Tijdschrift voor Medisch Onderwijs, juni 2000 | Vol. 19, nr.4, p. 128-141

Schriftelijke toetsen voor klinische probleemoplosvaardigheden: een overzicht

J. Beullens, H. Jaspaert

Samenvatting

Het toetsen van klinische probleemoplosvaardigheden is niet eenvoudig. Dit artikel beschrijft de resultaten van een literatuuronderzoek betreffende schriftelijke mogelijkheden voor het toetsen van klinische competentie: patient management problems (PMP's), modified essay questions (MEQ's) en de key features approach. Deze toetsvormen zijn realistischer dan mondelinge, essay- of meerkeuze-examens en bevorderen dat studenten zich bij het studeren richten op het toepassen van kennis. PMP's en MEQ's laten een betrouwbare meting toe, maar deze vereist wel ongeveer een dag toetstijd. Er zijn ook aanwijzingen voor de begripsvaliditeit van beide toetsvormen. De key features approach laat mogelijk kortere toetstijden toe, maar de psychometrische kwaliteit is minder duidelijk en er wordt uitgegaan van betwistbare veronderstellingen. Veelbelovend voor de toekomst lijken case-based simulations die via de computer worden afgenomen.

Inleiding

Een netelig probleem in het medisch onderwijs is het toetsen van klinische probleemoplosvaardigheden. Deze vaardigheden kunnen in de context getoetst worden door studenten een probleem te laten oplossen dat aangeboden wordt door een (simulatie)patiënt of in de vorm van een ziektegeschiedenis. Op het eerste gezicht is het objective structured clinical examination (OSCE) of stationsexamen hiervoor een valide toetsvorm. Een OSCE bestaat uit een aantal stations waarbij de student in elk station dezelfde tijd toegemeten krijgt om telkens een verschillende taak uit te voeren. De opdracht kan betrekking hebben op een probleem dat door een gestandaardiseerde patiënt naar voren wordt gebracht. Een gestandaardiseerde patiënt is een echte patiënt of een simulatiepatiënt die getraind is om een casus accuraat en consistent te presenteren. Doordat de organisatie van een OSCE veel tijd kost en door de kosten voor observatoren, gestandaardiseerde patiënten en verbruiksgoederen is de toets echter relatief duur.[1] Vooral als het aantal studenten groot is, wordt een grote organisatorische inspanning vereist. Haalbaarheidsproblemen spelen wellicht een minder grote rol bij de afname van schriftelijke toetsen.

Dit artikel vormt de neerslag van een literatuurstudie betreffende schriftelijke toetsen voor het meten van klinische probleemoplosvaardigheden, met name patient management problems (PMP's), modified essay questions (MEQ's) en de key features approach. Elke toetsvorm wordt bondig beschreven. Gezien de uitgebreidheid van een PMP of een MEQ (doorgaans de omvang van een boekje) worden in dit artikel geen voorbeelden opgenomen; wel wordt verwezen naar publicaties met voorbeelden. Verder worden voor- en nadelen en betrouwbaarheid en validiteit van elke examenvorm besproken. Ten slotte worden adviezen gegeven om de grootste bedreigingen voor betrouwbaarheid en validiteit aan te pakken.

Methode

In de Medline-bestanden van 1966 tot 1999 is gezocht naar literatuur over 'patient management problems', 'modi-

fied essay questions' en 'key features'. Deze trefwoorden werden gecombineerd met 'objectivity', 'reliability', 'validity', 'evaluation' en 'assessment'. Bij de rapportage zijn ook de artikelen uit de jaren zestig in aanmerking genomen, om een zo volledig mogelijk beeld te kunnen schetsen. De search werd beperkt tot het medisch onderwijs en medici. Abstracts en teksten uit proceedings van congressen zijn buiten beschouwing gebleven.

Patient management problems

PMP's werden ontwikkeld om op papier een arts-patiëntencontact na te bootsen. Ze werden in brede kring gebruikt om medische probleemoplosvaardigheden te toetsen.[2] Voorlopers van deze toets om diagnostisch probleemoplossen te meten zijn zowel in onderwijskundige als in medisch onderwijskundige kringen te vinden.[3 4] Deel III van het examen van de National Board of Medical Examiners (NBME) in de Verenigde Staten bestond reeds in de jaren zestig uit dit type vragen.[5] In de jaren zeventig werden ze zeer populair en werden ze opgenomen in de examens van de American National Boards.[6] Deze examenmethode werd verder ontwikkeld door McGuire en medewerkers in de Verenigde Staten en in Europa door Harden en medewerkers. Reeds in de jaren zeventig werden geautomatiseerde vormen uitgeprobeerd.[7] De NBME werkt momenteel aan de ontwikkeling van een computerversie voor gebruik op grote schaal.[8]

Een PMP werd oorspronkelijk gepresenteerd in een boekje. Het PMP bestaat uit een korte beschrijving van de patiënt en diens hoofdklachten, gevolgd door secties gewijd aan anamnese, klinisch onderzoek, diagnose en behandeling. Binnen elke sectie kiest de student uit een lijst van opties de meest geschikte. Op de linkerhelft van de bladzijde staan opties betreffende vragen die gesteld kunnen worden aan de patiënt, lichamelijk onderzoek, laboratoriumonderzoek, et cetera. Op de rechterhelft van de bladzijde staan de antwoorden/resultaten voor elke gekozen optie. Deze zijn onzichtbaar gemaakt, bijvoorbeeld doordat er een laagje over is aangebracht. Door antwoorden zichtbaar te maken (het laagje weg te krassen) geeft de student aan welke aanpak hij kiest voor het betreffende patiëntenprobleem. De antwoorden zijn in een zo realistisch mogelijke vorm gegoten. Elke beslissing wordt gevolgd door een beschrijving van het effect daarvan op de gezondheidstoestand van de patiënt, zodat de student in daaropvolgende stadia te maken heeft met de gevolgen van vorige beslissingen.[2 9] Uitgewerkte voorbeelden en informatie over de ontwikkeling van PMP's zijn elders te vinden.[10 11]

