

Summary:

Title: The condensation of amalgam in artificial cavities. A practical preclinical exercise.

A practical preclinical exercise for first year dental students in working with amalgam has been described. The exercise consisted of the condensation of amalgam in artificial cavities. The artificial cavity was made in a stainless steel plate, bounded on either side by two further steel plates capable of being screwed together. The student used a condensation technique of his own choice for this discovery learning exercise. After the model had been dismantled, the amalgam restoration in the middle plate was viewed using a stereo microscope. The student was presented with a photographic scale of re-

sults varying from good to poor. It was hoped that using the scale he would gain insight into his condensation method and the problems involved in the achievement of optimal condensation techniques necessary for his future clinical dentistry.

Literatuur:

1. Nicholls, J. I., Harding, D. C. (1977): Self-evaluation unit for teaching adequate amalgam condensation. *J Dent Educ* 41: 94-96.
2. Rookhuyzen, R. F. van, Plomp, Tj., Pilot, A. (1977): Individuele Studie Systemen in het tertiair onderwijs. Wolters Noordhof, Groningen.
3. Salvendy, G., Joost, M. G., Ferguson, G. W., Cunningham, P. R., Schiff, A., Wilko, R. A. (1976): Training in force exertions during condensation of dental amalgam. *J Dent Educ* 40: 370-372.

November 1980.

Ant. Deusinglaan 1
9713 AV Groningen.**EVALUATIE VAN MEERKEUZE-TOETSEN**

CH. PENNING

S. K. THODEN VAN VELZEN

Uit de vakgroep Cariologie, Endodontologie en Pedodontologie van de Universiteit van Amsterdam

Trefwoorden: Onderwijs – Cariologie – Meerkeuze-toetsen

Inleiding

Een van de problemen waarmee de opsteller van meerkeuze-toetsen wordt geconfronteerd, is de evaluatie van de onderscheiden vragen, in dit verband meestal items genoemd. Studenten klagen soms over onduidelijke, dubbelzinnige of zelfs onjuist geachte formuleringen en in een aantal gevallen blijken de klachten terecht te zijn. Het spreekt vanzelf dat de docent zich veel moeite moet geven om zijn items correct te formuleren. Maar ondanks een zorgvuldige voorbereiding kan na het afnemen van een toets blijken dat er onvolkomenheden zaten in sommige vragen. Het is om drie redenen van belang dergelijke onvolkomenheden op te sporen.

In de eerste plaats kunnen sommige studenten een onvoldoende hebben gescoord als gevolg van onjuist geformuleerde items. In de tweede plaats wil de docent de items op een later tijdstip wellicht weer gebruiken en tenslotte verschaffen de aan het licht getreden tekortkomingen nuttige feedback aan de itemschrijver.

De gangbare evaluatiemethode van meerkeuze-toetsen is de geautomatiseerde itemanalyse met behulp van een computerprogramma, dat een aantal scores oplevert die inzicht verschaffen in de kwaliteit van de toets als geheel en van de afzonderlijke

items. Men krijgt langs deze weg met geringe inspanning vrij veel informatie. Teneinde aanvullende informatie te verkrijgen van meer inhoudelijke aard kan men dan verder de toetsdeelnemers vragen om tijdens het afleggen van de toets schriftelijk commentaar op de afzonderlijke items te leveren. Het sorteren en analyseren van deze commentaren kost echter nogal wat tijd en de vraag moet worden gesteld of de verkregen informatie deze extra inspanning rechtvaardigt. Doel van het hier beschreven onderzoek was om de uitkomsten van de beide evaluatiemethoden te vergelijken en op grond daarvan vast te stellen of de extra informatie die met de tweede methode wordt verkregen van wezenlijke betekenis is.

De geautomatiseerde itemanalyse

Voor het scoren en evalueren van een afgelegde toets wordt binnen de Subfaculteit Tandheelkunde van de Universiteit van Amsterdam meestal gebruik gemaakt van een computerprogramma waarmee per item de drie hiernavolgende waarden worden berekend.

