apparaten die elders zijn verkregen.
– Slechts 6% van de hockeyers maakt consequent gebruik van een gebitsbeschermer.

Dit onderzoek werd ondersteund door het Nationaal Instituut voor de Sportgezondheidszorg (N.I.S.G.Z.).

Summary:

Title: Dental injuries in fieldhockey.

Keywords: Preventive dentistry – Traumatology – Mouthprotectors

Fieldhockeyplayers show an increasing interest in mouthprotectors. The possession of such a protector does not imply that it is frequently used. In a questionnaire sent to 3577 fieldhockeyplayers the question was posed: how many have one and how many wear it consequently. In addition, an estimation was made of the number of dental injuries that yearly occur during fieldhockey competition.

Literatuur:

1. *Groh H, Groh P.* Sportverletzungen und Sportschäden. München: Luitpold, 1975.
2. *Boersma-Slütter WGM, Broekman A, Lagro HAHM, Minderaa PH.* Sport, een riskante zaak? Een pilot-studie naar de incidentie van sportongevallen. Geneeskd en Sport 1979; 2:41.
3. *Stichting Consument en Veiligheid.* Jaarverslag 1984. Amsterdam: pors-Sport.
4. *Vanet R.* Gridirion challenge – can dentistry devise a mouthpiece that football-players will wear to prevent dental injuries? Dent Survey 1951; 27:1258.
5. *Bureau of Dental Health Education and Bureau of Economic Research and Statistics.* Mouth protectors: 1962 and the future. J Am Dent Assoc 1963; 66:539.
6. *Bureau of Dental Health Education.* Mouth protectors: a progress report. J Am Dent Assoc 1968; 77:632.
7. *Bureau of Dental Health Education and Council on Dental Materials and Devices.* Mouth protectors: 11 years later. J Am Dent Assoc 1973; 86:1365.
8. *Semmelink P, Bolhuis JHA.* Reactie op 'Tandvlees'. Hockeysport 1975; 21:4.
9. Schriftelijke opgave Koninklijke Nederlandse Hockeybond, 3 maart 1984.
10. *Kranenborg N.* Blessures bij micro-korfballers. Geneeskd en Sport 1981; 2:36–41.

Mei 1985.

Sorbonnelaan 16,
3584 CA Utrecht.

ONDERWIJS

DE BEOORDELINGSKWALITEIT BEOORDEELD

EEN STUDIE NAAR DE BETROUWBAARHEID VAN TWEE BEOORDELINGSMETHODEN VOOR DE KLASSE II-TWEEVLAKSPREPARATIE

G. J. J. M. STRAETMANS *Uit het Instituut Conserverende Tandheelkunde voor Volwassenen van de Katholieke Universiteit te Nijmegen.*

Trefwoorden: Beoordelingskwaliteit

Samenvatting:

De kwaliteit van het onderwijs is voor een belangrijk deel afhankelijk van de kwaliteit van de beoordeling. Zonder betrouwbare informatie over de vorderingen van de studenten kunnen geen goede beslissingen worden genomen over de voortgang van het onderwijsleerproces. In dit artikel wordt gerapporteerd over een studie die tot doel had om na te gaan of de betrouwbaarheid van een beoordelingsmethode voor preklinische werkstukken gebaat is bij gedetailleerde beoordelingsvoorschriften. De resultaten laten zien dat de inter-beoordelaarsovereenstemmingen toenemen als met een beoordelingsprotocol wordt gewerkt.