Er bestaan drie soorten PMP's: problemen met een lineaire structuur, problemen met een vertakte structuur en open problemen. Vragen moeten van het meerkeuzetype zijn in een vertakte structuur; in de lineaire structuur kan het ook om open vragen gaan. Terwijl in een lineaire structuur alle studenten hetzelfde pad volgen en ze in een vertakte structuur regelmatig terugkomen naar het hoofdpad, bepaalt bij het open probleem elke student zijn eigen pad.[10] Bij de scoring worden positieve punten toegekend voor nuttige opties, geen punten voor neutrale opties en negatieve punten voor schadelijke opties. Bij de beoordeling worden over het algemeen vijf scores toegekend: de efficiëntiescore (het aantal nuttige opties gedeeld door het totale aantal gekozen opties), de vaardigheidsscore (overeenstemming met experts, blijkend uit gekozen nuttige en vermeden schadelijke opties), de nalatigheidsscore (opties die gekozen moesten worden, maar niet gekozen werden), de overdaadscore (te vermij-

den opties die toch gekozen werden) en de competentiescore (berekend op basis van efficiëntie- en vaardigheidsscore, waarbij de laatste meer gewicht krijgt).[12 13]

Voordelen

PMP's meten de klinische competentie op een meer realistische wijze dan mondelinge, essay- of meerkeuze-examens.[14] Ziekte- en patiëntenfactoren zijn onder controle omdat de studenten dezelfde casus beoordelen. Een groot aantal studenten kan geëvalueerd worden op een relatief eenvoudige en goedkope wijze.[15] PMP's richten het leren en het evalueren op meer praktische en relevante doelstellingen. Ze stimuleren de student probleemoplosvaardigheden te ontwikkelen en ze voorzien de student van een situatie waaromheen hij zijn kennis kan opbouwen.[10]

Nadelen

De ontwikkeling van PMP's vereist tijd en inspanning en PMP's zijn duurder dan meerkeuzevragen. Studenten die geleerd hebben een zo volledig mogelijke anamnese af te nemen, zullen minder geneigd zijn selectief te werk te gaan in de secties anamnese en klinisch onderzoek. Sommige studenten krassen per ongeluk op de verkeerde plaats. De lange lijsten van mogelijk relevante keuzen bevatten aanwijzingen die in echte klinische situaties ontbreken.[16] Omgekeerd zijn tijdens een werkelijk consult visuele aanwijzingen aanwezig, die in schriftelijke simulaties ontbreken. Doordat in de simulatie de aandacht van de studenten gevestigd wordt op specifieke punten, kan de inhoud en organisatie van de schriftelijke simulatie arbitrair en kunstmatig lijken.[15]

Betrouwbaarheid

De betrouwbaarheid van een toets heeft betrekking op de nauwkeurigheid van de meting. Doorgaans wordt de betrouwbaarheid gemeten door van een toets de interne consistentie en de test-hertestbetrouwbaarheid te berekenen. De interne consistentie neemt toe met het aantal PMP's.[12] Deze bedroeg 0.52 bij twee problemen, 0.70 bij zes en 0.72 à 0.75 bij elf à zestien problemen.[17-19] Dit geeft de indruk dat wellicht al bij 20 of 25 problemen een voldoende betrouwbaarheid kan worden bereikt. De betrouwbaarheid bij hertoetsing na zestig à negentig dagen was vrij behoorlijk, zelfs bij gebruik van slechts één PMP. Ze bedroeg 0.72 voor een hypertensieprobleem en 0.73 voor een CARA-probleem.[20] Samenvattend kan gesteld worden dat een betrouwbaar examen de afname van meer PMP's vereist dan doorgaans worden gebruikt.

De belangrijkste bedreiging voor de betrouwbaarheid van klinische examens waarbij uitgegaan wordt van een casus of een patiënt, vormt de casusspecificiteit. De casusspecificiteit houdt verband met de medische casus als een uniek geheel. Ze verwijst naar de consistente vaststelling dat de prestatie op de ene medische casus weinig predictieve waarde heeft voor de prestatie op een andere casus. Uit een onderzoek betreffende een OSCE met gestandaardiseerde patiënten bleek bijvoorbeeld dat de prestaties op een welbepaalde component over verschillende casus heen minder overeenstemming vertoonden dan de prestaties op verschillende componenten binnen dezelfde casus.[21] Vermoedelijk speelt de casusspecificiteit een even belangrijke rol bij schriftelijke simulaties zoals het PMP. De correlaties tussen enkele PMP's onderling schommelden tussen 0.29 en 0.61.[22]

Er zijn echter ook aanwijzingen in tegengestelde richting. Een factoranalyse die uitmondt in enkele factoren, wijst meer op algemene vaardigheden dan op specifieke vaardigheden voor elke casus. Berner en medewerkers voerden een fac-

toranalyse uit op vijf PMP-casus en troffen vier factoren aan.[23] Ze beschouwden dit als steun voor de interpretatie dat de prestaties op dezelfde vaardigheid over verschillende klinische casus heen consistenter kunnen zijn dan de prestaties op verschillende vaardigheden binnen dezelfde casus. Juul en medewerkers voerden drie factoranalyses uit op 24 of 26 PMP's en vonden telkens dezelfde twee factoren, die 'gegevensverzameling' en 'beleid' werden genoemd.[2] De tweede aanwijzing leverde het onderzoek op waarbij voor twee onderwerpen telkens een PMP en een goed/fouttoets ontwikkeld werden. De scores van dezelfde toetsen voor verschillende onderwerpen lagen dichter bij elkaar dan die van verschillende toetsen voor hetzelfde onderwerp.[24]

Validiteit

De validiteit betreft de mate waarin een examen meet wat het hoort te meten. De *inhoudsvaliditeit* geeft de overeenstemming weer tussen de inhoud van het examen en de opleidingsdoelen. Niet alleen experts maar ook artsen en studenten vonden dat PMP's een realistische benadering vormden van de klinische praktijksituatie.[20] Als het examen uit slechts enkele PMP's bestaat, komt door de casusspecificiteit de inhoudsvaliditeit in het gedrang en is het lastiger de opleidingsdoelen te dekken. Er is dus nog niet aangetoond dat PMP-examens een voldoende inhoudsvaliditeit bezitten.

