

Het effect van een observatorentraining

M.T.A. Boumans, A.J.J.A. Scherpbier, A. van Ooy, C.P.M. van der Vleuten, R.J.J. Hoogenboom, L.W.T. Schuwirth

Samenvatting

Het is gebruikelijk dat observatoren voor stationsexamens getraind worden ten einde de interobservervariantie te verminderen en dus de interbeoordelaarsbetrouwbaarheid te vergroten. Uit onderzoek is gebleken dat training bij het gebruik van gedetailleerde beoordelingslijsten geen effect heeft op de interobservervariantie.

De vraagstelling van dit onderzoek was of training van observatoren wel effect heeft bij het gebruik van globale beoordelingslijsten. Hiertoe werd een op video opgenomen station gescoord door een controlegroep van tien observatoren en een experimentele groep van tien observatoren. De experimentele groep kreeg vervolgens een training, bestaande uit feedback en discussie met experts. Twee weken na de training scoorden alle observatoren op nieuw dezelfde videoband.

Uit de resultaten blijkt dat de training weliswaar enig effect had op de interbeoordelaarsbetrouwbaarheid, maar dat dit effect te gering was om er vergaande conclusies aan te kunnen verbinden. Bij een aantal observatoren bleek de training een averechts effect te hebben.

Inleiding

Sinds het begin van de jaren tachtig is het stationsexamen of Objective Structured Clinical Examination (OSCE) een veel toegepaste examenvorm om klinische competentie te meten. Deze examenvorm is ontworpen om een betrouwbaarder beoordeling van studenten te kunnen geven dan bij traditionele examens doorgaans het geval is. Tijdens een OSCE

moeten alle studenten dezelfde opdrachten (kennis en vaardigheden) uitvoeren, waarbij ze geobserveerd worden door inhouddeskundigen.¹ Inmiddels is er veel onderzoek verricht naar deze examenvorm.

De betrouwbaarheid van het stationsexamen kan beïnvloed worden door een aantal factoren waaronder verschillen tussen observatoren. Door het aanbrengen van enige structuur in beoordelingsprocedures kunnen aanzienlijke verbeteringen worden bereikt.²⁻³ In het algemeen is de interbeoordelaarsbetrouwbaarheid groter bij gebruik van een gedetailleerde beoordelingslijst (checklist) dan bij een globale beoordelingslijst (rating scales).²⁻⁴ De resultaten van een recent onderzoek van Cunningham, Neville en Norman gaven echter een vergelijkbare interbeoordelaarsbetrouwbaarheid te zien voor beide typen beoordelingslijsten.⁴

De gedetailleerde lijsten hebben meestal de vorm van een checklist, waarbij een beoordelaar (observer) op een twee- of driepuntsschaal aangeeft of een student een bepaalde vaardigheid goed, fout of niet heeft verricht.² Bij de globale lijst wordt gebruik gemaakt van een kwalitatief oordeel van de observator op een vijf-, zes- of zevenpuntsschaal.⁵ Observatoren zijn doorgaans stafleden die afkomstig zijn uit de disciplines waarover wordt getoetst.² Observatoren geven over het algemeen de voorkeur aan globale lijsten. Bij deze lijsten hebben ze meer vrijheid en kunnen ze meer van hun expertise kwijt.^{2,5} Met betrekking tot de betrouwbaarheid van de toets als geheel blijkt de interbeoordelaarsbetrouwbaarheid geen belangrijke verstoringe variabele te zijn; dit geldt zowel voor gedetailleerde lijsten als voor globale lijsten.² Uit eerder onderzoek

