Bepaling van het vermogen om te beslissen in onzekerheid met de Script Concordance Test methode

S. P. J. Ramaekers, W. D. J. Kremer, A. Pilot, P. van Beukelen en J. van Keulen

Samenvatting

Veel problemen uit de dagelijkse praktijk vergen beslissingen terwijl de beschikbare informatie onvolledig of onduidelijk is. De inbedding van deze onzekerheden in toetsen staat op gespannen voet met psychometrische kwaliteitseisen. Uitgaande van het Script Concordance Test (SCT)-format werd hiervoor een test ontwikkeld in de diergeneeskunde. Deze test is bij 148 studenten tweemaal afgenomen, aan het begin en eind van een studieonderdeel dat is gericht op het leren oplossen van klinische problemen. Hun oordelen en beslissingen in realistische praktijksituaties werden afgezet tegen die van een groep van 28 ervaren practici. Studentscores tonen bij de hertest een significante toename van gemiddeld 4,59 punt. Individueel correleren scores positief (r = 0,65, p < 0,001) en is het effect groot (d = 0,89). Uit de analyses van casus en resultaten wordt geconcludeerd dat de SCT een bruikbaar instrument vormt voor toetsing van het probleemoplossen in onzekerheid, dat bij grote groepen studenten kan worden gebruikt, zonder een onaanvaardbare overbelasting van patienten.

1 Inleiding

Oordelen en beslissen in situaties van relatieve onzekerheid, op basis van slechts beperkte informatie, of onder tijdsdruk, is voor velen – zeker in hogere functies en beroepen – onderdeel van hun professionele praktijk (Jonassen, 2004). Ervaren professionals doen dit onder andere door *tacit* kennis van contextuele aspecten mee te wegen in hun beoordeling van mogelijke oplossingen en onzekerheden (Thornton, 2006). Zij hebben bovendien kennisstructuren ontwikkeld die passen bij de samenhang waarin aspecten van problemen zich in de realiteit voordoen (Eraut, 2004). Die organisatie van kennis stelt hen in staat om op basis van beschikbare informatie snel eerste hypotheses over een probleem te vormen, patronen te herkennen of mogelijke oplossingsscenario's met relatief geringe cognitieve belasting te doordenken (Kirschner, 2002).

Om studenten hier adequaat op voor te bereiden wordt in opleidingen toenemend aandacht besteed aan authentieke vraagstukken en situaties. Het gaat dan om problemen en omstandigheden die in hoge mate realistisch zijn voor de professionele praktijk en dezelfde (mentale) activiteiten en processen oproepen. Voor het monitoren van de ontwikkeling die studenten daarin doormaken, als ook voor onderzoek naar de effectiviteit van dit onderwijs, is de beschikbaarheid van instrumentarium waarmee het probleemoplossend vermogen van studenten in realistische omstandigheden in kaart wordt gebracht, van bijzonder belang. Onzekerheden en ambiguïteit die deel uit kunnen maken van authentieke situaties zijn vanuit het perspectief van validiteit weliswaar belangrijk (Van der Vleuten & Schuwirth, 2005), tegelijkertijd worden deze aspecten in de toetsing veelal gezien als een bron van 'ruis' die de betrouwbaarheid bedreigt. Immers, hoe kunnen de antwoorden van studenten op toetsvragen worden beoordeeld wanneer het 'goede' antwoord niet met zekerheid gegeven kan worden? En hoe kan eventuele progressie van studenten dan betrouwbaar vastgesteld worden?

Deze studie richt zich op het verkrijgen van inzicht in de psychometrische kwaliteiten van een test waarmee progressie in de ontwikkeling van competentie op terrein van het professioneel oordelen en beslissen in onzekerheid moet kunnen worden bepaald. De test is ontwikkeld voor een omvangrijk studieonderdeel van de opleiding diergeneeskunde (Utrecht) dat is gericht op het leren oplossen van klinische problemen. De gebruikte methode is ontleent aan het Script

Concordance Test (SCT)-format, dat eind jaren negentig door Charlin en anderen (Charlin, Boshuizen, Custers, & Feltovich, 2007; Charlin, Brailovsky, Leduc, & Blouin, 1998a; Charlin & Van der Vleuten, 2004) werd ontwikkeld om het redeneren en beslissen in realistische klinische situaties te kunnen meten.

2 Theoretisch kader

Het format van de SCT is gestoeld op empirisch onderzoek en theorieën over de wijze waarop artsen patiëntproblemen analyseren, oordelen vormen en beslissingen nemen. Dit oplossen van klinische problemen en de ontwikkeling die artsen daarin doormaken, is al sinds het einde van de jaren vijftig een object van onderzoek en theorievorming (Ledley & Lusted, 1959; Rimoldi, 1961). Aanvankelijk werd het klinisch probleemoplossen vooral gezien als een cyclisch proces van het genereren en toetsen van hypothesen, totdat de juiste diagnose kan worden gesteld en alternatieven zijn uitgesloten. Verwacht werd dat inzicht in de wijze waarop experts dit doen, duidelijkheid zou verschaffen over de aanpak waarin beginners zouden moeten worden getraind. Toen bleek dat verschillen tussen beginners en experts slechts beperkt verklaard konden worden op basis van hun redeneerpatronen, en experts bovendien vaak nauwelijks of geen hypothesen formuleerden of expliciet testten, werd de focus van het onderzoek verlegd naar de verschillen in onderliggende kennisstructuren en de ontwikkeling daarvan (Elstein, Schulman, & Sprafka, 1978; Neufeld, Norman, Feightner, & Barrows, 1981; Norman & Schmidt, 1992).

De notie van 'ziektescripts' gaat ervan uit dat ervaren practici beschikken over kennisnetwerken, 'scripts', waarin allerlei aspecten van ziektes en aandoeningen in een voor de praktijk relevante samenhang zijn ondergebracht. Deze scripts worden met name gevoed door eigen praktijkervaringen en zijn rijkelijk voorzien van details over specifieke patiënten en concrete situaties (Boshuizen & Schmidt, 1992; Custers, Boshuizen, & Schmidt, 1996; Norman & Schmidt, 1992). Onderliggende kennis van medisch-biologische mechanismen of principes, evenals de bewuste redenaties die ten grondslag liggen aan oordelen, interpretaties en keuzes raken in de ontwikkeling van deze scripts op de achtergrond (*embedded*) en worden alleen nog geactiveerd wanneer daar expliciet de aandacht op wordt gericht (Rikers, Schmidt, & Moulaert, 2005). Naarmate practici meer ervaring hebben opgedaan gaan vergelijking met eerdere patiënten, herkenning van patronen en van zogenaamde *enabling conditions* in toenemende mate het klinisch probleemoplossen domineren (Norman, 2005). Het beoordelen van problemen en beslissen over oplossingen zou aldus niet alleen sneller verlopen maar ook beduidend minder cognitief belastend zijn (Custers et al., 1996; Kirschner, 2002).