1. De zogenaamde p-waarde, d.i. het gedeelte toetsdeelnemers dat het goede alternatief heeft gekozen, m.a.w. het aantal deelnemers met het goede alternatief ge-

Samenvatting:

Kritische opmerkingen van toetsdeelnemers zouden hun nut kunnen hebben bij de evaluatie van een tentamen. Bijeen toets voor derdejaars studenten over cariëspathogenese en -etiologie werd hen gevraagd op een formulier hun op- en aanmerkingen te noteren. Nadat met geautomatiseerde toetsverwerking gegevens waren verkregen over de kwaliteit van de toetsvragen, werden deze gegevens gelegd naast de opmerkingen van de toetsdeelnemers, teneinde vast te stellen of de combinatie van beide evaluatiemethoden tot een beter resultaat leidt. Naar de mening van de auteurs moet het antwoord bevestigend luiden.

deeld door het totale aantal deelnemers. Betreft het foutieve antwoord-alternatieven dan spreekt men meestal van a-waarde. De p-waarde dient niet te laag te zijn; de bepaling van de ondergrens is arbitrair, het zal echter zonder meer duidelijk zijn dat als de p-waarde gelijk of lager wordt dan de raad-kans, het aantal 'raders' bij de goede antwoorden proportioneel sterk gaat toenemen hetgeen afbreuk doet aan de betrouwbaarheid van de toets.

2. De zogenaamde r_{it} en r_{at} , deze zijn een maat voor de correlatie tussen elk der antwoord-alternatieven van het item en de score die is behaald voor de toets als geheel door hen die het bewuste alternatief hebben gekozen. De r_{it} geeft de correlatiewaarde van het goede item-alternatief; de r_{at} 's geven de waarden van elk der foutieve alternatieven. Het is gewenst dat de r_{it} positief uitvalt en de r_{at} 's negatief, m.a.w. een vraag wordt geacht goed te functioneren als de goede deelnemers hem goed en de

- De vijf lagen, die men bij cariës in het tandbeen kan onderscheiden, kunnen opnoemen en beschrijven; drie factoren kunnen noemen die mede bepalend zijn voor dikte en vorm van deze lagen.
- Kunnen beschrijven om welke twee redenen het veelal niet mogelijk is om bij excavatie alle bacteriën te elimineren.
- De twee lagen kunnen opnoemen die bij carieuze aantasting van elementen met een avitale pulpa ontbreken.
- De reacties in het dentine t.g.v. het cariësproces kunnen beschrijven.

Afb. 1. Enkele voorbeelden van leerdoelen op detail-niveau.

slecht voorbereide deelnemers hem fout beantwoorden.

3. De zogenaamde kleine *f* van Van Naerssen (1976); deze *f*-waarde geeft aan in welke mate het bewuste item heeft bijgedragen aan het discriminerend vermogen van het tentamen. De *f*-waarde wordt afgeleid van een waarde die op haar beurt wordt gebruikt voor het beoordelen van de betrouwbaarheid van het tentamen als geheel. Het betreft de zogenaamde signaalruisverhouding of anders gezegd de verhouding tussen de reële waarde en de ingesloten meetfout, onder andere ontstaan als gevolg van het raden van het goede antwoord of het niet invullen van het toetsformulier.

Tabel I. Overzicht van de items die op grond van een onbevredigende score voor een nadere beschouwing in aanmerking kwamen of die de toetsdeelnemers aanleiding hadden gegeven tot commentaar.

| Item nr. | p-waarde <.70 | r_{it} <0 | f-waarde <0 | aantal commentaren van toets- deelnemers | aantal mede- werkers dat commentaar valide en van invloed acht |
|----------|------------------|----------------|----------------|---|--|
| 2 | | | -.03 | 1 | - |
| 3 | .42 | | -.07 | 2 | 2 |
| 4 | | | | 3 | 1 |
| 5 | | | | 1 | - |
| 6 | | | | 1 | 1 |
| 7 | | | | 1 | - |
| 10 | | | | 1 | 1 |
| 13 | .69 | | | 5 | - |
| 14 | | | | 1 | - |
| 17 | | -.12 | -.16 | 3 | 2 |
| 19 | | | | 1 | 2 |
| 20 | | | | 1 | 2 |
| 22 | | | | | |
| 23 | | | | 2 | - |
| 24 | | | | 1 | - |
| 25 | | | -.06 | - | - |
| 26 | 50 | | -.13 | 12 | 3 |
| 30 | | | | 2 | 1 |
| 31 | | -.18 | -.28 | 2 | - |
| 32 | | | | 1 | - |
| 34 | | | | 2 | 1 |
| 36 | | | -.14 | 1 | 1 |
| 37 | | -.01 | -.07 | - | - |

- de bijbehorende leerstof is onvoldoende bestudeerd;
- de bijbehorende leerstof bevat onduidelijkheden.