M.T.A. Boumans, A.J.J.A. Scherpbier, A. van Ooy, C.P.M. van der Vleuten, R.J.J. Hoogenboom, L.W.T. Schuwirth

komt naar voren dat de variatie in stations (qua aantal, vorm en inhoud) als voornaamste bron van meefouten bij OSCE's beschouwd kan worden.^{2,3,6} Gedurende de opleiding tot basisarts aan de Universiteit Maastricht worden studenten in het eerste, tweede, derde, vierde en zesde studiejaar onderworpen aan een vaardigheidstoets ofwel OSCE. Bij deze toetsen worden de vaardigheden meestal beoordeeld met gedetailleerde beoordelingslijsten, maar vooral in het zesde studiejaar ook met globale beoordelingslijsten. Observatoren worden in Maastricht op hun taak tijdens de vaardigheidstoets voorbereid door een training.² De doelstelling van de training voor de vaardigheidstoets is het bewerkstelligen van uniformiteit bij de scoring van de beoordelingslijsten om zodoende de interbeoordelaarsbetrouwbaarheid te optimaliseren.⁶ Deze training wordt over het algemeen slecht bezocht om verschillende redenen, waaronder een matige kwaliteit van de gegeven training en het feit dat beoordelaars het niet nodig vinden om over de eigen vakspecifieke deskundigheid instructies te krijgen.²

Uit wetenschappelijk onderzoek op medisch onderwijskundig terrein blijkt echter dat training geen significant positief effect heeft op de interobservervariantie bij gebruik van gedetailleerde lijsten. Hierbij wordt uitgegaan van observatoren die een zekere mate van inhouddeskundigheid bezitten. Wel lijkt de betrouwbaarheid van de toetsobservaties verbeterd te worden door zorgvuldige selectie van inhouddeskundige observatoren die consistent scoren; ervaring alleen is hierbij niet genoeg.⁶⁻¹⁰

Eerder uitgevoerd onderzoek naar trainingseffecten had uitsluitend betrekking op gedetailleerde lijsten. Het effect van training bij gebruik van globale lijsten is nog niet onderzocht. De vraagstelling van het hier beschreven onderzoek was of training van observatoren bij gebruik van *globale* lijsten een positief effect heeft op de interobservervariantie. Cunningham et al. veronderstelt

(ook) dat training wellicht een groter effect heeft bij gebruik van globale lijsten dan bij gedetailleerde lijsten.⁴ Een positief trainingseffect zou bovendien de training voor observatoren zinvoller en daardoor wellicht aantrekkelijker kunnen maken.

Methode

Proefpersonen

De studie werd uitgevoerd aan de medische faculteit van de Universiteit Maastricht. De observatoren werden geselecteerd uit de vakgroepen Heelkunde, Orthopedie, Anatomie, Skillslab, Fysiologie en Huisartsgeneeskunde. De inclusiecriteria bij deze selectie waren inhouddeskundigheid en bekendheid met het observeren bij het station bewegingsapparaat in de vaardigheidstoetsen. Twintig van de 56 geselecteerde observatoren bleken inzetbaar als proefpersoon in de periode waarin het onderzoek werd uitgevoerd. De overige observatoren waren veelal bereidwillig, maar konden niet meedoen omdat het tijdstip van het onderzoek samenviel met patiëntenzorgactiviteiten.

Instrument

Een globale beoordelingslijst werd geselecteerd die gebruikt was als examenlijst voor een zesdejaarsstoets in het academisch jaar 1996/1997. Deze lijst hoort bij het station enkel/voetklachten en is gebaseerd op het onderwijs met betrekking tot het houdings- en bewegingsapparaat.¹¹ Er werd gekozen voor een lijst met uitsluitend medisch-technische aspecten, aanzien met name de validiteit en betrouwbaarheid van scoring met betrekking tot dit onderdeel te wensen over laten.^{7,10}

Figuur 1 toont de globale beoordelingslijst met de bijbehorende opdracht aan de student en ter illustratie de observatorinstructie voor het onderdeel 'passief bewegingsonderzoek'. Zoals in eerder onderzoek werd ook hier een

INFORMATIE VOOR DE STUDENT:

Patient (e), 25 jaar, bezoekt het spreekuur vanwege pijn ter hoogte van de rechtervoet. De klachten bestaan sinds een week. De vraag is: "Kunt U mij hiervan helpen?" Het beroep is administratief medewerks(st)er.

