaren mag de preparatiebreedte variëren van 1.0-1.3 mm'.

- Duidelijk moet zijn wat nog wel en wat niet meer acceptabel is. Het voorbeeld van de 'preparatiebreedte' is hier ook van toepassing.
- Met behulp van illustraties of modellen worden de criteria verduidelijkt.

ad 2. Voor elk subkenmerk is aangegeven hoe vastgesteld moet worden of het te beoordelen werkstuk aan de geformuleerde eisen voldoet. In het protocol zijn drie beoordelingsmethoden te onderscheiden: meten, schatten en vergelijken. De breedte en diepte van een preparatie worden gemeten met behulp van tandheelkundig instrumentarium. Hoeken, gevormd door op elkaar staande wanden, worden geschat. De afwerking van een preparatie wordt beoordeeld door die te vergelijken met de afwerking van een referentie-werkstuk. Referentie-werkstukken zijn door eerstejaars studenten vervaardigde klasse II preparaties, die volgens overeenstemmend oordeel van een aantal stafleden, nog net voldoende zijn afgewerkt. Een werkstuk is voldoende afgewerkt als de kwaliteit op dit aspect vergelijkbaar is met die van het referentie-werkstuk.

ad 3. Bij de keuze van een beoordelingsschaal wordt men tevens gedwongen te kiezen tussen zo hoog mogelijke overeenstemming en zo optimaal mogelijke terugkoppeling. De overeenstemming is gebaat bij een klein aantal schaalpunten. De terugkoppeling daarentegen bij een groot aantal (Houpt & Kress, 1973; Hinkelman & Long, 1973). Gekozen werd voor een driepunts-schaal. Alle subkenmerken, behalve die onder 'afwerking', worden gescoord op een nominale schaal en hebben alleen een classificatiefunctie. De '2'-score betekent dat aan de gestelde eisen voldaan is. Score '1' en '3' houden in dat niet aan de eisen voldaan is en verwijzen elk naar een andere geconstateerde fout. Voor de subkenmerken onder 'afwerking' wordt een ordinale schaal gehanteerd; hoe hoger de score hoe beter de afwerking van het werkstuk.

In een voorstudie werd het beoordelingsprotocol getest (Straetmans, 1982). Aan studenten en stafleden werd gevraagd werkstukken (afkomstig uit het werkstukkenbestand) te beoordelen met gebruikmaking van het beoordelingsprotocol. Bovendien kregen ze als opdracht mee om de omschrijvingen kritisch te bekijken en, indien nodig, van kanttekeningen te voorzien. Aan de hand daarvan werden de protocollen bijgesteld. Figuur 1 laat een pagina zien uit de laatste versie van het beoordelingsprotocol.

3 *Het werkstukkenbestand*

Uit een groot aantal door eerstejaars studenten vervaardigde practicumwerkstukken werden 35 werkstukken geselecteerd. In het bestand werden ongeveer even veel voldoende als onvoldoende werkstukken opgenomen. De geselecteerde werkstukken werden met bijbehorend buurelement gepositioneerd in kleine plastic bakjes. Op de bakjes werden identificatienummers aangebracht. Alle werkstukken werden door drie ervaren instructeurs beoordeeld. Uit deze oordelen werden referentie-oordelen gedestilleerd voor alle kenmerken en subkenmerken. De referentie-oordelen werden opgeslagen in het geheugen van een microcomputer.

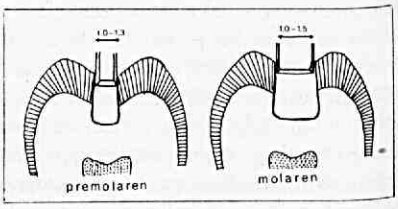
1.2. PREPARATIEBREEDTE VAN DE STEP

KENMERK : OUTLINE
TYPE : Klasse II

Bij premolaren mag de breedte van de step variëren van 1.0-1.3 mm.
Bij molaren mag de breedte van de step variëren van 1.0-1.5 mm.

De preparatie is te smal als glazuurmes 10-8-12 niet door de isthmus heen kan en te breed als glazuurmes 13-8-12 en 15-8-12 door de isthmus van resp. premolaar en molaar heen kunnen.