De erkenning van de complexiteit van klinische problemen, de omstandigheden waaronder ze moeten worden opgelost en van de beperkingen in de wijze waarop mensen met veel verschillende factoren tegelijkertijd of onzekerheden omgaan (Tversky & Kahneman, 1974), waren aanleiding voor nader onderzoek naar het hanteren van onzekerheden bij het beslissen. *Decision analysis research* is gericht op het optimaliseren van redeneringen in situaties van onzekerheid (Balla & Edwards, 1986; Lilford, Pauker, Braunholtz, & Chard, 1998). Daarbij wordt, op basis van kwantitatieve modellen, gestreefd naar standaarden voor een optimale (expert) aanpak van klinische problemen. Als methode voor analyse achteraf van genomen beslissingen inclusief mogelijke bronnen van bias wordt de waarde van deze benadering breed erkend (Hunink, 2001; Sarasin, 2001). Kritiek betreft met name de beperkte bruikbaarheid in een klinische setting: patiëntproblemen zijn specifiek en vergen maatwerk, de vele schattingen die de methode vraagt, blijken ook voor ervaren practici nauwelijks gelijktijdig met de uitvoering van patiëntonderzoek te maken, en 'standaarden' in de vorm van algoritmen en protocollen laten onvoldoende ruimte voor normale variatie en meeweging van andersoortige factoren zoals sociaalmaatschappelijke omstandigheden (Berg, 1997; Elstein, 2004).

In de huidige opvattingen over het klinisch probleemoplossen is sprake van een

zekere consensus dat het oplossingsproces in hoge mate domein- en contextspecifiek is en de transfer van het ene naar het andere probleem gering. Bij herhaling hebben studies laten zien dat succes in het oplossen van het ene probleem een weinig accurate voorspeller is voor succes bij een volgende (Elstein et al., 1978; Norman, 2005). Klinische problemen worden bovendien beschouwd als kennisintensief en het adequaat oplossen ervan niet zozeer als een kwestie van één optimale, generieke strategie, als wel van het beschikken over kennis van de veelheid aan relevante aspecten, die georganiseerd is in een op de praktijk toegesneden structuur (Elstein & Schwarz, 2002). Ervaren practici bewandelen in vergelijkbare situaties, afhankelijk van specifieke omstandigheden, vaker andere wegen om tot oplossingen te komen; ook bij dezelfde uitkomsten (diagnose, therapie, etc.) worden door verschillende practici soms andere benaderingen, redeneringen en afwegingen gevolgd (Grant & Marsden, 1988; Norman & Brooks, 1997). Wanneer het een probleem betreft waarmee de betreffende practicus veel ervaring heeft, is oplossen vaak gebaseerd op herkenning van patronen en 'weten', en nauwelijks op (bewust) beredeneren, deduceren of toetsen (Forde, 1998; Norman, Young, & Brooks, 2007).

Omstandigheden die bijdragen aan de complexiteit van het oplossen van klinische problemen zijn dat de beslissingen vaak genomen moeten worden op basis van beperkte gegevens (om de belasting van de patiënt en de kosten van het onderzoek binnen de perken te houden), binnen een beperkt tijdsbestek (om verergering en andere risico's te voorkomen), terwijl prognoses (over het verdere beloop, zonder en met interventies) soms weinig nauwkeurig voorspelbaar zijn (Eraut, 2004; Norman, 2005). Van de beschikbare informatie is niet altijd duidelijk hoe betrouwbaar die is en moet deze onzekerheid worden meegewogen (Forde, 1998). De effecten van onzekerheden worden snel groter naarmate beslissingen zijn geënt op langere reeksen van onderling afhankelijke beoordelingen (Eraut, 2004). Aangenomen wordt dat al deze omstandigheden mede verklaren waarom de oordelen en keuzes van experts in vergelijkbare situaties behoorlijk kunnen variëren (Eddy, 1990).

Ontwikkelingen in het beoordelen en toetsen van het probleemoplossend vermogen zijn sinds de negentiger jaren gedomineerd door het streven naar meer authentieke toetsvormen. Dit zijn vormen van assessment aan de hand van realistische vraagstukken en omstandigheden, die overeenkomstige cognitieve processen en activiteiten oproepen als in de praktijk van een beroep of functie plaatsvinden (Linn, Baker, & Dunbar, 1991; Newmann & Archbald, 1992; Swanson, Norman, & Linn, 1995; Van der Vleuten, 1996). De validiteit van toetsvormen die zijn geënt op gesloten vraagstukken met eenduidig goede en foute antwoorden voor het bepalen van competentie in het oplossen van complexere problemen, wordt betwist (Downing, 2003; Epstein & Hundert, 2002; Segers, Dochy, & De Corte, 1999; Van der Vleuten, 1996; Van der Vleuten & Schuwirth, 2005). Kritiek heeft betrekking op de beperkte representativiteit van de vraagstukken en de omstandigheden in beoordelingssituaties ten opzichte van de praktijk, en op de mate waarin de toetsvormen daadwerkelijk aanspraak doen op de cognitieve processen die het fundament vormen van deze competentie.

Specifiek voor de ontwikkeling van het probleemoplossend vermogen in de fase van coschappen of klinische stages kwam in eerder onderzoek naar voren dat – ondanks de toename van de praktijkervaring – de resultaten op conventionele tests waarmee het vermogen om casus op te lossen werd gemeten, niet verbeterden (Boshuizen, 2003; Patel, Arocha, & Zhang, 2005). Verklaringen voor dit fenomeen, dat bekend staat als de *intermediate dip*, worden gezocht zowel in veranderingen in de organisatie van kennis die studenten in deze fase ondergaan, als in tekortkomingen in de validiteit van gebruikte tests. De tests zouden niet zozeer het probleemoplossend vermogen en redeneren meten, als wel de voor die casus relevante parate kennis (Schmidt & Boshuizen, 1993).