1. *Inleiding*

Het is redelijk om te veronderstellen dat het discrimineren tussen prestaties van studenten moeilijker wordt naarmate die prestaties dichter bij elkaar liggen. Zo zullen docenten doorgaans geen moeite hebben om aan te geven waarin een uitstekend werkstuk verschilt van een slecht werkstuk, maar wél om aan te geven wat het verschil is tussen een 'juist voldoende' en een 'juist onvoldoende' werkstuk. Zeker als het gaat om het onderscheid voldoende-onvoldoende is dit een gevoelig probleem, omdat de consequenties voor de maker van het werkstuk vaak zeer aanzienlijk zijn. Met name verdient dit probleem onder de aandacht te komen van opleiders, omdat prestaties van studenten vaak in dit grensgebied van voldoende en onvoldoende liggen. In het studiejaar 1982-1983, bijvoorbeeld, werd aan de Subfaculteit Tandheelkunde te Nijmegen 47 procent van alle klasse II-tweevlakspreparaties met een vijf of een zes (tienpuntsschaal) gewaardeerd.[1]
De problemen die inherent zijn aan het beoordelen van tandheelkundige werkstukken zijn meermaals in dit tijdschrift besproken.[2 3] Kort gezegd komt het erop neer dat het bijzonder lastig is om een beoordelingsprocedure te ontwikkelen waarmee tandheelkundige werkstukken op objectieve wijze te evalueren zijn. De betrouwbaarheid van die instrumenten is laag, hetgeen zichtbaar is in de lage inter- en intra-beoordelaarsovereenstemmingen die met dergelijke instrumenten worden bereikt.[4]
In dit artikel wordt nagegaan of de betrouwbaarheid van de beoordeling inderdaad gebaat is bij een meer gestandaardiseerde en geobjectiveerde beoordelingsprocedure.

2. *Een beoordelingsprotocol voor de klasse II-tweevlakspreparatie*

Binnen het Instituut Conserverende Tandheelkunde voor Volwassenen van de katholieke Universiteit te Nijmegen werd een protocol ontwikkeld voor het beoordelen van preklinische preparaties van het type klasse II. In een eerder verschenen artikel in dit tijdschrift werden de eigenschappen van dit instrument uitvoerig besproken.[4]

Het beoordelingsprotocol werd afgeleid van de beoordelingsmethode zoals die tot op heden in gebruik is in het preklinisch onderwijs en waarvan is vast komen te staan dat ze weinig betrouwbare metingen oplevert.[5] Bij deze *kenmerkmethode* wordt een klasse II-preparatie beoordeeld op zes *kenmerken:* outline, diepte, caviteit-oppervlaktehoek, convergentie/divergentie, pulpo-axiale afschuining, afwerking. Deze zijn dermate veelomvattend dat een objectieve analytische beoordeling niet mogelijk is: het blijft globaal. In het beoordelingsprotocol wordt elk kenmerk geoperationaliseerd door middel van *subkenmerken.* Voor in totaal 32 subkenmerken wordt daarin omschreven wat de eisen zijn waaraan het werkstuk dient te voldoen (prestatiecriterium), hoe vastgesteld dient te worden of het werkstuk aan de gestelde eisen voldoet en hoe de daaruit voortvloeiende waarneming in een score moet worden omgezet (scoringsregel). Door middel van een illustratie worden de omschrijvingen verder verduidelijkt.

3. *Vaststellen van de beoordelingskwaliteit*

Om een beter inzicht te krijgen in de resultaten zoals die in de volgende paragraaf aan de orde komen, wordt eerst kort ingegaan op het begrip 'beoordelingskwaliteit' en op methodes om die kwaliteit te meten. De kwaliteit van beoordelingsinstrumenten kan worden bestudeerd via de betrouwbaarheid. Het is gebruikelijk in beoordelingsstudies de betrouwbaarheid te schatten door de overeenstemming tussen beoordelaars te berekenen. Hoe groter de overeenstemming is hoe betrouwbaarder het instrument. Een veel gebruikte maat hiervoor is het percentage overeenstemming. De voordelen van deze maat zijn de eenvoudige toepassing en het gemak waarmee hij geïnterpreteerd kan worden. Een nadeel is echter dat deze maat geen onderscheid maakt tussen echte overeenstemming en kansovereenstemming. Een maat die het probleem van de kansovereenstemming oplost is coëfficiënt Kappa.[6] Kappa geeft de proportie overeenstemming weer, gecorrigeerd voor kansovereenstemming. De kansovereenstemming tussen twee beoordelaars wordt geschat uit hun beider scoreverdelingen. Kappa neemt meestal een waarde aan tussen 0 en 1, maar negatieve Kappa's zijn ook mogelijk. Als de kansovereenstemming even groot is als de waargenomen overeenstemming, dan is Kappa gelijk aan 0. Als alle waargenomen overeenstemming echte overeenstemming is, dat wil zeggen als er geen kansovereenstemming is, dan is Kappa gelijk aan 1. Negatieve Kappa's doen zich voor als de kansovereenstemming groter is dan de waargenomen overeenstemming. Een Kappa-coëfficiënt wordt getoetst op significant afwijken van 0 door de gevonden waarde te delen door de standaardfout. Toetsing vindt plaats tegen de standaard normale verdeling door middel van z-waarden.