Vindt men nu een lage *p*-waarde gepaard gaande met een positieve r_{it} dan is de conclusie dat de goede toetsdeelnemers de vraag goed hebben beantwoord en dat derhalve de vraag en de bijbehorende leerstof moeilijk zijn. Vindt men echter een lage *p*-waarde in combinatie met een negatieve r_{it} dan moet men concluderen dat de vraag slecht is geformuleerd of betrekking heeft op een onduidelijk gedeelte van de leerstof.

Het verzamelen en beoordelen van het studententcommentaar

Bij een toets in het derde studiejaar werd tegelijk met het vragenformulier en het antwoordformulier nog een derde formulier uitgereikt, waarop de studenten bezwaren kenbaar konden maken tegen de formulering van de vragen. De toets bestond uit 37 tweekeuze-items^{*)}. De leerstof besloeg in totaal \pm 50 pagina's tekst uit drie leerboeken, waarvan twee in het Nederlands en één in de Engelse taal. De studenten hadden bij hun studie bovendien de beschikking over door de vakgroep geformuleerde leerdoelen op detail-niveau (afb. 1).

Van de 78 deelnemende studenten leverden er 27 een ingevuld commentaarformulier in. Nadat de commentaren waren gerangschikt volgens itemnummer bleek dat van de 37 items er 21 aanleiding hadden gegeven tot in totaal 43 commentaren, waarvan per item de strekking meestal gelijk was.

Om een zo objectief mogelijk oordeel te krijgen over de validiteit van de commentaren werd aan drie medewerkers van de vakgroep, die op geen enkele wijze betrokken waren geweest bij de voorbereiding van de toets, gevraagd om de commentaren in te delen in één der volgende categorieën:

- + Bezwaar valide, wellicht van invloed op de mogelijkheid het juiste antwoord te vinden.
- Bezwaar niet valide.

^{*)} Bij de analyse van een aantal meerkeuze-tentamens (Tromp, 1975) met vier antwoord-alternatieven bleek dat bij 57% van de vragen 2 van de 4 alternatieven elk door ten hoogste twee toetsdeelnemers waren gekozen. Kennelijk was het de itemschrijvers niet gelukt om vier gelijkwaardige antwoord-alternatieven aan te bieden en functioneerde het item feitelijk als een tweekeuzevraag. Op grond van deze waarneming wordt sindsdien meestal gebruik gemaakt van twee-keuzetoetsen.

- Bezwaar valide, echter *niet* van invloed op de mogelijkheid het juiste antwoord te vinden.
- ± Bezwaar valide, weet echter niet of dit bezwaar van invloed is op de mogelijkheid het juiste antwoord te vinden.

De genoemde medewerkers konden behalve over de commentaren beschikken over een exemplaar van het vragenformulier en over de drie genoemde leerboeken. De resultaten zijn vermeld in tabel I. Het bleek dat de commentaren steeds het karakter hadden van een bezwaar tegen het stellen van de bewuste vraag. Tegen 19 items werd bezwaar gemaakt, evenals tegen het in dezelfde toets combineren van twee andere items omdat de samenhang zodanig zou zijn dat men of beide fout of beide goed zou beantwoorden. Van deze 20 bezwaren werden er 9 niet valide of niet van invloed geacht. Van de resterende 11 werd één door 3 medewerkers, vier door 2 medewerkers en zes door slechts 1 medewerker valide en van invloed beoordeeld.

Het resultaat van de vergelijking van de beide evaluatiemethoden

Tabel I geeft een overzicht van de items die op grond van een onbevredigende p-waarde, danwel een negatieve r_{it} of f-waarde voor een nadere beschouwing in aanmerking kwamen of die de toetsdeelnemers aanleiding hadden gegeven tot commentaar, dat als een bezwaar tegen het stellen van de bewuste vraag kon worden opgevat. Het bleek dat de drie items met een lage p-waarde aan 2, 5 en 12 deelnemers aanleiding gaven tot bezwaar, dat door respectievelijk 2, 0 en 3 medewerkers werd gedeeld. Twee van de drie items hadden een negatieve f-waarde; de r_{it} was voor alle drie positief.