SCORING : Preparatie-breedte is te smal = 1
Preparatie-breedte is juist = 2
Preparatie-breedte is te groot = 3



Figuur 1 *Een pagina uit het beoordelingsprotocol*

4 *De trainingen*

Het doel van het trainingsprogramma is tweeledig:
- Beoordelaars leren omgaan met een gestandaardiseerde en geobjectiveerde beoordelingsprocedure.
- Beoordelaars terugkoppeling geven over geleverde beoordelingsprestaties, teneinde de overeenstemming te bevorderen.

Vijf docenten uit het preklinisch onderwijs namen deel aan de training die uit zes sessies bestond. Organisatorisch bleek het onmogelijk om de sessies met alle docenten tegelijk te laten plaatsvinden. Het geïndividualiseerde karakter van het trainingsprogramma maakte het echter mogelijk om de docenten afzonderlijk te trainen. Het gebruik van een microcomputer speelde hierin een belangrijke rol. Om terugkoppeling over de beoordelingsprestaties te krijgen waren docenten niet afhankelijk van de aanwezigheid van collega's. De computer gaf terugkoppeling door voor het beoordeelde werkstuk het referentie-oordeel op te zoeken en (eventueel) de eerder door collega-docenten gegeven oordelen. Iedere sessie bestond dus uit vijf individuele trainingen. Nadat alle beoordelaars aan bod waren geweest werd de sessie afgesloten met een plenaire zitting ter bespreking van de beoordelingsprestaties. Er was dan tevens gelegenheid om kritiek te leveren op de protocollen, de werkstukken en de trainingsopzet.

Het beoordelen in een individuele trainingssessie geschiedde in eerste instantie op de wijze zoals gebruikelijk in de Nijmeegse prekliniek; dat wil zeggen op kenmerkniveau. De scores werden ingevuld op schrapkaarten en vervolgens ingevoerd in de microcomputer via een kaartlezer. De microcomputer zocht voor elke werkstuk-kenmerkcombinatie het referentie-oordeel op en de, eventueel, door andere beoordelaars gegeven oordelen over dezelfde werkstuk-kenmerkcombinaties. Alle relevante informatie werd door een printer afgedrukt. Bij afwijkingen van het referentie-oordeel kreeg een beoordelaar opdracht, om voor het betreffende kenmerk de beoordeling te specificeren naar subkenmerken. In dat geval werd gebruik gemaakt van het beoordelingsprotocol. Ook over de beoordelingen op subkenmerkniveau kregen de beoordelaars onmiddellijke terugkoppeling via de printer.

5 *Resultaten*

Belangrijke vragen zijn:
1. Worden werkstukken betrouwbaarder beoordeeld aan de hand van het beoordelingsprotocol?
2. Is er sprake van trainings-effecten? De aandacht richt zich hierbij op de overeenstemming met het referentie-oordeel én op de benodigde tijd per subkenmerkoordeel.

Zoals in de probleemomschrijving werd vermeld, wordt de betrouwbaarheid gemeten door de overeenstemming tussen beoordelingen vast te stellen. De overeenstemming (tussen docent-oordeel en referentie-oordeel) wordt beschreven door het percentage identieke beoordelingen en door Cohens Kappa (Cohen, 1960). Laatstgenoemde maat corrigeert de waargenomen overeenstemming voor kans.

5.1 *Kenmerk- versus subkenmerkniveau*

Beide hiervoor genoemde overeenstemmingsmaten werden berekend op zowel kenmerk- als subkenmerkniveau, daarmee vergelijking van die twee beoordelingsmethoden mogelijk makend. In Tabel 1 en 2 worden respectievelijk de procentuele overeenstemming en Cohens Kappa gegeven voor de beoordelingen op kenmerk- en subkenmerkniveau. De berekeningen werden uitgevoerd over alle, in zes trainings-sessies, beoordeelde werkstukken en laten per beoordelingsaspect zien hoe groot de overeenstemming is voor elke beoordelaar apart en voor alle beoordelaars tesamen. Geconstateerd kan worden dat de overeenstemming met het referentie-oordeel op subkenmerkniveau groter is dan op kenmerkniveau. Dit geldt voor alle beoordelingsaspecten, behalve voor 'afwerking', waarvoor in de meeste gevallen het omgekeerde geldt. De vermoedelijke oorzaak daarvan is gelegen in het referentie-oordeel op subkenmerkniveau. In veel gevallen was dit aan de hoge kant en kwamen de beoordelaars unaniem tot afwijkende scores.