Kappa is een overeenstemmingsmaat voor data die van een 'laag', dat wil zeggen nominaal of ordinaal meetniveau zijn. De scores die het resultaat zijn van de kenmerkmethode als ook de scores van de subkenmerkmethode zijn van een dergelijk laag meetniveau. De kenmerkmethode levert data van ordinaal niveau op (hoe hoger de score hoe beter de kwaliteit van het werkstuk), de subkenmerkmethode produceert nominale data (de scores verwijzen naar categorieën die de kwaliteit van het beoordeelde aspect beschrijven).

In de meeste onderzoeken van het type dat in deze bijdrage wordt beschreven gaat men uit van de op zich aanvechtbare veronderstelling dat bij de cijfers zoals die toegekend worden in het onderwijs sprake is van intervalniveau. Dit is het geval als het verschil tussen opeenvolgende cijfers constant is. De aanvechtbaarheid van de veronderstelling zit in het feit dat het verschil tussen een vijf en een zes (voldoende/onvoldoende) groter is dan tussen een acht en een negen. De assumptie maakt het evenwel mogelijk gebruik te maken van 'sterkere' statistische maten. De betrouwbaarheid van de beoordelingen kan bijvoorbeeld op een andere wijze worden vastgesteld. Veelal wordt dan gebruik gemaakt van associatiematen. Deze geven de mate van samenhang aan tussen twee of meer beoordelaars. We spreken van een perfecte associatie als op basis van de door beoordelaar x toegekende scores de door beoordelaar y toegekende scores voorspeld kunnen worden. Opgemerkt dient te worden dat een perfecte associatie niet automatisch impliceert dat de toegekende scores door beoordelaar x en beoordelaar y identiek zijn. Associatiematen geven de mate van samenhang aan tussen beoordelaars, maar niet noodzakelijkerwijs de mate van overeenstemming!

Als men de associatie tussen méér dan twee beoordelaars in een getal wil uitdrukken, dan is de Intraklasse Correlatie Coëfficiënt (ICC) een bruikbare maat. De ICC wordt geschat op basis van variantieschattingen afkomstig uit een variantieanalyse en geeft de proportie weer van de totale variantie in de beoordelingen die toegeschreven kan worden aan de variantie in de werkstukkenkwaliteit. Waarden die de bovengrens van de ICC (=1.00) benaderen, geven een hoge associatie aan tussen de variantie in de werkstukkenkwaliteit en de totale variantie en wijzen dus op een hoge betrouwbaarheid van de beoordelingen.

Er zijn verschillende versies van de ICC, die zeer uiteenlopende resultaten kunnen geven als ze op dezelfde data worden toegepast. Elke versie is geschikt voor een specifieke situatie afhankelijk van de onderzoeksopzet en het onderzoeksdoel. Shrout en Fleiss geven op inzichtelijke wijze aan hoe bepaald kan worden voor welke versie gekozen dient te worden.[7]

In het in de volgende paragraaf te bespreken onderzoek naar de kwaliteit van twee beoordelingsinstrumenten wordt gebruik gemaakt van de hierboven beschreven overeenstemmings- en associatiemaat.