Tegen deze achtergrond werd het SCT-format ontwikkeld waarmee het competentieniveau op terrein van klinisch redeneren en probleemoplossen zou kunnen worden vastgesteld. SCT's worden verondersteld aan-

spraak te doen op het vermogen om beperkte patiëntgegevens goed te kunnen interpreteren (Sibert et al., 2002a). Ze zouden indicatief zijn voor de omvang en rijkdom in detail van aanwezige ziektescripts (Charlin, Roy, Brailovsky, Goulet, & Van der Vleuten, 2000) en zicht geven op het vermogen om professionele oordelen te vormen en adequate beslissingen te nemen ondanks onzekerheden (Charlin & Van der Vleuten, 2004). Uitgaande van levensechte problemen en situaties wordt deelnemers gevraagd om in een groot aantal casus te oordelen over diverse kwesties. De oordelen en beslissingen van een groep praktiserende experts bij dezelfde casus vormen hierbij de referentie voor weging van de antwoorden van deelnemers. De mate van overeenstemming met de experts geldt daarbij als indicator voor de mate van gevorderdheid van de deelnemer.

Sinds het eerste gebruik ervan in 1998 zijn er binnen diverse specialismen in de humane geneeskunde versies van een SCT ontwikkeld (Brazeau-Lamontagne, Charlin, Gagnon, Samson, & Van der Vleuten, 2004; Caire, Sol, Charlin, Isidori, & Moreau, 2004; Charlin et al., 1998b; Meterissian, Zabolotny, Gagnon, & Charlin, 2007; Sibert et al. 2002b). Nader onderzocht werden de effecten van variaties in testformats (Sibert et al., 2006), van verschillende momenten van afname, van mogelijke scoremodellen (Bland, Kreiter, & Gordon, 2005; Charlin, Desaulniers, Gagnon, Blouin, & Van der Vleuten, 2002) en van de samenstelling van het referentenpanel (Gagnon, Charlin, Coletti, Sauve, & Van der Vleuten, 2005; Nendaz et al., 2004). Testresultaten werden vergeleken tussen deelnemers op verschillende niveaus van klinische ervaring, en tussen deelnemers uit verschillende onderwijssystemen (Sibert et al., 2002a). Resultaten op de test werden gerelateerd aan andere aspecten en indicatoren van klinische competentie (Brailovsky, Charlin, Beausoleil, Cote, & Van der Vleuten, 2001; Brazeau-Lamontagne et al., 2004; Gagnon et al., 2006). Het achterwege blijven van een *intermediate dip* wordt gezien als bevestiging van de vooronderstelling dat deze tests meer aanspraak doen op redeneervermogen dan met de gebruikelijke (theoretische) toetsmethoden gerealiseerd kan worden (Charlin et al., 1998a; Charlin et al., 1998b).

De belangrijke verschillen van deze studie met eerder onderzoek betreffen de (preklinische) fase in de opleiding en de breedte van het domein waarvoor de test wordt ontwikkeld, en het herhaald gebruik bij eenzelfde cohort studenten. Afname van de test in de preklinische fase, dat wil zeggen wanneer de studenten nog weinig of geen praktijkstages (co-schappen) hebben gelopen, is vanuit theoretisch oogpunt omstreden; studenten zouden dan nog niet beschikken over geïnternaliseerde ziektescripts waarmee zij de informatie(patronen) en onzekerheden in een casus adequaat kunnen wegen. De test is bovendien bedoeld voor de gemeenschappelijke kern van alle afstudeerrichtingen binnen een opleiding en beperkt zich niet tot een goed afgebakend specialisme. Dit roept vragen op over de mogelijkheden om de test zowel representatief als uitvoerbaar te houden. En in eerdere studies, tenslotte, werden verschillende populaties (cohorten studenten op verschillende niveaus) met elkaar vergeleken. Deze studie volgt eenzelfde cohort in de tijd, waardoor mogelijk een onbetrouwbaarheid wordt geïntroduceerd: dezelfde test wordt bij de studenten in dit onderzoek twee keer afgenomen.

Gegeven deze achtergronden zijn de vragen waar we in deze studie een antwoord op zoeken:

1) Kan op basis van het SCT-format een test worden ontwikkeld die valide is voor het brede domein van de diergeneeskunde en de preklinische fase van de studie?
2) Is deze test voldoende betrouwbaar? Hierbij wordt de eis gesteld dat de test wél ambiguïteit in de klinische problemen en dus beslissen in onzekerheid omvat, maar niet tot onzekerheid over het niveau van de studenten mag leiden. In dit kader is ook gekeken naar mogelijke effecten van de verschillende manieren om de expertantwoorden te waarderen.
3) Is een dergelijke test voldoende sensitief om binnen het normale verloop van een studieonderdeel progressie te meten wat betreft het vermogen problemen op te lossen?
4) Is een dergelijke test uitvoerbaar in de

normale onderwijspraktijk, dus met grote aantallen (meer dan honderd) studenten?

Een positief antwoord op de bovenstaande vragen betekent niet alleen dat de ontwikkelde test voldoet aan de vereisten van kwaliteit; het realiseren van een voor deze context kwalitatief volwaardig instrument vormt ook een indicatie dat dit type tests in een eerdere fase in de ontwikkeling van studenten en voor een breder domein bruikbaar is, dan tot nog toe op theoretische gronden wordt verondersteld. Duidelijkheid over de kwaliteit van de test is bovendien van belang voor de mogelijke inzetbaarheid ervan als instrument in wetenschappelijk onderzoek naar die ontwikkeling van het (klinisch) redeneren en beslissen bij studenten.

3 Methode

3.1 Materialen

In een SCT worden klinische problemen en situaties beschreven in de vorm van een casus met bijbehorende vragen. Tabel 1 bevat een voorbeeld casusvignette met informatie over de casus. Deze informatie is relevant voor de vragen en vergelijkbaar met een 'script'. De gepresenteerde informatie geeft geen totaalbeeld van de onderliggende praktijksituatie; het is de eerste informatie waarover een dierenarts in de praktijk beschikt. Onrealistische beperkingen in de informatie en onzekerheden worden zorgvuldig vermeden. In principe wordt die informatie gegeven, die in de realiteit over een klinische probleem bekend is op het moment waarop de overweging of beslissing zich voordoet.