De *criteriumvaliditeit* betreft ondermeer de overeenkomst tussen de scores op het onderzochte examen en die op andere examens. Er was geen verband tussen het resultaat op een PMP en de beoordeling van een onderhoud met een gestandaardiseerde patiënt.[25] Er was ook nauwelijks samenhang met stagescores betreffende patiëntenzorg en beroepsattitudes (0.16).[26] Ook het verband tussen het oordeel van collega's en scores op PMP's was zwak (0.26).[19] Het is dus niet verrassend dat ook de correlatie met de klinische beoordeling laag was (0.19).[27] Vaak werd de samenhang berekend van PMP's met andere schriftelijke examens bestaande uit meerkeuzevragen. De correlaties met examens met meerkeuzevragen schommelden tussen 0.17 en 0.64.[25 28-30] Alhoewel sommige daarvan significant waren, waren de meeste toch vrij laag.[12] Interessant is nog dat de correlatie van PMP's met meerkeuzevragen van het probleemoplostype (0.44) hoger was dan met meerkeuzevragen van het herinnering/herkenningtype (0.25).[28] Met Deel I en II van de National Board Examinations werden correlaties gerapporteerd van respectievelijk 0.26 en 0.33 en met het American Board of Internal Medicine Examination van 0.54.[27 30] Bij pas afgestudeerde artsen werd een significante correlatie van 0.42 vastgesteld met de examenresultaten.[26]

Over het algemeen zijn de verbanden tussen de scores op PMP's en andere klinische of schriftelijke examens zwak, wat betekent dat ze wellicht iets anders meten.[7] De lage correlaties kunnen echter ook te wijten zijn aan een te lage betrouwbaarheid van de PMP's en andere klinische toetsen. De bevindingen inzake de criteriumvaliditeit laten dus geen eenduidige interpretatie toe.

De *begripsvaliditeit* kan bepaald worden door veronderstellingen ten aanzien van het te toetsen begrip of concept te onderzoeken. Met PMP's hoopt men probleemoplosvaardigheden te meten. Een factoranalyse op PMP's zou bijgevolg moeten resulteren in factoren die betrekking hebben op aspecten van probleemoplossen. Berner en medewerkers vonden factoren die te maken hadden met de adequaatheid van de initiële problemenlijst, de bekwaamheid diagnostische procedures te rangordenen en het vermogen tot

een diagnose te komen.[23] Juul en medewerkers vonden twee factoren die ze 'gegevensverzameling' en 'beleid' noemden.[2] Beide factoren bleven stabiel over de tijd en tussen studenten van verschillende jaren. Een andere aanwijzing voor begripsvaliditeit was dat de correlatie tussen beide factoren hoger was bij de 'seniors' dan bij de 'juniors'. De 'seniors', die meer klinische ervaring hadden, waren beter in staat dataverzameling en beleid op elkaar af te stemmen. Palchik en medewerkers voerden een factoranalyse uit op elk van veertien PMP's afzonderlijk en vonden daarbij telkens twee à drie factoren.[31] De vijf secties (anamnese, lichamelijk onderzoek, diagnostisch onderzoek, diagnose en beleid) laadden bij alle PMP's verschillend op de twee à drie factoren. De factorladingen suggereerden een verband tussen de sectie diagnose en ten minste één van de secties die betrekking hadden op gegevensverzameling. Het leek er dus op dat het diagnostisch beslissingsproces werd beïnvloed door de strategieën die gebruikt werden bij het verzamelen van informatie.

De begripsvaliditeit van PMP's kan ook getoetst worden door te onderzoeken of studenten die verder gevorderd zijn in hun opleiding of meer ervaring hebben, hogere scores behalen. Hierbij wordt aangenomen dat de probleemoplosvaardigheden toenemen in de loop van de medische opleiding en nadien in de praktijk. Enkele onderzoeken toonden aan dat ouderejaarsstudenten inderdaad significant hoger scoorden dan jongerejaarsstudenten.[2 13 32] Een analoge vaststelling werd ook gedaan bij assistenten.[29] Een acht weken durende stage bleek echter geen effect te hebben op de PMP-score. Studenten die stage gelopen hadden op een afdeling die verband hield met het PMP, hadden daarop geen hogere score dan studenten die stage liepen op andere afdelingen.[33] Een opleidingscomponent die een introductie inhield tot de concepten van klinisch probleemoplossen en de integratie daarvan met de praktijk op de afdelingen, had daarentegen wel een significant positieve invloed op de PMP-scores.[17] Een paar onderzoeken illustreerden dat artsen al dan niet in opleiding, een significant hogere score hadden dan studenten.[20 34] Newble en medewerkers daarentegen vonden dat artsen een lagere vaardigheidsscore behaalden dan studenten, maar zij voerden geen statistische analyse uit.[13] Wolf deed dit wel en kwam tot verschillende conclusies afhankelijk van de gebruikte analysetechniek. Met een variantieanalyse werd geen verschil gevonden, maar uit een trendanalyse bleek dat er een significante, lineaire stijging van de efficiëntiescore was van vierdejaarsstudenten tot artsen.[32 35] Op de vaardigheidsindex hadden artsen evenwel een lagere score dan zesdegraadsstudenten.[32] De artsen verschilden ook van de studenten in de hoeveelheid opgevraagde relevante informatie. Deze was bij de artsen kleiner.[13 32] Dit heeft alles te maken met de ontwikkeling van expertise: experts selecteren alleen die informatie die ze nodig hebben om het probleem op te lossen.[32] In slecht samengestelde PMP's wordt deze verhoogde efficiëntie in gegevensverzameling bestraft met lagere scores. Marshall en medewerkers signaleerden dat ervaren artsen hogere PMP-scores behaalden dan artsen in opleiding, maar gaven niet aan of dit verschil significant was.[24] Easterling daarentegen merkte op dat assistenten uit het vierde jaar in sommige jaren een significant hogere score behaalden dan artsen.[29] Samenvattend, kan besloten worden dat er aanwijzingen zijn voor de begripsvaliditeit van het PMP-examen.

De begripsvaliditeit wordt echter bedreigd door cueing: de invloed van de aangeboden opties op het aantal (op

papier) ondernomen acties tijdens het examen. Bij PMP's gaat men ervan uit dat de acties die de student onderneemt om het probleem op te lossen, overeenkomen met wat hij of zij zou doen bij een (gestandaardiseerde) patiënt of in een interviewsituatie. Met andere woorden de verwachting is dat de student in beide situaties evenveel informatie zal opvragen. In vergelijking met een onderhoud met een gestandaardiseerde patiënt kozen studenten bij een PMP 20 tot 150% meer opties in elk van de vier secties anamnese, klinisch onderzoek, laboratoriumonderzoeken en beleid.[36] Indien men alleen rekening hield met de kritische opties, werd een vergelijkbaar significant verschil vastgesteld. Het verschil was het meest uitgesproken bij anamnese en beleid. In twee andere onderzoeken waarin de opgevraagde informatie in een PMP werd vergeleken met die in een interview, werd in de eerste situatie 25 tot 74% meer informatie opgevraagd.[13 37] McCarthy noemde dit effect 'cueing' omdat de studenten door de beschikbare lijst opties ertoe worden aangezet meer informatie te selecteren dan ze in een andere situatie zouden kiezen.[37] Nog verontrustender voor de begripsvaliditeit was dat er geen overeenkomst bestond tussen de scores op de schriftelijk en de mondeling aangeboden casus.[37] In een van de zeldzame onderzoeken waarin de kosten werden geschat, werden de kosten die zouden voortvloeien uit de beslissingen van assistenten op PMP's, vergeleken met de kosten van hun werkelijke beslissingen in de dagelijkse praktijk, ontleend aan dossiers. De kosten bij de simulaties waren hoger dan die in de werkelijkheid en er was een significant positief verband ($r = 0.38$) tussen beide.[38]