Drie items toonden een negatieve r_{it} , die in alle drie de gevallen gepaard ging aan een negatieve f-waarde. Deze drie items ontlokten 3, 2 en 0 deelnemers commentaar; de commentaren werden door 2, respectievelijk geen der medewerkers valide én van invloed geacht.

Van de overige items van de toets toonden drie een negatieve f-waarde; twee van deze gaven aan steeds één deelnemer aanleiding tot commentaar, dat in één geval werd bevestigd door één der medewerkers. Verder werd tegen nog 12 afzonderlijke en één combinatie van 2 items bezwaar aangetekend door 1-3 toetsdeelnemers; deze 13 bezwaren werden 7 keer gedeeld door 1 of 2 medewerkers.

In afbeelding 2 zijn de teksten en commentaren vermeld van de 7 items die óf aanleiding gaven tot bezwaren die door tenminste 2 medewerkers werden gedeeld

| | |
|------------|--|
| Item 3 | Het toepassen van corticosteroïden in de tandheelkunde kan als gevolg hebben: |
| Commentaar | <ol style="list-style-type: none"> 1. stimulatie van de periapicale botgenezing; 2. het ontstaan van een apicaal abces. <ol style="list-style-type: none"> 1. Staat volgens mij niet in de tekst. 2. Stof is niet behandeld in de opgegeven hoofdstukken. |
| Item 17 | Wanneer men kiemvrije ratten een cariogeen dieet per maagsonde toedient: |
| Commentaar | <ol style="list-style-type: none"> 1. kan men een correlatie tussen voeding en cariës aantonen; 2. neemt de cariësfrequentie niet toe. <ol style="list-style-type: none"> 1. Kiemvrije ratten hebben geen cariës → kan dus nooit toenemen dus nooit een correlatie tussen voeding + cariës. 2. Ad 1: men kan dit afleiden na vergelijk met orale toediening. Ad 2: de cariësfrequentie neemt af. 3. Beide antwoorden zijn mogelijk. Bij zulk een dieet per maagsonde zal de cariësfrequentie inderdaad niet toenemen zodat er een correlatie moet bestaan tussen voeding en cariës. |
| Item 19 | Dextraan kan worden gevormd uit: |
| Commentaar | <ol style="list-style-type: none"> 1. saccharose; 2. lactose. <ol style="list-style-type: none"> 1. Dextraan kan uit zowel saccharose als uit lactose gevormd worden, beide alternatieven zijn juist. |
| Item 20 | Primaire dentinecariës is: |
| Commentaar | <ol style="list-style-type: none"> 1. het beginnende cariësproces in het dentine, nadat het proces juist door het glazuur is 'heengebroken'; 2. het cariësproces van het dentine dat niet door het glazuur was bedekt. |
| Item 22 | Primaire dentinecariës manifesteert zich: |
| Commentaar | <ol style="list-style-type: none"> 1. in de driehoekige ruimten apicaal van de contactpunten; 2. op de occlusale vlakken van postcaniene elementen. <p>Vragen afhankelijk van elkaar → 20 fout → 22 fout.</p> |
| Item 26 | Indien de DMFS-index wordt gebruikt in plaats van de DMFT-index dan wordt daarmee: |
| Commentaar | <ol style="list-style-type: none"> 1. de meetgevoeligheid verhoogd ten koste van de betrouwbaarheid; 2. de reproduceerbaarheid van de testresultaten vergroot. <ol style="list-style-type: none"> 1. Niet behandeld. 2. Staat volgens mij niet in de tekst. 3. In de opgegeven leerstof werd niets vermeld over DMFS en DMFT-index. 4. Niet behandeld in de opgegeven stof. 5. Wat betekent meetgevoeligheid? 6. Niet in tekst gezien. 7. DMFS-index, DMFT-index niet genoemd. 8. Het antwoord op deze vraag heb ik niet in de tekst aangetroffen. 9. DMFS DMFT niet bekend. 10. Stond niet in dictaat. 11. Worden niet besproken in de opgegeven literatuur! 12. Staat dit in de te leren stof? |
| Item 31 | Cariëuze demineralisatie van glazuur betreft in eerste instantie: |
| Commentaar | <ol style="list-style-type: none"> 1. de kernen der prismata; 2. de wanden der prismata. <p>- Niet behandeld. - Wanden der kristallieten in plaats van prismata.</p> |
| Item 37 | Een gebied van carieus dentine, dat relatief groot is in verhouding tot de bijbehorende glazuurlesie, wordt aangetroffen bij: |
| Commentaar | <ol style="list-style-type: none"> 1. fissuurcariës; 2. cariës van de gladde vlakken. <p>Geen</p> |

Afb. 2. De tekst van de items 3, 17, 19, 20, 22, 26, 31 en 37, met de daarop betrekking hebbende studentcommentaren.