Door de opzet van de training wordt een vergelijking tussen de overeenstemmingen op kenmerk- en subkenmerkniveau bemoeilijkt. Immers, doordat beoordeling op subkenmerkniveau alleen maar plaatsvond als géén overeenstemming bereikt werd met het referentie-oordeel op kenmerkniveau, is op subkenmerkniveau meestal geen volledige informatie be-

Tabel 1 *Percentage overeenstemming met het referentie-oordeel op kenmerk- (K) en subkenmerkniveau (SK), uitgesplitst naar beoordelingsaspecten (OU tot en met AF[3]) en beoordelaars (1 tot en met 5)*

	OU		DI		CA		CO		PA		AF	
	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	53	74	55	75	49	86	45	75	51	68	55	54
2	51	79	40	69	47	83	45	76	47	68	62	43
3	62	72	53	72	53	86	43	74	47	75	51	48
4	47	74	34	66	60	71	45	75	47	60	55	44
5	62	74	49	71	45	74	36	77	53	50	64	49
TOT.	55	75	46	70	51	80	43	75	49	64	57	48

schikbaar over werkstukken. Een ander nadeel van de besproken opzet betreft de invloed van de terugkoppeling op kenmerkniveau op de beoordelingsprestaties op subkenmerkniveau. De extra kritische instelling van een beoordelaar als gevolg van die terugkoppeling én de informatie die verkregen wordt middels het verstrekte referentie-oordeel over de kwaliteit van het werkstuk op het betreffende aspect, kunnen in het voordeel werken van de subkenmerkmethode. De gesignaleerde hogere overeenstemmingen op subkenmerkniveau mogen, op basis van deze gegevens, dan ook niet enkel en alleen worden toegeschreven aan het gebruik van de beoordelingsprotocollen.

In een poging om een meer directe vergelijking te kunnen maken tussen de kenmerk- en de subkenmerkmethode werd een maand na afloop van de laatste trainings-sessie opnieuw een beroep gedaan op de trainees. Hun werd gevraagd om zes, at random uit het werkstukkenbestand getrokken, werkstukken op alle subkenmerken te beoordelen. De werkstukken werden dus direct op subkenmerkniveau beoordeeld, zonder voorafgaande kennis van de kwaliteit op kenmerkniveau. De op deze subkenmerk-beoordelingen gebaseerde overeenstemmingen kunnen direct vergeleken worden met de uit de trainings-sessies stammende beoordelingen en overeenstemmingen op kenmerkniveau. Tabel 3 geeft per beoordelaar de overeenstemmingen weer met het referentie-oordeel, berekend over zes werkstukken. Beoordelaar 3 kon door langdurige afwezigheid niet aan deze beoordelingsronde deelnemen.

De verschillen in overeenstemming tussen de kenmerk- en de subkenmerkmethode zijn groot als de overeenstemming wordt uitgedrukt in percentages. Wordt de gevonden overeenstemming gecorrigeerd voor kansovereenstemming (coëfficiënt Kappa) dan zijn de verschillen tussen de beoordelingsmethoden veel geringer en soms (bij beoordelaar 2) zelfs afwezig. Alleen bij beoordelaar 5 is het verschil in overeenstemming tussen de kenmerk- en de subkenmerkmethode statistisch significant op

Tabel 2 *Reële overeenstemming (Kappa × 100) met het referentie-oordeel op kenmerk- (K) en subkenmerkniveau (SK), uitgesplitst naar beoordelingsaspecten (OU tot en met AF) en beoordelaars (1 tot en met 5)*