OPDRACHTEN:

- I. Neem een korte relevante anamnese af.
- II. Verricht een relevant fysisch diagnostisch onderzoek en benoem wat u doet.
- III. Hoe luidt uw differentiaaldiagnose en wat is uw meest waarschijnlijke diagnose?
Wat is uw verdere therapievoorstel en evt. aanvullende diagnostiek?

OBSERVATORINSTRUCTIE:

Deze observatorinstructie geeft een globale beschrijving van de handelingen die de student dient uit te voeren. In deze crierijlijst wordt gebruik gemaakt van een 6-puntsschaal. U dient hiervan te beoordelen of de student dit goed, voldoende, matig, onvoldoende, op slechte wijze dan wel niet doet. U vindt aanvullende informatie in deze observatorinstructie over de itemnummers die in de crierijlijst met een * gemarkeerd zijn.

De onderstaande items zijn omschreven voor de score goed.

PASSIEF BEWEGINGSONDERZOEK:

- Item 34: Minimaal de volgende bewegingen dienen passief gecontroleerd te worden:
- dorsale en plantaire flexie
 - inversie en eversie
 - pro- en supinatie
 - waarbij de goede handvatting gekozen wordt d.w.z.:
 - voor dorsale en plantaire flexie: één hand fixeert het onderbeen distaal en de andere hand beweegt de middenvoet.
 - voor inversie en eversie: één hand fixeert het onderbeen en de andere hand beweegt de middenvoet.
 - voor pro- en supinatie: één hand fixeert de calcaneus en de andere hand beweegt de middenvoet.

Item 35: De juiste conclusie dient te zijn (rechter enkel/voet):

- plantaire flexie pijnlijk in de eindstand

Figuur 1. Opdracht en beoordelingsvoorschriften station enkel/voetklachten.

video-opname gebruikt.^{6 7 9 10} Deze video-opname toont een quasi-realistische toetsstation-setting: een basisarts analyseert de enkel/voetklachten bij een vooraf geïnstrueerde proefpersoon zoals bij een normaal toetsstation gebruikelijk is.

Procedure

Voorafgaand aan het experiment werd deze videoband door drie experts, een orthopedisch chirurg, een anatoom en een huisarts, onafhankelijk van elkaar gescoord. Op basis van consensus werd hierna een ideaal gescoorde globale lijst opgesteld, welke als uitgangspunt zou dienen voor de training van de experimentele

groep én als gouden standaard bij de statistische analyse.

De twintig observatoren werden at random ingedeeld in een experimentele en een controlegroep, waarbij rekening gehouden werd met hun voorkeursdata. Uitgangspunt hierbij was dat observatoren van verschillende disciplines - orthopeden, huisartsen, anatomen, fysiologen en Skilslabmedewerkers - zo gelijk mogelijk over beide groepen verdeeld waren. Voorafgaand aan het onderzoek ontvingen beide groepen de globale lijst met de vraag deze door te nemen. Alle observatoren bekeken vervolgens in groepsverband de videoband en gaven individueel hun beoordeling. Aansluitend namen alleen de observatoren van

	nr.	goed	vol- doende	matig	onvol- doende	slecht	niet gedaan
OPDRACHT I: Relevante anamnese							
- Essentiële vragen	27	0	0	0	0	0	0
- Aanvullende vragen	28	0	0	0	0	0	0