Afgezien van een onbedoeld gunstig effect op de eindscore kan cueing ook leiden tot een negatief effect. Bij de sectie laboratoriumonderzoeken werd bijvoorbeeld een lage niet-significante correlatie (0.16) gevonden tussen het aantal gekozen opties en de score. Dit betekent dat de studenten er door de optielijst ook toe werden aangezet potentieel schadelijke procedures te selecteren. Bij vergelijking van de tien beste met de tien zwakste studenten bleek de puntenwinst van deze laatste groep groter. Ten opzichte van een mondeling aangeboden casus zonder aanwijzingen ('cues') speelt de PMP dus in het voordeel van de zwakkere studenten.

De bevindingen inzake de validiteit samenvattend kan besloten worden dat de inhoudsvaliditeit goed is bij een voldoende aantal PMP's maar dat doorgaans te weinig PMP's worden gebruikt. Mede daarom zijn de bevindingen inzake de criteriumvaliditeit moeilijk te interpreteren. De bevindingen betreffende de begripsvaliditeit zijn inconsistent. Onderzoeken met factoranalyses of groepen met verschillend opleidingsniveau leverden positieve aanwijzingen. PMP's geven echter geen correct beeld van wat de student in een interviewsituatie zal doen en dit is waarschijnlijk te wijten aan het cueing-effect.

Adviezen ter verbetering van de PMP-toets

De grootste bedreiging voor de betrouwbaarheid (en inhoudsvaliditeit) vormt de casusspecificiteit en voor de begripsvaliditeit het cueing-effect. De remedie tegen casusspecificiteit is theoretisch eenvoudig maar praktisch soms moeilijk haalbaar: gebruik een groot aantal problemen met uiteenlopende casus. Dit levert echter weer haalbaarheidsproblemen op: de ontwikkeling van PMP's is erg tijdrovend en bij een groot aantal PMP's is de afname niet meer realiseerbaar op een halve dag. Een andere oplossing voor het probleem van de casusspecificiteit biedt mogelijk de key features approach, waarbij de vragen beperkt blijven tot de meest relevante elementen in de oplossing van het medisch

probleem, zodat in eenzelfde tijdsbestek meer casus aan bod kunnen komen. Op de key features approach wordt verderop in dit artikel nader ingegaan.

De oplossing bij uitstek om het cueing-effect te voorkomen is om geen antwoord-alternatieven te voorzien. Bij schriftelijke PMP's betekent dit dat open vragen gesteld worden waarop de student binnen de gereserveerde ruimte antwoordt. In een onderzoek werd dit uitgeprobeerd voor de sectie differentiële diagnose en niet voor de secties anamnese, klinisch onderzoek, diagnostische onderzoeken en therapeutische procedures. De betrouwbaarheid bleef hetzelfde, met of zonder 'uncued' sectie. De spreiding van de scores was evenwel groter als de 'uncued' sectie in de berekening betrokken werd en dit kwam de begripsvaliditeit ten goede.[17]

Modified essay questions

De MEQ werd in Engeland ontwikkeld, in eerste instantie voor vervolgopleidingen, maar al snel werd deze toetsvorm ook gebruikt voor medisch studenten.[39] Later is de MEQ ook toegepast in ondermeer Australië, Canada en de VS. Het gaat om korte essayvragen. Door de open vragen is er geen cueing-effect: er zijn geen ongewenste aanwijzingen. Een uitvoerige lijst van voorbeelden met mogelijke, korte antwoorden is elders te vinden.[41] In Nederland is de term 'gestructureerde open vraag' gangbaar.[42]

MEQ-toetsen gaan uit van een ziektegeschiedenis en worden aan de student voorgeschoteld in stadia.[39] Deze stadia betreffen zowel het verloop van de consultatie als het beloop van de ziekte. De vragen worden in de vorm van een boekje aangeboden. Het boekje begint met een kort scenario met daarin beschreven de rol van de student, de medische faciliteiten waarover hij beschikt, en details van de casus. Elke bladzijde bevat nieuwe informatie over de patiënt gevolgd door een vraag. Deze informatie maakt deel uit van het probleem van de patiënt en de vraag zet de student ertoe aan om de gegevens aan te wenden om een beslissing te nemen.[40] De vragen dienen zowel de vaardigheden herinnering en interpretatie als probleemoplossing te meten. Om de toets levensecht te maken, kan de student gevraagd worden röntgenfoto's, elektrocardiogrammen, microscopie-preparaten of gedragingen op videoband te interpreteren.[43]

Soms wordt onderscheid gemaakt tussen ongerichte vragen, die algemeen van aard zijn en een halve tot volledige bladzijde antwoord vereisen, en gerichte vragen, waarbij het gewenste aantal antwoordpunten wordt aangegeven. De afname van het examen zelf wordt uiteraard gesuperviseerd. Er dient namelijk op gelet te worden dat de studenten niet verder bladeren in het boekje en evenmin teruggaan om vorige antwoorden te corrigeren.[39] Om de scoring zo objectief mogelijk te maken, worden op voorhand modelantwoorden en een puntenverdeling voorzien.