	OU		DI		CA		CO		PA		AF	
	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	18	30	33	41	25	46	20	27	27	46	5	16
2	19	42	10	25	17	44	18	24	21	42	13	10
3	33	32	29	37	30	39	11	20	20	60	8	11
4	5	39	6	31	40	7	17	26	21	31	11	10
5	25	32	20	28	17	4	5	22	31	17	20	13
TOT.	20	36	19	32	26	27	14	24	24	39	11	12

Tabel 3 *Overeenstemming met het referentie-oordeel, berekend over zes werkstukken*

	KENMERKEN		SUBKENMERKEN	
	%	KAPPA	%	KAPPA
BEOORDELAARS				
1	56	0.32	78	0.43
2	58	0.35	74	0.35
4	54	0.29	73	0.31
5	44	0.13	76	0.35

het 5%-niveau. Deze vrij kleine verschillen in reële overeenstemming worden veroorzaakt door het conservatieve karakter van Cohens Kappa. Dit is een gevolg van een assumptie van Cohens Kappa, die luidt dat beoordelaars weten welke marginale waarden ze in een kruistabel moeten reproduceren. Een beoordelaar zou bijvoorbeeld weten dat van de tien te beoordelen werkstukken er zes voldoende zijn, één goed en drie onvoldoende. Naar alle waarschijnlijkheid zal deze beoordelaar zich in zijn beoordeling voor een belangrijk deel laten leiden door die informatie. Bij gevolg zullen de marginale proporties nauw met elkaar overeenstemmen, wat weer leidt tot een hoge kansovereenstemming en dus tot een lage Kappa. In het concrete geval van het beoordelen van tandheelkundige practicumwerkstukken, echter, is geen sprake van aan beoordelaars bekende marginalen. Zij worden dus ten onrechte gestraft voor het bereiken van marginale overeenstemming. Dit in overweging genomen kan de conclusie niet anders luiden dan dat beoordelingen op subkenmerkniveau betrouwbaarder zijn dan beoordelingen op kenmerkniveau.

5.2 *Trainings-effecten*

Uit de tweede vraagstelling blijkt wat in het concrete geval verstaan wordt onder trainingseffect. In de eerste plaats een steeds hogere overeenstemming tussen docent-oordeel en referentie-oordeel. In de tweede plaats een systematische reductie van de benodigde beoordelingstijd, zonder dat dit ten koste gaat van de overeenstemming. Training heeft zin als voor één van de omschreven effecten bewijs kan worden gevonden. Slechts in het geval dat de overeenstemming tussen docent-oordeel en referentie-oordeel niet stijgt en bovendien de benodigde beoordelingstijd niet gereduceerd wordt, is training niet zinvol.

5.2.1 *Systematische toename in de overeenstemming?*

In Tabel 4 worden per trainings-sessie de overeenstemmingen met het referentie-oordeel gegeven van elke beoordelaar. Aangezien de Kappa-coëfficiënten hetzelfde beeld laten zien wordt volstaan met de procentuele overeenstemmingen.

Tabel 4 *Percentage overeenstemming met het referentie-oordeel op kenmerkniveau (K) en subkenmerkniveau (SK), uitgesplitst naar beoordelaars (1 tot en met 5) en trainings-sessie (I tot en met VI)*

	I		II		III		IV		V		VI	
	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	39	75	45	66	63	70	50	72	46	80	61	68
2	42	79	43	59	52	74	56	80	48	70	48	73
3	56	74	50	61	46	71	63	79	46	70	50	66
4	47	80	60	53	42	70	44	67	50	63	46	64
5	47	61	50	66	44	71	60	76	56	75	50	64
GEM.	46	74	50	61	49	71	55	75	49	72	51	67

Alle gemiddelde overeenstemmingspercentages per sessie liggen op kenmerkniveau rond de 50 procent. Afwijkingen naar boven en naar onder zijn gering en moeten aan het toeval worden toegeschreven. Op subkenmerkniveau schommelen de gemiddelde overeenstemmingen per sessie rond de 70 procent. Ook hier is geen sprake van systematische toename in de overeenstemming als de beoordelaars meer training hebben gehad. Een opvallende daling in de gemiddelde overeenstemming doet zich voor bij de tweede sessie. Nadere bestudering van de beoordelingsscores leerde, dat in die sessie vooral voor de subkenmerken van het beoordelingsaspect 'afwerking' de overeenstemming erg laag was. Zoals al eerder werd gezegd, moet getwijfeld worden aan de kwaliteit van de referentie-oordelen met betrekking tot dit beoordelingsaspect.