OPDRACHT II: Relevante fysische diagnostiek van de enkels/voeten

I Inspectie							
- Botten en gewrichten	29	0	0	0	0	0	0
- Weke delen	30	0	0	0	0	0	0
- Looppatroon	31	0	0	0	0	0	0
II Actief bewegingsonderzoek							
- laat bewegingen door de patiënt uitvoeren	32	0	0	0	0	0	0
- komt tot de juiste conclusie	33	0	0	0	0	0	0
III Passief bewegingsonderzoek							
- voet passief bewegingen bij de patiënt uit	34	0	0	0	0	0	0
- komt tot de juiste conclusie	35	0	0	0	0	0	0
IV Spier testen							
- test de spiergroepen isometrisch	36	0	0	0	0	0	0
- test spieren selectief	37	0	0	0	0	0	0
- komt tot de juiste conclusie	38	0	0	0	0	0	0
V Palpatie							
- palpeert botten en gewrichten	39	0	0	0	0	0	0
- palpeert weke delen	40	0	0	0	0	0	0
- komt tot de juiste conclusie	41	0	0	0	0	0	0
VI Specifieke testen							
- voert specifieke testen uit	42	0	0	0	0	0	0
- komt tot de juiste conclusie	43	0	0	0	0	0	0

OPDRACHT III:

- Wat is uw differentiaaldiagnose?	44	0	0	0	0	0	0
- Wat is op grond van deze bevindingen de meest waarschijnlijke diagnose?	45	0	0	0	0	0	0
- Wat is uw therapievoorstel en evt. aanvullende diagnostiek?	46	0	0	0	0	0	0

Figuur 1. Vervolg: Globale beoordelingslijst station enkel/voetklachten.

de experimentele groep deel aan een training. Deze training bestond uit feedback op de beoordeling van de band door twee van de drie experts (de orthoped en de huisarts) aan de hand van de consensuslijst; de scores werden met elkaar vergeleken en indien nodig door de experts toegeelicht of bediscussieerd.

Na een interval van twee weken werd dezelfde videoband door beide groepen opnieuw bekeken en gescoord. Mocht er na de data-analyse een duidelijk positief effect meetbaar zijn, dan zou na twee maanden een reëntentemeting plaatsvinden.

Analyse

Bij de analyse van de data werd gebruik gemaakt van Cohens kappa.¹² De mate van overeenstemming tussen de observatoren onderling en de mate van overeenstemming met het expertoordeel werden apart berekend. De formule voor kappa is:

$$\text{kappa} = \frac{(p^0 - p^2)}{(1 - p^2)}$$

Hierbij is p⁰ de proportie geobserveerde overeenstemming en p² de proportie overeenkomst

Tabel 1. Kappa totaal en Pearson-correlatie voor controlegroep en experimentele groep van meetmoment 1 en meetmoment 2. IBB = interbeoordelaarsbetrouwbaarheid van observatoren onderling; IBE = interbeoordelaarsbetrouwbaarheid van observatoren ten opzichte van het expertoordeel.

	meetmoment 1		meetmoment 2		moment 1 / 2 correlatie
	IBB	IBE	IBB	IBE	
controlegroep (n=10)	.18	.17	.16	.10	.42*
experimentele groep (n=10)	.16	.19	.24	.30	-.04

* = significant

die verwacht kan worden op basis van het toeval. De propooritie overeenkomst (p^e) wordt berekend aan de hand van de randfrequenties uit de overeenstemmingsmatrix. Oorspronkelijk is kappa een maat voor de overeenstemming tussen twee observatoren. Om voor iedere, uit tien observatoren bestaande, groep een "kappa totaal" te bepalen werden van de 45 mogelijke observatorcombinaties de gemiddelde p^e en de gemiddelde p^e bepaald.

Met behulp van de Pearsoncorrelatie ($p < 0.1$) werden tenslotte voor elke groep afzonderlijk de waarden van kappa totaal op beide meetmomenten met elkaar vergeleken om het verband tussen de mate van overeenstemming op meetmoment 1 en meetmoment 2 te bepalen.

Resultaten

Geen van de twintig observatoren viel uit gedurende het onderzoek. In tabel 1 is de kappa totaal voor beide groepen aangegeven evenals de Pearson-correlatie voor iedere groep afzonderlijk.