Tabel I. Cijfers op basis van de kenmerk- (K) en subkenmerkmethode (SK).

werk-	beoordelaar A		B		C		D	
stuk	K	SK	K	SK	K	SK	K	SK
I	5	5	5	7	7	5	4	5
II	5	6	7	6	5	4	4	6
III	4	4	4	4	4	4	3	4
IV	5	5	7	7	4	4	4	4
V	5	5	5	5	8	5	6	5
VI	5	5	6	8	6	7	5	7

Tabel II. Variantiebronnen bij een tweeweg variantie-analyse (mixed model: werkstukkeneffect = random; beoordelaarseffect = fixed).

bron	kwadratensom	df	variantie	F
Kenmerkmethode				
werkstukken	11.38	5	2.28	
beoordelaars	7.80	3	2.60	2.52
error	15.46	15	1.03	
Subkenmerkmethode				
werkstukken	16.21	5	3.24	
beoordelaars	6.46	3	2.15	3.12
error	10.29	15	0.69	

4. *Kenmerk- versus subkenmerkmethode*

4.1. *Vraagstelling*

Een voor de hand liggende vraag is natuurlijk of met het beoordelingsprotocol (subkenmerkmethode) prestaties van studenten beter beoordeeld worden dan met de huidige beoordelingsmethode (kenmerkmethode). Concreter: zijn cijfers toegekend op grond van subkenmerkscores betrouwbaarder dan cijfers gebaseerd op kenmerkscores?

Aangezien een bepaald cijfer op geheel verschillende manieren tot stand kan komen is overeenstemming op cijferniveau geen garantie voor overeenstemming op itemniveau. En juist dit laatste is van zeer groot belang voor het onderwijs. Als studenten geen betrouwbare informatie krij-

gen over welke deelvaardigheden ze beheersen en welke niet, dan kunnen nodeloze herhaling en langdurig falen het gevolg zijn. De efficiëntie van het onderwijs is hier in het geding. De tweede vraag luidt derhalve: worden de diverse deelvaardigheden betrouwbaarder beoordeeld met de subkenmerk- dan met de kenmerkbeoordelingsmethode?

4.2. *Materiaal en methode*

Vier tandartsen uit het Instituut Conserverende Tandheelkunde voor Volwassenen van de Katholieke Universiteit te Nijmegen beoordeelden onafhankelijk van elkaar zes klasse II-tweevlakspreparaties in kunststof elementen. De eerste keer met gebruikmaking van de vigerende beoordelingsmethode (kenmerk-beoordelingsmethode), de tweede keer met behulp van de subkenmerk-beoordelingsmethode. Tussen de twee beoordelingsmomenten zat minimaal een maand tijd. De aangeboden preparaties waren op strikt toevallige wijze getrokken uit een werkstukkenbestand van klasse II-tweevlakspreparaties, dat speciaal voor beoordelingstrainingen is aangelegd.[1] De door eerstejaarsstudenten geprepareerde elementen waren met bijbehorende buurelementen opgesteld in kleine plastic bakjes. De werkstukken werden door elke tandarts in dezelfde volgorde beoordeeld met gebruikmaking van dezelfde tandheelkundige instrumenten en met een standaard lichtbron.
De tandartsen waren niet op de hoogte van de doelstelling van het onderzoek en wisten bij de eerste beoordelingsronde niet dat er nog een tweede zou volgen.

4.3. *Resultaten*

4.3.1. *Betrouwbaarheid van cijfers*
In tabel I staan de cijfers die het resultaat zijn van transformaties (naar een tienpunts-schaal) van de ruwe itemscores, zoals die zijn toegekend door de tandartsen aan de hand van de kenmerk- en de subkenmerk-beoordelingsmethode.
De resultaten van de tweeweg variantie-analyses, afzonderlijk toegepast op de cijfers gebaseerd op de kenmerk- en de subkenmerk-beoordelingsmethode, worden gepresenteerd in tabel II.
De varianties zijn gebruikt om de intraklasse correlatie-coëfficiënten te berekenen voor de cijfers gebaseerd op de twee beoordelingsmethodes. De versie van de ICC die in dit specifieke geval gebruikt dient te worden ziet er als volgt uit:

$$ICC = \frac{MS_w - MS_e}{MS_w + (k-1)MS_e}$$

waarbij MS_w staat voor het werkstukkeneffect, MS_e voor de foutenvariantie en k voor het aantal beoordelaars.[7] De ICC's voor de cijfers op basis van de kenmerk- en subkenmerk-beoordelingsmethode bedragen respectievelijk 0.23 en 0.48. Alleen de laatstgenoemde waarde is statistisch significant ($p<.01$).
De berekende ICC's schatten de betrouwbaarheid van één enkele beoordelaar. Naarmate er meer beoordelaars naar hetzelfde werkstuk kijken en het uiteindelijke oordeel tot stand komt door middeling van deze beoordelingen, stijgt de betrouwbaarheid. Een indicatie voor de betrouwbaarheid van de gemiddelde beoordeling wordt gegeven door:[7]

$$ICC = \frac{MS_w - MS_e}{MS_w}$$

Toegepast op de variantie-schattingen in tabel II resulteert deze formule in ICC's van 0.55 en 0.79 voor de cijfers berekend op grond van beoordelingen met de kenmerk- respectievelijk de subkenmerk-beoordelingsmethode.

4.3.2. *Betrouwbaarheid van itemscores*
Voor elke beoordelaar is berekend hoe groot de overeenstemming is die bereikt wordt met alle overige beoordelaars. De overeenstemming wordt uitgedrukt in Cohen's Kappa. De berekeningen werden uitgevoerd per beoordelingsaspect, zodat het mogelijk is om na te gaan of bepaalde onderdelen van de prestatie (deelvaardigheden) moeilijker te beoordelen zijn. Tabel III geeft de beoordelingsprestaties weer van elke docent op elk beoordelingsaspect. Per cel zijn er twee Kappa-coëfficiënten; een voor de overeenstemming behaald met de kenmerkmethode en een voor de overeenstemming behaald met de subkenmerkmethode. De schuingedrukte waarden in de cellen zijn z-waarden; ze zijn het resultaat van toetsingen van het verschil tussen de Kappa's op kenmerk- en subkenmerkniveau.
Een voorbeeld kan het lezen en interpreteren van tabel III verduidelijken. In de cel die gevormd wordt door rij 1 en kolom 1 staan drie waarden. De waarde 0 geeft aan dat beoordelaar A niet in staat is geweest om tot echte overeenstemming te komen met alle andere beoordelaars over de kwaliteit van de aangeboden werkstukken met betrekking tot het aspect 'outline'.
Alle overeenstemming met de andere beoordelaars moet aan het toeval worden toegeschreven. Beoordeling van dezelfde werkstukken op hetzelfde beoordelingsaspect en door dezelfde beoordelaars, maar aan de hand van de subkenmerk-beoordelingsmethode, leidde tot een redelijke overeenstemming tussen beoordelaar A en de overige beoordelaars (Kappa x 100=66). De z-waarde 8.07, tenslotte, is het resultaat van een toetsing van het verschil tussen de twee besproken Kappa's. Als deze z-waarde groter of gelijk is aan 1.96, dan is de kans kleiner dan vijf procent dat het verschil tussen de Kappa's een toevallig verschil is. De waarde 8.07 duidt dus op een zeer significant verschil in overeenstemming tussen de kenmerk- en subkenmerk-beoordelingsmethode voor wat betreft de beoordelingen van het aspect 'outline'.
Uit tabel III kan worden afgeleid dat voor alle beoordelaars geldt dat op bijna elk beoordelingsaspect de bereikte overeenstemming met de overige beoordelaars groter is als gebruik gemaakt wordt van de subkenmerk-beoordelingsmethode. Daar-

Tabel III. Inter-beoordelaarsovereenstemming (Kappa ×100) per beoordelingsaspect op kenmerk- (K) en subkenmerkniveau (SK) en toetsingsresultaten (z).