Voordelen

Rabinowitz somt de volgende voordelen op: modified essay questions (1) simuleren de klinische volgorde van de beslissingen in de klinische praktijk, (2) kunnen naast objectieve feitenkennis en het vermogen van de kandidaat om gegevens te verzamelen, te selecteren en te gebruiken, ook herinnering toetsen evenals inzicht, oordeel en selectiviteit in het probleemoplossingsproces, (3) kunnen de effecten van biologische, psychologische en sociologische factoren omvatten, (4) bevatten geen openlijke aanwijzingen en (5) geven onmiddellijk feedback aan de kandidaten.[44]

Nadelen

De samenstelling van het vragenboekje en de ontwikkeling van scoringsschema's zijn

tijdrovende taken. Ook het scoren zelf neemt veel tijd in beslag.[44] De kandidaten kunnen er voordeel bij hebben het boekje door te nemen vóór ze beginnen (wat dan ook ten strengste verboden is). Ze kunnen de vragen ook beantwoorden volgens wat ze denken dat de examinator zou willen horen eerder dan wat ze werkelijk zouden doen.[41] Met het volledig doorwerken van alle stadia van een casus in een MEQ is erg veel tijd gemoeid. De respondenten weten bovendien niet hoe gedetailleerd hun antwoord moet zijn.[42]

Betrouwbaarheid

De betrouwbaarheid van het examen kan gemeten worden door de interne consistentie en de interbeoordelaarsbetrouwbaarheid te berekenen. Bij dertien examens voor eerste- en tweedejaarsstudenten geneeskunde varieerde Cronbach's alfa, die de interne consistentie aangeeft, van 0.26 tot 0.81 met als mediaanwaarde 0.54.[43] De interne samenhang laat bijgevolg te wensen over. In een ander onderzoek werden de antwoorden op gerichte essayvragen (opsommingen) en op ongerichte (algemene) vragen gescoord door twee beoordelaars. De interbeoordelaarsbetrouwbaarheid bedroeg 0.84 à 0.97 voor de gerichte vragen en 0.80 à 0.91 voor de ongerichte.[45] De MEQ's voldeden bijgevolg wel aan deze vorm van betrouwbaarheid.

Validiteit

Om de *inhoudsvaliditeit* te meten wordt de inhoud van het examen vergeleken met de inhoud van het opleidingsonderdeel. Als het scoringsschema wordt samengesteld door een groep van experts, wordt verondersteld dat de inhoudsvaliditeit hoog is.[41]

De *criteriumvaliditeit* kan bepaald worden door de overeenkomst met de scores op andere examens. De MEQ's vertoonden een significante samenhang met de uitslag op een klinisch examen (0.42).[46] Sommige aspecten van het oordeel van de klinische supervisor hingen samen met de MEQ-score, andere niet. Het verband met het oordeel van de klinische supervisor was significant voor beroepsattitudes (0.23) maar niet voor medische kennis, gegevensverzameling en klinisch oordeel.[44] MEQ's correleerden significant met andere essayvragen(0.23 à 0.50) en met meerkeuzevragen (0.36 à 0.43).[46] Significante correlaties werden vastgesteld van meerkeuzevragen met gerichte MEQ's (0.48) of ongerichte MEQ's die de inhoud betreffen (0.46), maar niet met ongerichte MEQ's die het proces betreffen (0.20).[41] In een ander onderzoek was er een significante samenhang met meerkeuzevragen (0.36) en met Deel II van het examen van de NBME (0.30), maar niet met Deel I en III (respectievelijk 0.25 en 0.19).[45] Over het algemeen is de criteriumvaliditeit bijgevolg niet erg overtuigend aangetoond. Er is een zwak verband tussen de uitslagen op MEQ's en uitslagen op klinische examens, maar ook tussen MEQ's en uitslagen op theoretische examens. De lage correlaties kunnen echter ook te wijten zijn aan een te lage betrouwbaarheid van de MEQ's of van de andere toetsen.

De *begripsvaliditeit* is onderzocht door na te gaan of ook op itemniveau de probleemoplosvaardigheden die men meent te toetsen, aantoonbaar zijn. Er zouden dus minder items van het herinneringstype moeten zijn en meer van het interpretatie- en probleemoplossingstype. Bij nadere bestudering van een MEQ-examen bedroeg het percentage van de drie itemtypes respectievelijk 31, 50 en 18%. Dit betekent dat 68% van de items het herinneringsniveau overstegen.[43] Om de begripsvaliditeit te ondersteunen zouden de factoren resulterend uit een factoranalyse uitgevoerd op MEQ's, aspecten van probleemoplossen moeten betreffen. Twee factoranalyses

mondden uit in telkens tien factoren, die vaker items van het interpretatie- en probleemoplossingstype omvatten dan van het herinneringstype.[46] Als de probleemoplosvaardigheden toenemen in de loop van de medische opleiding, moet dit aan de MEQ-scores te merken zijn. De scores op zowel de gerichte als de ongerichte essayvragen namen significant toe van het eerste tot het derde jaar geneeskunde.[45] In een onderzoek werd een aanzienlijke toename (bijna een verdubbeling) van de MEQ-score opgemerkt na een vijf weken durende probleemgerichte cursus huisartsgeneeskunde.[47] Verschillende onderzoeken leverden door middel van verschillende methoden dus aanwijzingen voor de begripsvaliditeit van MEQ's.

Problemen en adviezen

Na jarenlange ervaring met het samenstellen van klinische essay-examens stelden Feletti en Smith de volgende problemen vast: (1) het aantal te beantwoorden vragen nam van jaar tot jaar toe, maar de daarvoor voorziene examentijd niet en (2) het percentage probleemoplosvragen verminderde van jaar tot jaar. De studenten klaagden over dubbelzinnige items. Ze zochten duidelijk naar aanwijzingen hoe deze te beantwoorden en naar de discipline van de corrector. Ze waren onzeker over de vereiste diepte van de uitleg.[48] Feletti en Smith gaven de volgende adviezen:[48]

- Er moeten regelmatige en vroegtijdige procedures worden ontwikkeld om de kwaliteit van de vragen te controleren.
- Elke vraag zou moeten opgesteld en gescoord worden door een kleine groep geïnteresseerde docenten.
- De items zouden het nemen van beslissingen, het oplossen van problemen of het toepassen van kennis moeten benadrukken.
- Voor elk item moet er ofwel een tijdslimiet (bijvoorbeeld tien minuten) voorzien worden, ofwel 30% extra tijd bovenop de door de auteur geschatte tijd voor dat item.
- De correctoren zouden de antwoorden meer volgens de geest dan volgens de letter van het modelantwoord moeten interpreteren.
- Een meer uniform criterium of een standaard is wenselijk.