5.2.2 *Systematische reductie van de beoordelingstijd?*

Aangenomen mag worden, dat, als gevolg van de jarenlange ervaring die docenten hebben met het beoordelen op kenmerkniveau, de beoordelingstijd van de kenmerken niet verkort kan worden door training. Voor de subkenmerken ligt dit anders. In het begin zullen docenten de aanwijzingen in het protocol aandachtig moeten lezen. Na enige training zal men zich passages kunnen herinneren en na nog meer training is het mogelijk dat men het protocol 'van-buiten-kent'. Doordat steeds de tijdstippen werden genoteerd waarop een beoordelaar met de training aanving en weer eindigde, kan voor elke beoordelaar nagegaan worden hoeveel tijd gemiddeld besteed werd aan één subkenmerk-oordeel. Tabel 5 bevat de gemiddelde beoordelingstijden per subkenmerk.

Een duidelijk trainings-effect tekent zich af; beoordelingstijden worden korter als de beoordelaars meer sessies achter de rug hebben. De kleiner wordende standaardafwijking geeft aan dat het trainings-effect zich bij alle beoordelaars voordoet en niet alleen wordt veroorzaakt door het middelen van de beoordelingstijden.

6 *Discussie*

De volgende conclusies lijken gerechtvaardigd:

- De overeenstemming met het referentie-oordeel is op subkenmerkniveau aanzienlijk groter dan op kenmerkniveau.
- Noch op kenmerkniveau noch op subkenmerkniveau wordt de overeenstemming met het referentie-oordeel groter als gevolg van training.
- De benodigde tijd voor het beoordelen op subkenmerkniveau wordt door training flink gereduceerd.

Theoretisch gezien verdient het beoordelen aan de hand van het beoordelingsprotocol de

Tabel 5 *Gemiddelde beoordelingstijden per subkenmerk in seconden, berekend per sessie (I tot en met VI) en per beoordelaar (1 tot en met 5)*

	I	II	III	IV	V	VI
1	37	21	24	19	25	21
2	65	48	45	42	35	29
3	65	42	54	41	37	28
4	75	69	59	37	41	40
5	61	36	34	37	35	23
GEM.	61	43	43	35	35	28
ST. AFW.	14	18	14	9	6	7

voorkeur. Door de grotere objectiviteit wordt de overeenstemming tussen beoordelaars (en dus de betrouwbaarheid) bevorderd. Voor de student betekent het beoordelen op subkenmerkniveau, dat hij/zij uitvoerige en betrouwbare terugkoppeling krijgt over de geleverde prestatie. In samenspraak met de instructeur kan op basis van die beoordeling een goede beslissing genomen worden over de continuering van het onderwijsleerproces. Nodeloze herhalingen worden daardoor beperkt. Deze kwantitatieve winst biedt studenten de mogelijkheid om zich breder te oriënteren en te bekwamen. Daarmee wordt de kwaliteit van het onderwijs bevorderd.

Praktisch gezien stuit het beoordelen op subkenmerkniveau op bezwaren. De methode vergt veel meer tijd dan de beoordeling op kenmerken. Mogelijke oplossingen hiervoor zijn:

- Alleen toetswerkstukken worden op subkenmerkniveau door de staf beoordeeld. Oefenwerkstukken worden door de studenten zelf beoordeeld.
- Alleen van kenmerken die als onvoldoende worden beoordeeld, worden de subkenmerken beoordeeld.

Onderzoek moet uitwijzen of hier sprake is van bruikbare alternatieven en welk alternatief de voorkeur geniet.