óf waarvan twee van de drie waarden onbevredigend waren. In drie gevallen zijn beide criteria van toepassing. Bij de afzonderlijke items valt het volgende op te merken.

Item 3

Dat het juiste alternatief inderdaad niet in de opgegeven leerstof was te vinden wordt duidelijk weerspiegeld in de p-waarde (0.423) die zelfs lager is dan de raadkans (0.5).

Item 17

Hier gaat een tamelijk hoge p-waarde (0.846) gepaard aan een negatieve r_{it} (-0.123). Onder degenen die het onjuiste alternatief hebben gekozen zijn dus blijkbaar meer goede dan slechte studenten. De vraag moet dus inderdaad verwarring hebben veroorzaakt, zoals vooral uit het derde commentaar op deze vraag blijkt.

Item 19

Dextraan kan zowel uit saccharose als uit lactose worden gevormd; beide alternatieven zijn derhalve juist. Het is opvallend dat slechts één deelnemer bezwaar maakte. Alle deelnemers zonder uitzondering hebben echter gekozen voor het door de itemschrijvers bedoelde antwoord-alternatief. De f-waarde is dan ook ± 0 in overeenstemming met het ontbreken van het discriminerend vermogen.

Items 20/22

De afhankelijkheid van beide items heeft niet geleid tot opvallend hoge of lage scores. Er kan aan worden getwijfeld of deze

afhankelijkheid invloed heeft gehad op de keuze van het alternatief.

Items 26

Evenals bij item 3 werd dit item niet gedekt door de opgegeven leerstof, hetgeen duidelijk tot uiting komt in de p-waarde (0.50), die hier precies gelijk is aan de raadkans.

Item 31

Hoewel de medewerkers op grond van de tekst van de desbetreffende leerstof het studentenbezwaar niet valide achtten, moet toch worden geconstateerd dat het tweede alternatief minder gelukkig is geformuleerd (wanden der prismata i.p.v. interprismatische substantie) en aanleiding heeft gegeven, zoals uit een der commentaren blijkt, tot verwarring tussen de glazuurprismata en de prisma-kristallieten. De zeer lage f-waarde en r_{it} zijn hiervan waarschijnlijk het gevolg geweest.

Item 37

In wezen zijn beide antwoorden goed, al is alternatief 1 duidelijk het betere. Slechts 7 toetsdeelnemers kozen voor alternatief 2; onder hen moeten zich enkelen hebben bevonden met een hoge score voor de toets als geheel, getuige de negatieve item-waarden.

Conclusie

Het blijkt dat elk der evaluatiemethoden gegevens biedt die met de andere methode niet werden verkregen. Voorts is het deelnemerscommentaar een hulp gebleken bij

de interpretatie van de onbevredigende item-waarden. Het lijkt er op dat de extra inspanning die het verzamelen en analyseren van de commentaren van de toetsdeelnemers kost ruimschoots wordt goedgemaakt door de voordelen die deze enquête biedt bij de analyse van de toets en de afzonderlijke items.

Summary:

Title: Evaluation of multiple choice items

Critical remark of test participants could well be useful when evaluating a test. Third year students who did a test about caries pathogenesis and caries aetiology were asked to give their criticisms and other remarks on a form provided for that purpose. After computer-assisted analysis of the quality of the test-items, the results were compared with the remarks of test-participants. The aim of the comparison was to determine whether putting both evaluations methods together would lead to a better result. The authors feel that the answer is in the affirmative.

Literatuur:

1. Van Naerssen, R. F. (1967): Itemsselectie bij studietoetsen. Ned Tijdschr Psychologie 22: 345-359.
2. Tromp, Th. J. M. (1975) Intern rapport.

December 1980.

Louwesweg 1,
1066 EA Amsterdam.