Key features approach

In reactie op de problemen betreffende de psychometrische kwaliteit van PMP's en MEQ's werden twee toetsvormen ontwikkeld die toegespitst waren op de sleutelelementen (key features) in het klinisch probleemoplossingsproces. In Nederland werd de 'simulation of initial medical problem-solving' (SIMP) ontwikkeld en in Canada de 'Q4'. De bespreking wordt beperkt tot een beschrijving van de toets en een rapportering van de bevindingen betreffende de psychometrische kwaliteit. De literatuur laat niet toe uitvoerig in te gaan op voordelen en nadelen, laat staan adviezen te geven ter verbetering van de toetskwaliteit.

Simulation of initial problem-solving (SIMP)

De SIMP poogt een oplossing te bieden voor de twee grootste problemen van de PMP: de casusspecificiteit en het cueing-effect. Om dit laatste te voorkomen eindigt elke ziektegeschiedenis met de open vraag: 'Wat zou u doen als u een arts was die geconfronteerd werd met dit probleem?' Om in korte tijd zoveel mogelijk casus te kunnen aanbieden, wordt de toets toegespitst op de essentiële aspecten van de klinische competentie. In de literatuur van het hypothetico-deductief redeneren wordt een centrale rol toegekend aan de ontwikkeling van de initiële diagnostische hypothesen. De SIMP's meten daarom uitsluitend dit beginmoment in het probleemoplossingsproces. De toets

bestaat uit tien à twintig casus die elk vijf à tien minuten afname vereisen. Om een subjectieve beoordeling van de antwoorden te vermijden, werden antwoordsleutels samengesteld. Een voorbeeld van een SIMP met bijhorende antwoordsleutel wordt gegeven door De Graaff et al.[25]

De Graaff en co-auteurs gingen de betrouwbaarheid na van twee toetsen bestaande uit respectievelijk twaalf en tien casus.[25] De correlatie tussen twee beoordelaars was gemiddeld 0.75 voor de eerste toets en 0.82 voor de tweede. De interne consistentie daarentegen was matig: de correlaties tussen de casus en de toetsscore bedroegen gemiddeld respectievelijk 0.48 en 0.34. De generaliseerbaarheidscoëfficiënt berekend op zes andere casus bedroeg 0.74.[25]

In hetzelfde onderzoek werd voor dezelfde twee toetsen de criteriumvaliditeit berekend. De correlatie van de toetsen met twaalf en tien casus met de score gegeven door een simulatiepatiënt was hoog (respectievelijk 0.43 en 0.74), de correlatie met het oordeel over de globale bekwaamheid gegeven door een tutor was laag (0.14 en 0.27), evenals de correlatie met twee voortgangstoetsen (-0.16 à 0.25). De toets met zes casus correleerde significant met het globale oordeel van een simulatiepatiënt (0.38), laag met diens score op een gedetailleerde checklist (0.19) en negatief met de voortgangstoets (-0.25). Een verband met een andere meting van de klinische bekwaamheid en geen verband met de kennistoetsen bevestigen de criteriumvaliditeit.

Q4

De Medical Council of Canada verving de PMP-toets door een toetsvorm die Q4 genoemd werd, omdat deze het vierde onderdeel was van het qualifying examination.[49] In de veronderstelling dat de intercasuscorrelatie bij PMP's gemiddeld 0.10 bedroeg, schatten Bordage en Page dat een examen uit ongeveer veertig casus moet bestaan om een betrouwbaarheidscoëfficiënt van 0.80 op te leveren.[50] Om een Q4-examen met evenveel problemen te kunnen ontwikkelen, dienden de aangeboden problemen beperkt te worden tot de sleutelelementen in het probleemoplossingsproces. Bovendien moest het gaan om een representatieve en voldoende steekproef uit het domein van de klinische problemen. De 'key features' werden gedefinieerd als de kritische elementen bij de oplossing van een probleem en/of de elementen die het vaakst aanleiding geven tot fouten bij studenten of moeilijkheden bij artsen.[50] De vraag zelf kan de vorm aannemen van een meerkeuzevraag met een beperkt aantal opties of van een korte open vraag.[51] In die zin ligt de Q4 niet enkel in het verlengde van de PMP maar ook van de MEQ. De auteurs geven zelf enkele voorbeelden.[50 51]

Hoeveel problemen er precies nodig zijn om een betrouwbaar Q4-examen op te leveren, is niet duidelijk. Er zijn wel aanwijzingen voor de inhoudsvaliditeit. Ten aanzien van 71% van de 171 key features die door de toetscommissie werden ontwikkeld, waren drie à zes beoordelaars het unaniem eens dat deze sleutelelementen betroffen. De beoordelaars werden op hun beurt gevraagd key features te ontwikkelen; 94% hiervan kwam overeen met de oorspronkelijke sleutelelementen. Volgens de beoordelaars werden de studenten gedurende de stage één of meerdere malen geconfronteerd met de 59 door de toetscommissie geselecteerde problemen.[52] Rapporten over de criterium- en begripsvaliditeit van de Q4-toets werden in de literatuur niet aangetroffen.

Discussie

De vier beschreven toetsen zijn schriftelijke toetsen die werden ontwikkeld om klini-

sche probleemoplosvaardigheden te meten. Ze laten toe een grote groep studenten tegelijkertijd op een relatief eenvoudige en goedkope wijze te examineren. Bovendien meten ze de klinische competentie op een meer realistische wijze dan mondelinge, essay- of meerkeuze-examens. Daarenboven zetten ze de studenten er toe aan zich bij het leren te richten op het toepassen van kennis in plaats van louter op het uit het hoofd leren van feitenkennis. Wat zijn deze toetsvormen echter waard?

Met de betrouwbaarheid van PMP's en MEQ's is het nog niet zo slecht gesteld: een voldoende betrouwbaarheid is haalbaar indien voldoende problemen worden aangeboden. Hoeveel dit er minimaal moeten zijn, is vooralsnog niet duidelijk. In elk geval zijn meer dan zestien problemen noodzakelijk. Of dit er veertig moeten zijn, zoals Bordage en Page stellen, kan betwijfeld worden.[50] Zij gaan uit van een intercasuscorrelatie van 0.10, terwijl er ook waarden van 0.29 tot 0.61 werden gerapporteerd.[22] Vast staat dat een halve dag PMP-toetsing niet volstaat, een gehele dag (zes tot acht uur) is wellicht vereist.[6] Dit probleem doet zich evenwel voor bij alle toetsen die medische competentie meten, ook het OSCE. Alleen daarom deze lange toetsvormen verwerpen is bijgevolg niet correct. De inhoudsvaliditeit van PMP's en MEQ's is niet bewezen. Niet alleen vereist dit een groot aantal problemen, maar bovendien moeten deze zo gekozen worden dat kan worden gemeten in welke mate de onderwijsdoelstellingen worden bereikt. De criteriumvaliditeit van de lange toetsvormen is niet overtuigend aangetoond: ze correleren zwak zowel met klinische als met theoretische examens. Zowel bij PMP's als bij MEQ's leverden onderzoeken met factoranalyses of vergelijkingen van groepen met verschillend opleidingsniveau aanwijzingen op voor de begripsvaliditeit.