De toepassing van de besproken methode om de betrouwbaarheid van werkstukbeoordelingen te vergroten beperkt zich niet tot het tandheelkundig onderwijs. In principe is de methode bruikbaar voor alle vormen van onderwijs die gebruik maken van 'work-sample tests', om na te gaan of de aan te leren vaardigheden beheerst worden. Enkele voor de hand liggende voorbeelden zijn technische scholen en scholen voor huishoud- en nijverheidsonderwijs. Bij deze onderwijsvormen liggen de voorbeelden voor het opscheppen:

- Aan de hand van een opgetrokken muurtje wordt vastgesteld of een leerling de vaardigheid van het metselen onder de knie heeft.
- Een gestreken overhemd kan een indicatie zijn voor de mate waarin het omgaan met een strijkijzer beheerst wordt.
- Een saucijzebroodje kan uitsluitsel geven over de vraag of een leerling bladerdeeg kan bereiden.

Legio andere voorbeelden zijn te geven.

In dit artikel zijn twee manieren besproken om de betrouwbaarheid van werkstukbeoordelingen te bevorderen. In de eerste plaats het opstellen van een beoordelingsprotocol en in de tweede plaats het trainen van beoordelaars. Het opstellen van een beoordelingsprotocol lijkt over het algemeen haalbaar voor elk denkbaar werkstuk. De volgende, globale, aanwijzingen kunnen behulpzaam zijn bij het construeren van zo'n beoordelingsprotocol:

1. De doelstelling moet in zo veel mogelijk zinnige deeldoelen ontleed worden. 'Zinnig' betekent dat de deeldoelen betrekking moeten hebben op de kwaliteit van het werkstuk. Zo is het, bijvoorbeeld, niet van belang of bij een gestreken overhemd de linker mouw over de rechter gevouwen is, of andersom.
2. Elk deeldoel moet geoperationaliseerd worden. In zo eenduidig mogelijke terminologie dient elk deeldoel meetbaar te worden geformuleerd. Het beste gebeurt dit met kwantiteiten. Bijvoorbeeld door de onder- en bovengrens te bepalen van een afmeting, een volume, een gewicht, een hoek, een temperatuur, etc. Waar dit niet mogelijk is moet een zo nauwkeurig en objectief mogelijke beschrijving in kwalitatieve termen gegeven worden.
3. Voor elk geoperationaliseerd deeldoel moet aangegeven worden hoe vastgesteld kan worden of werkstukken aan de gestelde eisen voldoen. Als gebruik gemaakt kan worden van meetinstrumenten, dan verdient het aanbeveling om daarvoor de instrumenten te nemen die nodig zijn voor de vervaardiging van het werkstuk. Voor de leerling biedt dat de mogelijkheid van tussentijdse evaluatie en dus van tijdige bijsturing.
4. Per deeldoel moet een beoordelingsschaal geconstrueerd worden waarmee de kwaliteit van het werkstuk beschreven kan worden. Aan de score moet de maker van het werkstuk kunnen zien of voor het betreffende deeldoel aan de eisen voldaan is en zo niet, welke fout begaan is. Het aantal schaalpunten moet afhankelijk zijn van het aantal zinnige kwaliteitsonderscheidingen.
5. Vaak kan een illustratie, in de vorm van een schematische weergave, de omschrijvingen verduidelijken. Vooral bij het aangeven van de plaats waar 'gemeten' moet worden schieten woorden alleen dikwijls tekort.

In tegenstelling tot het opstellen van een beoordelingsprotocol, behoort het opzetten van een geïndividualiseerd trainingsprogramma lang niet altijd tot de mogelijkheden. In de eerste plaats wordt dit veroorzaakt door de noodzaak van een werkstukkenbestand. Geïndividualiseerde trainingen kunnen niet functioneren zonder de beschikking over werkstukken waarvan de kwaliteit bekend is in termen van de te hanteren beoordelingsmethode. Het opzetten van zo'n werkstukkenbestand is niet altijd mogelijk in verband met de beperkte houdbaarheid (bijvoorbeeld voedsel), de omvang (bijvoorbeeld metselwerk) en de hoge kostprijs (bijvoorbeeld bij edelsmeden) van sommige werkstukken. In de tweede plaats is een computer nodig om voor snelle terugkoppeling te kunnen zorgen en om de beoordelingsprestaties te administreren. Ten slotte moet opgemerkt worden dat het ontwikkelen van een geïndividualiseerd trainingsprogramma kostbaar is en dat niet elke doelstelling zo'n investering rechtvaardigt. Alleen centrale vaardigheden (vaardigheden met vele toepassingen) en vaardigheden die kostbare oefening vereisen (bijvoorbeeld het vergroten van foto's) komen hiervoor in aanmerking.