De bij een casus behorende vragen zijn onderling onafhankelijk. Ze bevatten een aanvullend casusgegeven en een te toetsen hypothese of suggestie voor de verdere aanpak. Deelnemers wordt gevraagd te beoordelen in hoeverre de gesuggereerde hypothese of aanpak bevestigd of verworpen moet worden op basis van de totale casusinformatie in het vignette en de aanvulling.

Ontwikkeling van de SCT

Aan de hand van epidemiologische gegevens over frequent voorkomende problemen in de eerstelijns praktijk is eerst een toetsmatrijs samengesteld met een representatieve selectie van klinische problemen. Op basis hiervan zijn casusvignettes en vraagitems geconstrueerd met specifieke informatie over de te beoordelen situaties. Bij de keuze van items is gestreefd naar selectie van die casusingre-

Tabel 1

Voorbeeld casusvignette en vraagitems

Casus: Voor de tweede keer binnen één maand wordt u deze 9-jarige merrie aangeboden i.v.m. koliekverschijnselen. Bij de eerste consultatie leek sprake van obstipatie in de linker ventrale colon, die u met pijnstillers en parafine laxatie heeft behandeld. De verstopping leek na 2 dagen volledig opgeheven; het paard bleef niettemin last houden van slappe mest en onvoldoende eetlust. Sinds gisteren lijkt opnieuw sprake van koliek (krabben met de voorbenen, kijken naar de buik, afwisselend liggen-staan).
Uit uw onderzoek komen o.a. naar voren: enige abdominale défence musculaire en hoorbare borborygmi. Rectale exploratie: geen bijzonderheden. Algemeen: het paard is onrustig, transpireert licht, pols 52 /min , temp 38,2°, heeft gele oogslijmvliezen en zit slecht in het haar.

Stel, u overweegt dat hier sprake is van:	en er blijkt in uw onderzoek:	dan is deze hypothese: -2	-1	0	+1	+2
a. een strangulerende obstructie van het colon	Geen recente rantsoenwisselingen m.u.v. enkele dagen vasten na de vorige koliekbehandeling	0	0	0	0	0
-2 = zeer onwaarschijnlijk -1 = minder waarschijnlijk 0 = niet minder noch meer waarschijnlijk +1 = meer waarschijnlijk; +2 = zeer waarschijnlijk.						

Stel, u overweegt als aanpak voor behandeling:	en er blijkt bovendien sprake van:	dan is deze aanpak: -2	-1	0	+1	+2
d. buikpunctie	Toename van koliek na het rijden	0	0	0	0	0
-2 = gecontraïndiceerd -1 = minder wenselijk 0 = niet minder noch meer zinvol +1 = wenselijk +2 = noodzakelijk.						

diënten, die in de professionele praktijk ten grondslag liggen aan overwegingen en keuzen (de zogenaamde key features) op het terrein van diagnostiek en diergeneeskundige interventies. Deze zijn overwegend van medisch-biologische aard maar kunnen ook betrekking hebben op andere aspecten zoals de beschikbaarheid van bepaalde faciliteiten, opvattingen en wensen van de eigenaar, het tijdstip waarop of de omstandigheden waaronder het probleem zich voordoet, een spoedeisend karakter, et cetera. De representativiteit en de formulering zijn gecontroleerd door drie onafhankelijke inhoudelijk deskundigen die niet betrokken waren bij de constructie van de casuïstiek.

Om na te gaan of onbekendheid met dit type casusvignettes, het specifieke vraagformat van de SCT en tendenties naar het midden (het 'indifferente' antwoord) van invloed zouden zijn op de scores van de studenten, werden conceptversies van de test in drie sessies via een think-aloud protocol afgenomen bij 4e-jaars studenten uit het voorafgaande cohort. Ook is nagegaan of de beoogde cognitivieve processen (onder andere informatie combineren, hypotheses toetsen aan de hand van de informatie, wegen van waarschijnlijkheden) inderdaad werden opgeroepen. De resultaten hiervan hebben niet geleid tot aanpassingen in het format van de casus, vraagitems of antwoorden; wél gaven ze aanleiding tot aanpassingen in specifieke formuleringen en aanvullingen op de bijbehorende testinstructie.

In eerdere studies (Charlin, Tardif, & Boshuizen, 2000) bleek dat een voor een medisch subdomein representatieve SCT tenminste 50 - 60 vraagitems nodig heeft voor een betrouwbaarheid (Cronbach's α) in de orde van 0,80. Rekening houdend met de breedte van het domein is uiteindelijk gekozen voor een SCT bestaande uit 30 casus met 120 vraagitems. Deze casus hebben betrekking op de gebruikelijke diersectoren: gezelschapsdieren, paard en landbouwhuisdieren. De items omvatten de diagnostiek, de aanpak van het patiëntonderzoek, therapeutische en preventieve interventies, en de prognose.

Om de test te kunnen gebruiken voor verschillende deskundigheidsniveaus (van begin 4e jaars tot en met ruim ervaren practici) is bovendien enige spreiding in vraagniveaus aangebracht. Van alle items in de test is 1/6 zodanig gekozen dat zij, naar verwachting van de samenstellers, ook bij de 4e jaars studenten slechts beperkte spreiding zullen opleveren. Eveneens is 1/6 van de items zodanig geformuleerd dat zij naar verwachting, ook bij de referentiegroep van ervaren practici, aanleiding zijn voor een grotere variatie in responses. Tabel 2 toont de verdeelsleutels voor de casus en items, aanvullend op de lijst van frequent voorkomende aandoeningen in de eerstelijns praktijk.