Het eerste wat uit deze gegevens blijkt is dat *alle* kappa's opvallend laag zijn. De onderlinge verschillen tussen de diverse kappa's zijn bovendien weinig imponerend. De kappa-totaalwaarden van beide groepen op meetmoment 1 blijken niet veel te verschillen, zodat aangemomen kan worden dat beide groepen vergelijkbaar waren voordat de interventie plaats vond. Wel opvallend zijn de correlaties tussen de beide meetmomenten. Er is een duidelijk verschil tussen de controlegroep en de experimentele groep. De correlatie in de controlegroep suggereert enig verband tussen beide meetmomenten. In de experimentele groep is van enige correlatie geen sprake, hegeen op een trainingseffect zou kunnen wijzen. Een opvallend verschijnsel bij de data-analyse op observatorniveau was dat zowel in de controlegroep als in de experimentele groep op meetmoment 1 bij zes van de 45 observatorcombinaties een negatieve kappa werd gevonden (IBB). Op meetmoment 2 bleken in de controlegroep negen observatorcombinaties een negatieve kappa te vertonen; daarentegen waren in de experimentele groep alle negatieve kappa's verdwenen. Na de training bleken drie van de tien observatoren in de experimentele groep minder overeenstemming met het expertoordeel te bereiken (IBE) dan voor de training. In de controlegroep viel op dat op meetmoment 2 acht observatoren minder en twee observatoren meer overeenstemming vertoonden met het expertoordeel (gouden standaard) dan op meetmoment 1; met andere woorden, geen enkele observator scoorde consistent.

Conclusie en beschouwing

De correlatieverschillen tussen de controlegroep en de experimentele groep suggereren een trainingseffect. Het resultaat is echter niet opmerkelijk, aangezien in de wetenschappelijke literatuur kappawaarden <0.40 als een mi-

nieme overeenstemming beschouwd worden.¹³ Er werd derhalve besloten om geen re-entemering uit te voeren. De lage kappa's bij deze studie nopen echter wel tot enige kritische kanttekeningen. Allereerst is Cohens kappa een strenge maat voor de interbeoordelaarsbetrouwbaarheid, aangezien bij de berekening gecorrigeerd wordt voor toeval door rekening te houden met de frequenties in de randtotaal. Verder wordt de mate van overeenstemming op itemniveau bekeken, waarna deze op de totale overeenstemming wordt gemiddeld. Op het niveau van de totale toets zullen de verschillen tussen de diverse beoordelaars zich uitmiddelen, hegeen de toetsbetrouwbaarheid ten goede komt. De uitkomst van dit onderzoek is conform soortgelijke bevindingen in de literatuur en is dan ook niet verrassend te noemen. Ook uit deze studie blijkt dat training nagenoeg geen effect heeft op de interobservervariantie.

Opvallend was dat drie observatoren na de training minder overeenstemming met het expertoordeel vertoonden. Deze observatoren behoorden tot drie verschillende vakgroepen. Hierbij moet rekening gehouden worden met de mogelijkheid dat de training toch onvoldoende structuur heeft geboden. Een ander fact is het inconsistent scoren door alle observatoren van de controlegroep, zoals blijkt uit de vergelijking van de gegevens van beide meetmomenten met het expertoordeel. Tegen deze achtergrond dringt de vraag zich op of een interval van twee weken tussen meetmoment 1 en meetmoment 2 wellicht niet te lang is geweest. Dit kan zowel gelden voor de controlegroep als voor de experimentele groep.

Mogelijk is er sprake van andere storende variabelen gedurende het experiment waarover geen controle bestond. Daarbij valt te denken aan een motivaatprobleem, aangezien de controlegroep opnieuw dezelfde videoband diende te scoren. Daarnaast vonden alle melingen aan het einde van de werkdag plaats; hegeen de concentratie wellicht niet ten goede kwam. Het is evenwel ook mogelijk dat de observato-

ren uit de controlegroep kritischer zijn gaan kijken.