De korte toetsvormen bevatten slechts één (SIMP) of enkele vragen per probleem (Q4), zodat in een examen veel problemen kunnen worden aangesneden, wat de betrouwbaarheid en inhoudsvaliditeit ten goede zou moeten komen. De bevindingen omtrent de betrouwbaarheid van de SIMP's zijn moeilijk te interpreteren: het gaat om drie verschillende indicatoren die niet in overeenstemming zijn met elkaar en met de verwachting dat de betrouwbaarheid stijgt naarmate het aantal casus toeneemt. Alleen voor de criteriumvaliditeit werden aanwijzingen gegeven, namelijk een verband met andere klinische scores. Deze toets staat of valt met de veronderstelling dat alleen de initiële diagnostische hypothese er toe doet. Hoewel is gebleken dat bij experts het hypothetico-deductief redeneren minder belangrijk is dan patroonherkenning, blijft ook dan de eerste beslissing de belangrijkste. Geneeskundestudenten zijn echter nog geen experts. Zelfs als ze de diagnose correct kunnen stellen, betekent dit nog niet dat ze het klinisch onderzoek goed uitvoeren, de juiste laboratoriumtesten aanvragen of de juiste beleidsbeslissing nemen. Met de Q4-toets is het wel de bedoeling vragen te stellen over de sleutelbeslissingen bij het oplossen van een probleem. De bevindingen wijzen er ook op dat de key features inderdaad sleutelelementen betreffen, althans volgens experts (inhoudsvaliditeit). Of de key features inderdaad de elementen zijn die bij de studenten het vaakst aanleiding geven tot fouten, is niet aangetoond en moeilijker te bewijzen. Verder dient de kwaliteit van de Q4 nog bewezen te worden: betrouwbaarheid, criterium- en begripsvaliditeit zijn niet bekend.

Ten slotte

PMP's en MEQ's hebben een tijdje een slechte naam gehad. De problematiek van

de casusspecificiteit geldt echter voor alle toetsvormen. Voor betrouwbare uitspraken zijn daarom veel casus noodzakelijk met als gevolg lange toetstijden. De key features approach (SIMP, Q4) biedt interessante mogelijkheden, maar er is nog weinig onderzoek over deze benadering gepubliceerd in internationale tijdschriften. Voor de toekomst mag veel verwacht worden van 'case-based simulations' die via de computer worden afgenomen.[8 53] Een aantal problemen bij de afname van schriftelijke simulaties en waarschijnlijk ook bij het nakijken lijkt oplosbaar te zijn.

Literatuur

1. Luijk SJ van, Vleuten CPM van der. Het stationsexamen. In: Metz JCM, Scherpbier AJJA, Vleuten CPM van der, redactie. Medisch onderwijs in de praktijk. Assen: Van Gorcum; 1995. p. 202-7.
2. Juul DH, Noe MJ, Nerenberg RL. A factor analytic study of branching patient management problems. Med Educ 1979;13:199-203.
3. Glaser R, Damrin D, Gardner FM. The tab item: a technique for the measurement of proficiency in diagnostic problem-solving. In: Lumsdaine AA, Glaser R, redactie. Teaching machines and programmed learning: a source book. Washington: National Education Association; 1960. p. 275-85.
4. Rimoldi HJA. A technique for the study of problem solving. Educ Psychol Meas 1955;15:450-61.
5. Hubbard JP. Programmed testing. In: Examinations and their role in evaluation of medical education and qualification for practice. Philadelphia: National Board of Medical Examiners; 1964. p. 102-14.
6. Vleuten CPM van der, Newble DI, Case S, Holsgrove G, McCann B, McRae C, Saunders N. Methods of assessment in certification. In: Newble DI, Jolly B, Wakeford R, redactie. The certification and recertification of doctors: issues in the assessment of clinical competence. Cambridge: Cambridge University Press; 1994. p. 105-25.
7. Taylor WC, Grace M, Taylor TR, Fincham SM, Skakun EN. The use of computerized patient management problems in a certifying examination. Med Educ 1976;10:179-82.
8. Clyman SG, Melnick DE, Clauser BE. Computer-based case simulations by the National Board of Medical Examiners of the United States. In: Jong PGM de, Bloemendaal PM, redactie. Toetsing in de basisopleiding en het postacademisch onderwijs. Leiden: Boerhave Commissie voor Postacademisch Onderwijs in de Geneeskunde, Rijksuniversteit Leiden; 1997. p. 133-47.
9. Andrew BJ. An approach to the construction of simulated exercises in clinical problem-solving. J Med Educ 1972;47:952-8.
10. Harden RM. Preparation and presentation of patient-management problems (PMPs). Med Educ 1983;17:256-76.
11. Marshall JR, Fabb WE. The construction of patient-management problems. Med Educ 1983;15:126-35.
12. McGuire CH, Babbott D. Simulation technique in the measurement of problem-solving skills. J Educ Meas 1967;4:1-10.
13. Newble DI, Hoare J, Baxter A. Patient management problems: issues of validity. Med Educ 1982;16:137-42.
14. Goran MJ, Williamson JW, Gonnella JS. The validity of patient management problems. J Med Educ 1973;48:171-7.
15. Jones TV, Gerrity MS, Earp JA. Written case simulations: do they predict physicians' behavior? J Clin Epidemiol 1990;43:805-15.
16. McCarthy WH, Gonnella JS. The simulated patient management problem: a technique for evaluating and teaching clinical competence. Br J Med Educ 1967;1:348-52.
17. Wolf FM, Allen NP, Cassidy JT, Maxim BR, Davis WK. A criterion-referenced approach to measuring medical problem solving: validity of patient management problems. Eval Health Prof 1985;8:223-40.
18. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. Med Educ 1985;19:238-47.
19. Norcini JJ, Meskauskas JA, Langdon LO, Webster GD. An evaluation of a computer simulation in the assessment of physician competence. Eval Health Prof 1986;9:286-304.
20. McLaughlin FE, Carr JW, Delucchi KL. Measurement properties of clinical simulation tests: hypertension and chronic obstructive pulmonary disease. Nurs Res 1981;30:5-9.
21. Colliver JA, Markwell SJ, Vu NV, Barrows HS. Case specificity of standardized-patient examinations: consistency of performance on components of clinical competence within and between cases. Eval Health Prof 1990;13:252-61.
22. Marshall J. Assessment of problem-solving ability. Med Educ 1977;11:329-34.
23. Berner ES, Bligh TJ, Guerin RO. An indication for a process dimension in medical problem-solving. Med Educ 1977;11:324-8.
24. Marshall JR, Fleming P, Heffernan M, Kasch S. Pilot study on use of PMPs. Med Educ 1982;16:365-6.