Context en deelnemers

Deze diergeneeskundige variant van de test (SCT) is ontwikkeld als instrument om na te gaan welke progressie studenten maken in de zogenaamde Klinische lessen. Dit studieonderdeel vormt de hoofdmoot van het 4e jaar van de opleiding diergeneeskunde en is specifiek gericht op het leren klinisch redeneren en probleemoplossen. In de vorm van klinische practica, werkgroepen en demonstraties worden wekelijks een aantal patiënten onderzocht, besproken c.q. bekeken. Inhoudelijk vervullen de klinische lessen een brugfunctie tussen de preklinische vakken in de eerste drie jaar van de studie en de klinische fase met co-schappen (jaar 5 en 6). De SCT werd

Tabel 2

Verdeelsleutels voor de casus en items in de SCT diergeneeskunde

Diersoort / - sector		Oordeel / beslissing		Beoogd vraagniveau	
- Gezelschapsdieren (hond, kat, etc)	9	- Diagnose + prognose	52	- Instroomniveau 4e jaar	20
- Landbouwhuisdieren (varken, pluimvee)	8	- Behandeling	35	- Intermediate	80
- Landbouwhuisdieren (runderen)	9	- Preventie	15	- expert	20
- Paard	8	- Aanpak onderzoek	18		

bij dezelfde studenten twee keer afgenomen. De afname betrof een onderwijsexperiment en verving geen ander studieonderdeel. Studenten namen deel op vrijwillige basis en konden in principe afzien van deelname.

Om de invloed van onbekendheid van de studenten met het format van de test en de casus op de resultaten in de eerste testafname te minimaliseren, vond de eerste afname (in december 2007) plaats nadat hen gelegenheid was geboden om met enkele (maximaal zeven) casusbesprekingen in klinische werkgroepen te oefenen. De tweede afname (in juni 2008) vond plaats na afsluiting van alle klinische lessen aan het eind van het studiejaar.

Referentiegroep

Aangezien de SCT uitgaat van open casus over specifieke praktijksituaties, wordt de scoresleutel van een SCT-test bepaald via een referentiegroep van ervaren practici. De inclusiecriteria voor de referentiegroep waren: praktiserend dierenarts, werkzaam in 1e-lijns praktijk, ten minste 10 jaar ervaring, géén docent, een goede reputatie op basis van (tweedelijns) kliniekcontacten over verwezen patiënten en als deskundig benoemd door ten minste twee collega dierenartsen. Op basis van de bevindingen in eerdere SCT-testen (Gagnon et al., 2005; Nendaz et al., 2004) werd uitgegaan van ten minste tien referenten per diersector. In totaal 35 referenten die aan de inclusiecriteria voldeden ontvingen vooraf een uitnodiging en informatie over de bedoeling en achtergronden van de test. Referenten hebben expliciet ingestemd met deelname, alvorens zij de test zelf ontvingen.

Van de uitgenodigde referenten stemden 32 vooraf in met deelname; de redenen om niet te participeren bleken te maken te hebben met privé-omstandigheden. Praktische en triviale omstandigheden verhinderden verder in enkele gevallen daadwerkelijke participatie; uiteindelijk werd van 28 referenten de ingevulde test retour ontvangen. Voor bepaling van de individuele scores van de studenten en verdere kwantitatieve analyse van de test werd op basis van de antwoorden van de practici een antwoordsleutel vastgesteld. Omdat de meeste practici uit de referentiegroep niet (meer) op alle deelterreinen van de diergeneeskunde werkzaam zijn, zijn alleen de scores van de referenten op wiens terrein de betreffende casus ligt en zij als ter zake deskundig mogen worden beschouwd, meegewogen in de antwoordsleutel. De nonrespons van referenten bleek niet selectief ten aanzien van de deelterreinen in de test.

Evaluatie

In aanvulling op de test kregen alle deelnemende studenten en referenten enkele evaluatieve vragen voorgelegd over hun ervaringen met deelname aan de test en oordelen over de kwaliteiten van de casus en het format van de vraagitems in de test.

4 Resultaten en analyse

4.1 Vaststellen van het scoremodel

Aan de hand van de retour ontvangen responses kon een antwoordsleutel worden vastgesteld; het aantal referenten dat per deelterrein kon worden meegewogen, bedroeg respectievelijk 12 (Gezelschapsdieren, GD), 12 (Landbouwhuisdieren, LHD) en 11 (Paard). Deze aantallen voldoen aan de norm. De antwoordsleutel die hieruit werd geconstrueerd, toonde bij 2/3 van de referenten overeenstemming op één alternatief in 22 vraagitems, en op twee naastgelegen alternatieven (bijv. *zeer onwaarschijnlijk* en *minder waarschijnlijk*) in 71 vraagitems. Deze variatie is in lijn met de verwachtingen.

In 17 van de 120 items was de spreiding in antwoorden van referenten veel groter, zodat deze items op inhoud en formulering opnieuw zijn bekeken. Daarbij werd door twee inhoudsdeskundigen (senior-docenten van de Faculteit Diergeneeskunde) onafhankelijk van elkaar beoordeeld of er sprake was van fouten in het construct of de precieze formulering van casus en vragen, die dergelijke spreiding in responses konden verklaren. Op grond van deze analyse werden geen items uit de test verwijderd. Van deze 17 items behoorden 12 tot de 'expert'-vraagcategorie (zie Tabel 2).

Bepalend voor de weging van elk van de antwoordalternatieven is de mate waarin de referenten onderling overeenstemmen over de juistheid van dat alternatief. Het gebruike-

lijke scoremodel voor de SCT gaat uit van een (modus)score van 1,0 voor het door de referenten meest gegeven antwoord en een gewogen score (tussen 0 en 1), overeenkomstig het aantal referenten dat het betreffende alternatief koos, voor elk van de andere antwoorden.

Een nadere beschouwing van de spreiding in responses binnen de referentiegroep en mogelijke patronen daarin resulteerde in twee hypotheses over het beste scoremodel. De eerste hypothese stelt dat een vijfpuntsschaal onrealistisch is: er is een 'meest waarschijnlijk antwoord' omgeven door twee iets minder plausibele mogelijkheden. De tweede hypothese stelt dat het onjuist is om aan het door experts meest gekozen antwoord een weegfactor van 1 toe te kennen: alle weegfactoren gezamenlijk moeten optellen tot 1. Deze hypotheses werden getoetst via analyse van vijf alternatieve scoremodellen. Tabel 3 toont de modellen en het effect van de scoringsregels op de ranges en gemiddelden van de studenten en de ervaren practici in de eerste testafname. De scores op de modellen 2, 3 en 5 correleren sterk (r = 0,98 resp. 0,89 en 0,96; $p < 0{,}01$; n = 164) met die op model 1, het gebruikelijke SCT-scoringsmodel. Alleen voor model 4 is de correlatie matig (0,66). Dit model is theoretisch echter ook het meest omstreden: het minderheidsexpertoordeel levert hierbij geen punten op en daar zullen zeker experts moeite mee hebben.