beoordelaar	outline		diepte		cav.opp.		conv.		p.a.a.		afwerking	
	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
A	0	66	5	22	19	52	38	55	−13	−15	0	34
		8.07)*		*0.88*		*1.78*		*0.85*		*−0.10*		*1.53*
B	37	66	− 4	17	14	38	22	38	−27	0	12	31
		1.45		*0.47*		*1.55*		*0.98*		*1.54*		*0.81*
C	39	64	10	22	36	29	38	36	−14	0	−24	31
		1.30		*0.87*		*−0.38*		*−0.15*		*0.78*		*2.24*)*
D	52	73	−23	35	25	37	32	49	4	−11	−24	46
		0.92		*3.67*)*		*0.68*		*0.87*		*−0.74*		*2.88*)*

*) = $p<.05$.

bij hoort echter wel de kanttekening dat in slechts drie gevallen sprake is van een significant grotere overeenstemming.

4.4. *Discussie*

Uit tabel II blijkt dat het gebruik van de subkenmerk-beoordelingsmethode de overeenstemming tussen beoordelaars bevordert. De variantie die toegeschreven wordt aan beoordelaars is afgenomen van 2.60 naar 2.15. De grotere overeenstemming tussen de beoordelaars heeft tot gevolg dat de verhouding tussen de variantie die toegeschreven wordt aan de verschillen tussen de werkstukken en de foutenvariantie, positiever is voor de subkenmerkmethode (3.24/0.69 = 4.70) dan voor de kenmerkmethode (2.28/1.03 = 2.21). Met andere woorden: de beoordelaars zijn met het beoordelingsprotocol beter in staat geweest om onderscheid te maken tussen de werkstukken dan aan de hand van de kenmerk-beoordelingsmethode. Dit wordt ook geïllustreerd met de ICC-waarden die afzonderlijk berekend werden over de cijfers op kenmerk- en subkenmerkniveau. De ICC op subkenmerkniveau is twee keer zo groot als die op kenmerkniveau. Toch kan ook op subkenmerkniveau niet gesproken worden van betrouwbare cijfers. Slechts 48 procent van de totale variantie in de beoordelingen kan worden toegeschreven aan de kwaliteitsverschillen tussen de werkstukken. Het is niet onwaarschijnlijk dat de vrij geringe variatie in de kwaliteit van de werkstukken de betrouwbaarheid gedrukt heeft.
De betrouwbaarheid van het gemiddelde oordeel op subkenmerkniveau is aanvaardbaar te noemen. Maar het feit dat deze aanvaardbare waarde pas bereikt wordt als vier beoordelaars worden ingeschakeld laat de betrekkelijkheid zien van dit gegeven. In het onderwijs zal het veelal niet haalbaar zijn om meer beoordelaars naar hetzelfde werkstuk te laten kijken.

De Kappa-coëfficiënten in tabel III geven aan dat de beoordelingen op itemniveau betrouwbaarder zijn als het item een subkenmerk is in plaats van een kenmerk. Voor de studenten betekent dit dat zij meer betrouwbare informatie krijgen over hun prestaties en daardoor beter in staat zullen zijn om passende maatregelen te treffen ten einde gesignaleerde tekortkomingen op te heffen. Het onderwijsleerproces kan zich daardoor efficiënter voltrekken.
Opvallend in tabel III zijn de zeer lage Kappa's met betrekking tot de beoordelingsaspecten 'diepte' en 'pulpo-axiale afschuining'. Dit geldt zowel voor de overeenstemmingen op kenmerkniveau als voor die op subkenmerkniveau. Nadere bestudering van berekeningen resulteerde in de constatering dat de variatie in de toegekende scores met betrekking tot deze beoordelingsaspecten zeer beperkt was. Het gevolg van deze beperkte scorevariatie is dat de scoreverdelingen van de beoordelaars in de kruistabellen grote overeenkomst vertonen. En dat leidt weer tot hoge kansovereenstemming en dus tot lage Kappa-coëfficiënten. De lage Kappa's met betrekking tot de aspecten 'pulpo-axiale afschuining' en 'diepte' zijn dus een gevolg van de definitie van kansovereenstemming in Cohen's Kappa en geen indicatie dat het beoordelingsprotocol met betrekking tot deze aspecten minder goed zou functioneren.