Er zijn genoeg vragen gerezen om verder onderzoek te entameren. De vraag is echter of hoge prioriteit gegeven moet worden aan het verbeteren van de interbeoordelaarsbetrouwbaarheid, gezien de beperkte invloed van deze factor op de totale toetsbetrouwbaarheid. Een argument om wel aandacht aan het gedrag van en het beoordelen door observatoren te besteden is dat er klachten van studenten over observatoren zijn.¹⁴ Bovendien is het onbekend wat voor de individuele student het gevolg is van een minder goede prestatie op een station met betrekking tot de prestaties op volgende stations. Aangezien elke student een eerlijke beoordeling verdient op elk station, is het de moeite waard om tijd en mankracht te blijven spenderen aan training van observatoren. Gezien de unanieme waardering van de observatoren uit de experimentele groep voor de opzet van de training voor het experiment, zou deze trainingsopzet wellicht ook toegepast kunnen worden bij de 'gewone' observatortrainingen.

Literatuur

1. Harden RM, Gleason FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;13:41-54.
2. Luijk SJ van. Al doende leert men [proefschrift]. Maastricht: Universiteit Maastricht; Universitaire Pers; 1994.
3. Vleuten CPM van der. Toetsing van medische competentie. In: Metz JCM, Schephier AJJA, Vleuten CPM van der, redactie. *Medisch onderwijs in de praktijk*. Assen: Van Gorcum, 1995:151-64.
4. Cunningham JPW, Neville AJ, Norman GR. The risk of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education* 1997;1:227-336.
5. Streiner DL. Global rating scales. In: *Neurfield VR, Norman G, redactie. Assessing clinical competence*. New York: Springer Publishing Company, 1985:119-41.
6. Vleuten CPM van der, Luijk SJ van, Ballegooijen AMJ, Swanson DB. Training and experience of medical examiners. *Med Educ* 1989;22:290-6.

7. Newble DJ, Hoare J, Speldrake PF. The selection and training of examiners for clinical examinations. *Med Educ* 1980;14:345-9.
8. Mouloupoulos SD, Stamatiopoulos S, Nanas S, Economides K. Medical education and experience affecting intra-observer variability. *Med Educ* 1986;20:133-5.
9. Luddbrook J, Marshall VR. Examiner training for clinical examinations. *Br J Med Educ* 1971;5:152-5.
10. Wilson GM, Lever R, Harden RM, Robertson II, MacRitchie J. Examination of clinical examiners. *Lancet* 1969;Jan 4:37-40.
11. Boumans MTA, Ooy A van. Het onderzoek van de onderste extremiteiten. Utrecht: Wetenschappelijke uitgeverij Bunge, 1995;47-68.
12. Cohen J. A coefficient of agreement for nominal scales. *Educational & Psychological Measurement* 1960;20(1):37-46.
13. Schouten HJA. Een praktische inleiding in methodologie en analyse. Houten/Diegem: Bohn Stafleu Van Loghum, 1995;33-8.

14. Visser K, Louw A de, Luijk SJ van, Scherpbier AJJA. De observator geobserveerd. In: Houtkoop E, Pols J, Pollenans MC, Scherpbier AJJA, Verwijnen GM, redactie. *Gezond Onderwijs - 3*. 's-Gravenhage: Haagse Hogeschool, 1994;119-24.

DE AUTEURS

- M.T.A. Boumans, arts, vaardigheidsdocente Skillslab, Universiteit Maastricht.
Dr. A.J.J.A. Scherpbier, arts, hoofd van het Skillslab, Universiteit Maastricht.
Dr. A. van Ooy, orthopedisch chirurg, Vakgroep Orthopedie, Universiteit Maastricht.
Prof. dr. C.P.M. van der Vleuten, psycholoog, Vakgroep Onderwijsontwikkeling en Onderzoek, Universiteit Maastricht.
R.J.I. Hoogenboom, research assistent, Vakgroep Onderwijsontwikkeling en Onderzoek, Universiteit Maastricht.
M.L.W.T. Schuurman, arts, medisch-onderwijskundige Vakgroep Onderwijsontwikkeling en Onderzoek, Universiteit Maastricht.

Correspondentieadres:

M.T.A. Boumans, Skillslab, Faculteit der Geneeskunde, Postbus 616, 6200 MD Maastricht.