Beide typen aanpassingen van het scoremodel (transformatie van de 5-punts- naar een 3-puntsschaal, relatieve modusscore in

Tabel 3

Het effect van verschillende scoremodellen op de resultaten in de eerste afname (met de gemiddelde, hoogste en laagste scores, ten opzichte van het theoretisch maximum)

Model 1 (α = 0,79)					
(5-p schaal, modus=1, alle andere waardes gewogen t.o.v. moduswaarde)	max	hoog	laag	*M*	*SD*
Studenten	120	86	52	75	5,54
Practici		107	73	94	7,99
Model 2 (α = 0,69)					
(5-p schaal, alle waardes gewogen t.o.v. totaal)	max	hoog	laag	*M*	*SD*
Studenten	64	46	27	39	3,21
Practici		61	38	51	5,62
Model 3 (α = 0,64)					
(3-p schaal, modus=1, naastgelegen waarde in zelfde richting gewogen t.o.v. totaal)	max	hoog	laag	*M*	*SD*
Studenten	120	66	30	51	6,01
Practici		91	45	69	10,74
Model 4 (α = 0,68)					
(3-p schaal, transformatie: waardes in zelfde richting samengevoegd)	max	hoog	laag	*M*	*SD*
Studenten	82	66	41	60	3,70
Practici		78	57	70	4,12
Model 5 (α = 0,70)					
(3-p schaal, modus=1, naastgelegen waarde gewogen t.o.v. moduswaarde)	max	hoog	laag	*M*	*SD*
Studenten	120	75	41	61	5,85
Practici		97	57	80	9,44

plaats van score = 1) leiden in de praktijk tot schaalreductie en een (ogenschijnlijk) hogere mate van overeenstemming tussen alle referenten. De interne consistentie van de toets (Cronbach's α) neemt echter in alle gevallen af. Dit is het gevolg van het met de reductie gepaard gaande informatieverlies. Gegeven het doel om de SCT te gebruiken voor periodieke monitoring van competentieontwikkeling in de laatste fase(n) van de studie en de daarin verwachte progressie, is het onwenselijk dat de onderlinge verschillen en daarmee de sensitiviteit kleiner wordt. De uiteindelijke analyses zijn dan ook gebaseerd op model 1, het oorspronkelijke scoremodel.

Bij één van de referenten bleek sprake van 40% *outlier responses* (= geen modus of direct naastgelegen waarde) ten opzichte van de overige referenten en een individuele score lager dan gemiddeld minus twee standaarddeviaties. Ook scoorde deze referent lager dan het student-gemiddelde. Besloten werd deze referent uit de definitieve antwoordsleutel te verwijderen.

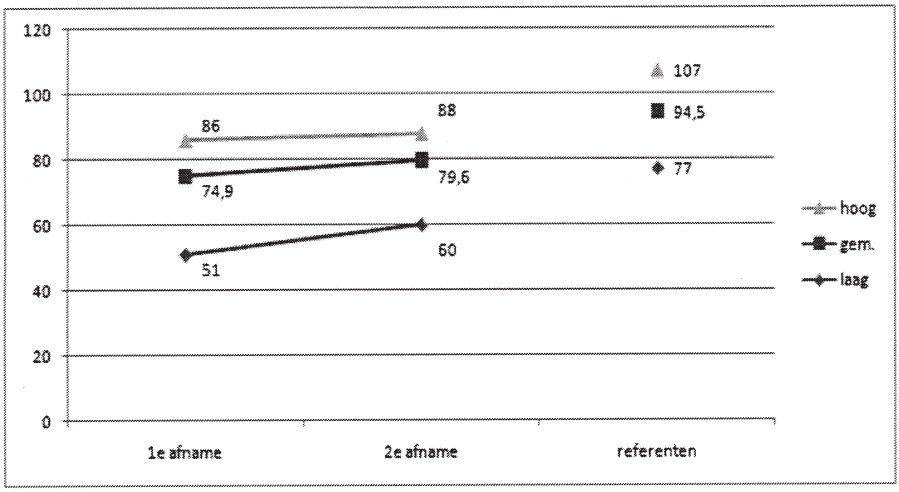
4.2 Scores van studenten

Van de ingeschrevenen studenten hebben 168 studenten (97,7%) de eerste test gemaakt. Daarvan namen 148 (86%) deel aan beide afnames. Uitval had vooral te maken met ziekte en studieonderbreking. Figuur 1 toont hun uiteindelijke scores op de eerste en tweede afname, vergeleken met die van de referentiegroep. Daarbij blijkt een scoretoename van de tweede (gemiddelde = 79,6; *SD* = 4,9) ten opzichte van de eerste afname (gemiddelde = 74,9; *SD* = 5,6). Dat verschil is significant (t = 12,753, df = 147, p < 0,00025). De individuele resultaten van de studenten op beide afnames correleren, bovendien, positief (r = 0,653, n = 148, p < 0,001) en het effect van de klinische lessen is groot (Cohen's d = 0,89).

4.3 Psychometrische analyses

De referentiescores en de studentantwoorden zijn statistisch geanalyseerd om zicht te krijgen op de psychometrische kwaliteiten van de test. Waar – op basis van grotere spreiding in de gekozen antwoordalternatieven dan wel een lage itemrestcorrelatie – daar aanleiding voor bleek, zijn casus en vraagitems opnieuw inhoudelijk bekeken om na te gaan of zij vanuit oogpunt van validiteit (Borsboom, Mellenbergh, & Heerden, 2004) in aanmerking kwamen voor verwijdering uit de test. Cronbach's α voor de eerste en de tweede afname zijn respectievelijk 0,79 en 0,78. Voor geen van de items zou verwijdering ervan, vanwege een lage itemrestcorrelatie, leiden tot een stijging van α met meer dan 0,004.

Tenslotte is bepaald hoe het niveau van de studenten en de verschillen tussen de eerste en tweede afname, binnen de diverse subsets (afstudeerrichting, type patiënten, domein van oordelen en beslissingen, vraagniveau),



Figuur 1. De scores van studenten in de eerste en tweede afname, vergeleken met het referentiepanel van ervaren practici.