4.5. *Conclusie en aanbevelingen*

De resultaten in de tabellen I, II en III leiden tot de conclusie dat het beoordelingsprotocol de beoordelingskwaliteit van de klasse II-tweevlakspreparatie bevordert. De betrouwbaarheid van beslissingen over zakken of slagen en over de beheersing van deelvaardigheden is aanzienlijk groter als de subkenmerk-beoordelingsmethode wordt gebruikt in plaats van de kenmerk-beoordelingsmethode. De prijs die betaald moet worden voor de toename van de beoordelingskwaliteit is een sterke toename van de tijd die gemoeid is met het gebruik van het beoordelingsprotocol. Gemiddeld genomen kunnen zeven werkstukken worden beoordeeld met de kenmerkmethode tegen één met de subkenmerkmethode.[1] Het verdient daarom aanbeveling om te zoeken naar wegen om het beoordelingsprotocol in te voeren zonder dat dit betekent dat docenten geen tijd meer overhouden voor instructie. Eén van die manieren is het invoeren van zelf-evaluatie voor oefenwerkstukken. Dit heeft als extra voordeel dat studenten expliciet nota nemen van de prestatiecriteria. De operationeel gedefinieerde prestatiecriteria vormen een goede leidraad voor het verwervingsproces van de betreffende vaardigheid. Toetswerkstukken moeten door de staf beoordeeld blijven worden, ten einde het gevaar van zichzelf bevoordelende studenten te voorkomen. Door de omvangrijkheid van de beoordelingstaak kan de student echter niet meer rekenen op onmiddellijke uitslag. Na het verstrijken van de toetstijd moeten de werkstukken worden ingenomen en gecodeerd en vervolgens ter hand worden gesteld aan docenten of student-assistenten die voor de beoordeling zorgdragen. Deze gang van zaken heeft als bijkomend voordeel dat de beoordelaar tijdens het beoordelingsproces niet weet wiens werkstuk beoordeeld wordt. Ook dit levert een positieve bijdrage aan de objectiviteit van de prestatiemeting.

Summary:

Title: Judging the quality of judgements of cavity preparations.

Keywords: Quality of judgements

A major factor that influences the quality of education is the quality of the evaluation instruments with which student performances are assessed. Making decisions concerning the student's progress assumes the availability of reliable information about his strong and weak points.
This article presents evidence to support the need for very detailed performance criteria, assessment methods and scoring rules. Interrater agreements increased if judgements were made using a socalled 'raters protocol'.

Literatuur:

1. *Straetmans GJJM.* Evaluatie in het tandheelkundig onderwijs. Beoordelen van practicumwerkstukken en meten van probleemoplosvaardigheid. Nijmegen: Katholieke Universiteit, 1985. Academisch proefschrift.
2. *Penning Ch, Steures RWR, Thoden van Velzen SK, Tromp ThJM.* Een klinisch instructie- en beoordelingssysteem voor caviteitspreparatie en -restauratie. Ned Tijdschr Tandheelkd 1980; 87: 34-43 en 88-94.
3. *Wiegman JE, Van Groeningen G, Van de Poel ACM.* Het belang van duidelijke criteria bij de beoordeling van tandheelkundige werkstukken. Ned Tijdschr Tandheelkd 1985; 92: 72-5.
4. *Straetmans GJJM.* De ontwikkeling van een trainingsprogramma voor het beoordelen van preklinische tandheelkundige werkstukken. Ned Tijdschr Tandheelkd 1984; 91: 115-20.
5. *Sanders AJ.* Evaluatierapport blok 155 studiejaar 1977-1978. Intern Rapport, Katholieke Universiteit te Nijmegen, 1980.
6. *Cohen J.* A coefficient of agreement for nominal scales. Educ Psych Measurement 1960; 20: 37-46.
7. *Shrout PE, Fleiss JL.* Intraclass correlations: uses in assessing rater reliability. Psych Bull 1979; 86: 420-8.

Juni 1985. Adres: Dr. G. J. J. M. Straetmans, Louis van Gasterenstraat 196, 7558 SZ Hengelo.