Noten

1. Het in was reconstrueren van een geheel of gedeeltelijk verloren gegaan gebitselement.
2. Prepareren betekent in de tandheelkunde dat voorbereidingen worden getroffen voor het leggen van de vulling (restauratie). Een belangrijk aspect van het prepareren is het boren van een caviteit. De plaats en de grootte van de aantasting bepalen de vorm van de te boren caviteit. De verschillende vormen worden aangeduid met klassen. Een klasse II preparatie wordt gemaakt als er sprake is van een aantasting in een wand die grenst aan een naburig element.
3. OU = Outline, ofwel de vorm van de caviteit. DI = diepte van de caviteit. Om breuken te voorkomen en om de vulling op de plaats te houden moeten de wanden in de caviteit een bepaalde hoek vormen ten opzichte van elkaar. De aspecten CA (caviteitoppervlakte hoek) en CO (convergentie/divergentie) hebben hierop betrekking. PA = pulpo-axiale afschuining. Dit aspect heeft betrekking op een bepaalde hoek in de caviteit die afgeschuind moet worden om breuk van de vulling te voorkomen. AF = afwerking.

Literatuur

Cohen, J., A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 1960, *20*, 37-46.

Hinkelman, K. W. & N. K. Long, Method for decreasing subjective evaluation in preclinical restorative dentistry. *Journal of Dental Education,* 1973, *37*, 13-18.

Houpt, M. I. & G. Kress, Accuracy of measurement of clinical performance in dentistry. *Journal of Dental Education,* 1973, *37*, 34-46.

Instituut Conserverende Tandheelkunde Voor Volwassenen, *Syllabus bij de geïndividualiseerde kursus Preparatie/Restauratie I en II.* Blok 155, Blok 255. Katholieke Universiteit, Nijmegen, 1981.

Mackenzie, R. S., Factors essential to evaluation of clinical performance. *Journal of Dental Education,* 1974, *38*, 214-223.

Otto, F. L., *Beoordelingsprocedures beoordeeld: Een systematische evaluatie van beoordelingsprocedures van de motorische onderwijsblokken eerste cursusjaar Tandheelkunde.* Rapport van de Subcommissie Toetsing en Beoordeling Motorische Vaardigheden, Katholieke Universiteit, Nijmegen, 1981.

Patridge, M. A. & T. A. Mast, Dental clinical evaluation: A review of the research. *Journal of Dental Education,* 1978, *42*, 300-305.

Steures, R. W. R. & Th. J. M. Tromp, Vernieuwing van een practicum voor tandheelkundige handvaardigheden (Deel I, II en III). *Nederlands Tijdschrift voor Tandheelkunde,* 1978, *85*, 421-426; 1980, *87*, 225-230, 258-264.

Straetmans, G. J. J. M., *Het onderwijsstimuleringsproject 'Een geïndividualiseerd trainingsprogramma voor het beoordelen van praktikumwerkstukken'.* Intern rapport CE 82-04, Katholieke Universiteit, Nijmegen, 1982.

Wiegman, J. E., *Assessment of dental skills using specific criteria.* Paper presented at the 1981 meeting of the Association for Dental Education in Europe, Groningen, 1982.

Curriculum vitae

G. J. J. M. Straetmans (1953) studeerde na zijn onderwijzersopleiding (1974) pedagogiek aan de Katholieke Universiteit te Nijmegen (afstudeerrichting Onderwijskunde in 1980). Sinds 1981 is hij als wetenschappelijk medewerker verbonden aan het Instituut Conserverende Tandheelkunde voor Volwassenen van genoemde universiteit.

Adres: Instituut Conserverende Tandheelkunde voor Volwassenen, Katholieke Universiteit, Postbus 9101, 6500 HB Nijmegen

Manuscript aanvaard 28-2-'84