Tabel 4

G-studie: variantie-analyse met de participanten, vraagitems en afnamemomenten als componenten, in een gekruisd twee-facet design

GENERALISEERBAARHEIDSANALYSE: (P * F1 * F2)

Aantal deelnemers: 160
Aantal items / Facet 1: 120
Aantal afnamen / Facet 2 : 2

Bron	*df*	*SS*	*MS*	Variantie-component	Proportie
Deelnemers (P)	159	156.641	.985	.0004	2,5%
Items (F1)	119	666.981	5.605	.016	11,6%
Afnamen (F2)	1	12.484	12.484	.001	0,5%
P*F1	18921	2749.176	.145	.027	19,2%
P*F2	159	12.522	.079	.000	0%
F1*F2	119	38.523	.324	.001	1,0%
P*F1*F2	18921	1728.250	.091	.091	65,1%
Foutenvariantie	Relatief 0,001	Absoluut 0,001			
G-coefficiënten	G 0,854	*φ* 0,679			

zich verhouden tot het niveau van de ervaren practici. Binnen deze subsets kwamen geen significante verschillen tussen groepen of vraagcategorieën naar voren.

Om meer zicht te krijgen op de betrouwbaarheid van de test en de proportionele bijdrage daaraan van de verschillende variantiecomponenten (de studenten, de items, en de momenten van afname) en hun onderlinge interacties, is een generaliseerbaarheidsstudie (Brennan, 2001) uitgevoerd in een zogenaamd gekruist-twee-facet design. De resultaten hiervan zijn te vinden in Tabel 4. De gevonden generaliseerbaarheidscoëfficient (G) indiceert dat ruim 85% van de totale variantie voortkomt uit werkelijke verschillen tussen deelnemers, dus een hoge mate van generaliseerbaarheid van resultaten. De twee momenten van afname bij dezelfde groep blijken nauwelijks bij te dragen aan de ruis, dus het feit dat dezelfde studenten dezelfde test opnieuw maken levert geen vertekening op. Voor zover er sprake is van ruis, is die met name toe te schrijven aan de interactie van deelnemers, items en afnamemomenten. De items lijken ook niet alle 120 geheel onafhankelijk van elkaar zijn. Dat is in een inhoudelijk samenhangend kennisdomein ook niet verwonderlijk en nauwelijks te vermijden.

4.4 Aanvullende oordelen van deelnemers

Uit de aanvullende evaluatievragen (zie Tabel 5) komt onder andere naar voren dat beide groepen, studenten en referenten, het realiteitsgehalte van de casus hoog beoordelen (4,4 respectievelijk 4,0). De studenten gaven aan meer moeite met de casus te hebben gehad (4,3 voor de studenten versus 3,3

Tabel 5

Vergelijking van uitkomsten op overeenkomstige evaluatievragen bij beide groepen

	studenten		practici	
	M	*SD*	*M*	*SD*
1. Moeilijkheidsgraad inhoud van de casus	4,34	0,72	3,30	0,82
2. Moeilijkheidsgraad vorm van de vragen	3,89	0,94	3,71	1,27
3. Realiteitsgehalte problemen en omstandigheden in deze casus	4,40	0,81	3,96	1,14
4. Overeenkomst van casus met casus in Klinische Lessen (*practici: eigen praktijk*)	3,41	1,02	3,93	1,14
5. aanspraak van test op klinisch redeneren (versus parate kennis)	2,41	0,97	3,11	0,89
6. variatie in niveau van afzonderlijke casusitems	3,21	0,88	4,04	0,79
7. helderheid instructietekst	3,78	1,06	4,00	1,22

voor de ervaren practici), ervoeren meer aanspraak op parate kennis dan op het klinisch redeneren (2,4 versus 3,2) en minder variatie in het niveau van de casusvragen (3,2 versus 4,0).

Deze verschillen tussen relatieve beginners en ervaren practici zijn niet meer dan logisch. Aanvullend op deze vragen gaven zowel practici als studenten aan moeite te hebben met het oordelen in casus waarin een relevante, maar niet de meest voor de hand liggende diagnose of behandeling werd gesuggereerd.

5 Discussie en conclusies

Beoogd werd om een test te ontwikkelen waarmee kan worden bepaald hoe goed het klinisch probleemoplossen van studenten diergeneeskunde is ontwikkeld om realistische vraagstukken en omstandigheden uit de professionele praktijk, adequaat te kunnen hanteren. Met name oordelen en beslissen in onzekerheid, dat een intrinsiek onderdeel is van de authentieke beroepspraktijk maar in de gangbare vormen van toetsing niet of nauwelijks deel uitmaakt van het beoordelingsproces, zou hierin een belangrijke plaats moeten innemen. De onderzoeksvragen van deze studie hebben betrekking op de validiteit, betrouwbaarheid en sensitiviteit van de test, en op de mogelijkheid de test in te zetten voor veel studenten in een studiefase die voorafgaat aan praktijkstages.

5.1 Validiteit

De eerste onderzoeksvraag is of de hier onderzochte test inderdaad valide is voor wat betreft het brede domein van de Diergeneeskunde in de preklinische fase van de studie. We constateren dat een belangrijk verschil tussen schriftelijke tests, zoals deze SCT, en het klinisch probleemoplossen in de praktijk de aanwezigheid betreft van het dier en de eigenaar. Hun aanwezigheid vergt parallel aan het proces van probleemanalyse, aandacht voor de uitvoering van allerlei noodzakelijke handelingen, aspecten van welzijn en veiligheid, en communicatie met de eigenaar. Bovendien worden ook andere aspecten van het klinisch probleemoplossen aangesproken: gericht selecteren van de benodigde informatie, interpreteren van eigen waarnemingen en zelf genereren van hypothesen. De toetssituatie is een reductie (ten behoeve van de uitvoerbaarheid) en is dus per definitie minder valide dan wanneer de studenten alle casus in echte praktijksituaties moeten oplossen. De vraag is dus niet of de test *in zijn algemeenheid* valide is, maar of zij valide is voor het nemen van beslissingen over het niveau van de studenten.