25. Graaff E de, Post GJ, Drop MJ. Validation of a new measure of clinical problem-solving. Med Educ 1987;21:213-8.
26. Al-Chalabi TS, Al-Na'Ama MR, Al-Thamery DM, Alkafajei AMB, Mustafa GY, Joseph G, Sugathan TN. Critical performance analysis of rotating resident doctors in Iraq. Med Educ 1983;17:378-84.
27. Ramsey PG, Shannon NF, Fleming L, Wenrich M, Peckham PD, Dale DC. Use of objective examinations in medicine clerkships: ten-year experience. Am J Med 1986;81:669-74.
28. Joorabchi B, Chawhan AR. Multiple choice questions: the debate goes on. Br J Med Educ 1975;9:275-80.
29. Easterling WE. In-training examinations for residents in obstetrics and gynecology, 1975 to 1978. Am J Obstet Gynecol 1979;133:733-41.
30. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. Ann Intern Med 1989;110:719-26.
31. Palchik NS, Wolf FM, Cassidy JT, Ike RW, Davis WK. Comparing information-gathering strategies of medical students and physicians in diagnosing simulated medical cases. Acad Med 1990;65:107-13.
32. Gruppen LD, Wolf FM. Expertise, efficiency, and the construct validity of patient management problems. J Med Educ 1985;60:878-80.
33. Papp KK, Williams SD, Goldman MH. Relationship between type of surgical clerkship, order of completion, and achievement on patient management problems. Surgery 1984;96:102-7.
34. Mazzuca SA, Cohen SJ, Clark CM. Evaluating clinical knowledge across years of medical training. J Med Educ 1981;56:83-90.
35. Wolf FM. Validity of patient-management problems re-examined. Med Educ 1984;18:222-5.
36. Norman GR, Feightner JW. A comparison of behaviour on simulated patients and patient management problems. Med Educ 1981;15:26-32.
37. McCarthy WH. An assessment of the infuence of cueing items in objective examinations. J Med Educ 1966;41:263-6.
38. White RE, Quimby BB, Skipper BJ, Webster GD. Cost of residents' decisions on actual patients and in simulated encounters. J Med Educ1984;59:833-5.
39. The Board of Censors of the Royal College of General Practitioners. The modified essay question. Proc R Coll Gen Pract 1971;21:373-85.
40. Feletti GI, Engel CE. The modified essay question for testing problem-solving skills. Med J Aust 1980;1:79-80.
41. Walker JH, Stanley IM, Venables TL, Gambrill EC, Hodgkin GK. The MRCGP examination and its methods. IV: MEQ paper. J R Coll Gen Pract 1983;33:804-8.
42. Graaff E de. Essayvragen. In: Metz JCM, Scherpbier AJJA, Vleuten CPM van der, redactie. Medisch onderwijs in de praktijk. Assen: Van Gorcum; 1995. p. 182-8.
43. Feletti GI. Reliability and validity studies on modified essay questions. J Med Educ 1980;55:933-41.
44. Rabinowitz HK. The modified essay question: an evaluation of its use in a family medicine clerkship. Med Educ 1987;21:114-8.
45. Norman GR, Smith EK, Powles AC, Rooney PJ, Henry NL, Dodd PE. Factors underlying performance on written tests of knowledge. Med Educ 1987;21:297-304.
46. Irwin WG, Bamber JH. The cognitive structure of the modified essay question. Med Educ 1982; 16:326-31.
47. Irwin WG, Bamber JH. An evaluation of a course for undergraduate teaching of general practice. Med Educ 1978;12:20-5.
48. Feletti GI, Smith EK. Modified essay questions: are they worth the effort? Med Educ 1986;20:126-32.
49. Page G, Bordage G. The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. Acad Med 1995;70:104-10.
50. Bordage G, Page G. An alternative approach to PMPs: the key features concept. In: Hart I, Harden R, redactie. Further developments in assessing clinical competence. Montreal: Can-Heal Publications; 1987. p. 57-75.
51. Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. Acad Med 1995;70:194-201.
52. Bordage G, Brailovsky C, Carretier H, Page G. Content validation of key features on a national examination of clinical decision-making skills. Acad Med 1995;70:276-81.
53. Schuwirth L. An approach to the assessment of medical problem solving: computerised case-based testing [proefschrift]. Maastricht: Datawyse Universitaire Pers Maastricht; 1999.

De auteurs:

J. Beullens is psycholoog en wetenschappelijk medewerker aan de Onderwijskundige Dienst van de Faculteit Geneeskunde van de KU Leuven.
Prof. dr. H. Jaspaert is pedagoog en hoofddocent en werkt in de Onderwijskundige Dienst van de Faculteit Geneeskunde van de KU Leuven.

Correspondentieadres:

Johan Beullens, Onderwijskundige Dienst Faculteit Geneeskunde, KU Leuven, Minderbroedersstraat 17, B-3000 Leuven.
E-mail: Johan.Beullens@med.kuleuven.ac.be.

Summary

Clinical problem-solving skills are not easy to assess. This article present the results of a Medline search of the literature on paper-and-pencil tests of clinical competence, specifically patient management problems (PMP), modified essay questions (MEQ) and the key features approach. These tests are more realistic than oral examinations, essay questions or multiple-choice items and they provide an incentive for students to concentrate on the application rather than on the reproduction of knowledge in preparation for a test. PMPs and MEQs have adequate reliability with testing times of about one day. There are indications of construct validity of both test formats. The key features approach may allow shorter testing times, but its psychometric qualities still need to be proven. In the near future, computerised case-based testing may prove to be a promising approach. (Beullens J, Jaspaert H. Paper-and-pencil test formats for the assessment of clinical problem-solving skills: an overview. Dutch Journal of Medical Education 2000;19(4):129-42)