Kenmerkend voor deze SCT is dat deze wel uitgaat van authentieke situaties, problemen en omstandigheden waarbij – net als in de klinische praktijk – meerdere opties mogelijk zijn. Het geheel aan casus en items (120 in totaal) is gebaseerd op epidemiologische gegevens en vormt in de ogen van experts een representatieve selectie van de in de eerstelijns praktijk veel voorkomende klachten en aandoeningen. In dit opzicht doet de SCT niet onder voor traditionele gesloten kennistoetsen, maar overtreft ze de mogelijkheden van beoordelen in de ('echte') praktijk zeer ruim. Binnen dit aantal vraagitems was het mogelijk om op alle terreinen waar in de praktijk oordelen en beslissingen worden gevraagd, sets van representatieve vragen aan de orde te stellen. Ondanks de typische, en als lastig ervaren formulering van vraagitems in de SCT geven de referenten aan dat het realiteitsgehalte van de casus en vragen voor hun praktijken hoog is. Studenten bevestigen evenzeer de overeenkomsten tussen de SCT-casus en het type casus waarmee zij in de klinische lessen werden geconfronteerd.

De verschillen in het klinisch probleemoplossen van experts zijn verdisconteerd in de 'standaard' die gehanteerd wordt voor beoordeling van het door deelnemers gekozen antwoord. Overeenstemming met de oordelen en beslissingen van een groep ervaren practici geldt als de standaard waartegen de antwoorden van de studenten worden afgezet. Hierdoor zijn complexere, open vraagstukken, reële dilemma's en variatie in mogelijke wijzen van oplossen, niet uitgesloten in deze test. Inhoudelijk bood het format voldoende ruimte voor vragen variërend van oorzakelijke en beïnvloedende factoren tot en met afwegingen vanuit het perspectief van kosten, voorkeuren van de diereigenaar,

of risico's voor andere dieren en omgeving.

Uit analyse van de hardop-denkprocessen tijdens de pretest en uit de studentenevaluatie na afloop van de testafname kan worden geconcludeerd dat de SCT inderdaad cognitieve activiteiten vergt die overeenkomen met klinisch probleemoplossen in de praktijk: interpreteren en combineren van patiëntgegevens, schatten van de betrouwbaarheid van informatie en informatiebronnen, toetsen van hypothesen, wegen van waarschijnlijkheden bij beperkt beschikbare informatie etc.

We concluderen dat de SCT aan eisen van validiteit voldoet. De test biedt betere mogelijkheden dan gebruikelijke gesloten theorietoetsen vanwege het beslissen in onzekerheid. Ook is hij representatiever dan de gebruikelijke beoordelingen omdat in de werkelijke klinische praktijk studenten doorgaans slechts een gering aantal observatiebeoordelingen krijgt.

5.2 Betrouwbaarheid

De tweede onderzoeksvraag is of de test voldoende betrouwbaar is. Omvat de test inderdaad ambiguïteit in de klinische problemen en dus beslissen in onzekerheid, maar leidt hij niet tot onzekerheid over het niveau van de studenten? Onze redenering is dat, wanneer de casus en vraagitems elk afzonderlijk voldoende valide zijn, een betrouwbare vaststelling van het niveau van de student een kwestie is van testlengte. Het sterke punt van het SCT-format is dat het in principe mogelijk is een voldoende lange test op te stellen. Dit onderzoek laat zien dat dit ook praktisch gerealiseerd kan worden: de test bevat een voldoende aantal vragen (120) en de betrouwbaarheid (Cronbach's α) is naar wens. Ook heeft het herhaald afnemen van dezelfde test geen invloed op de uitkomsten. Een belangrijk aandachtspunt is de noodzakelijke spreiding in expertoordelen. Allereerst is daar de vraag hoe groot deze spreiding mag zijn, wat zich vertaalt naar de vraag hoe 'open' de vraagstukken moeten zijn aan de hand waarvan wordt getest. Wanneer de casusinformatie en de bijbehorende vragen te gesloten zijn (slechts één correct antwoord hebben) dan wordt de aanspraak op redeneren, oordelen en beslissen minder en vervangen door al dan niet kennis hebben van het feit. Wanneer ze te open zijn (vele mogelijk goede antwoorden), dan is niet meer te bepalen wat de waarde van elk antwoord is, en zijn leek en deskundige niet meer onderscheiden. De items in deze test bleken in dit opzicht te voldoen, maar een hard criterium voor de mate van spreiding heeft zich niet opgedrongen. Nader onderzoek hiernaar is een interessante mogelijkheid.

Belangrijk is verder dat de spreiding veroorzaakt wordt door normale variatie in aanpak en focus bij experts, en niet voortkomt uit fouten in het construct (de formulering van de casus en vragen) of fouten bij de experts (deskundigheid op dit onderwerp/vakgebied die mogelijk tekort schiet; inadequate selectie van de experts). De maatregelen die getroffen zijn om onwenselijke bronnen van spreiding te elimineren (een procedure waarin de casus en antwoordmogelijkheden onder constructie regelmatig tussentijds zijn beoordeeld; zorgvuldige selectie van experts; voldoende groot aantal experts; eliminatiecriteria voor verwijdering) zijn effectief gebleken.

Aanvullend onderzoek naar alternatieve scoremodellen laat zien dat het oorspronkelijke model (een vijfpuntsschaal waarbij het door experts meest gekozen antwoord de waarde 1 krijgt en andere door experts gekozen antwoorden een waarde krijgen recht evenredig met het percentage) het beste voldoet.

De voornaamste conclusie is dat, ondanks dat zowel experts als studenten in essentiële zin beslissen in onzekerheid en er geen 100% zekerheid is over het goede antwoord, het niveau van studenten onomstreden kan worden vastgesteld.

5.3 Sensitiviteit

De derde vraag betreft de sensitiviteit: is de test inderdaad voldoende sensitief om binnen het normale verloop van een studieonderdeel progressie te meten wat betreft het vermogen problemen op te lossen? De analyse van de testresultaten en de onderlinge vergelijking van resultaten op de twee afnamen laten zien dat met de SCT inderdaad een significante verbetering in de resultaten van studenten kan worden vastgesteld na deelname aan het reguliere onderwijs (de Klinische lessen). Iets dergelijks is met een SCT niet eerder

binnen één populatie vastgesteld. De studenten verbeteren zich gemiddeld van 74,9 naar 79,6, waarbij de experts zich bevinden op 94,5. Het gaat om relatief beginnende studenten, en deze scores impliceren dat er voldoende ruimte is om tussenliggende niveaus te bepalen, bijvoorbeeld tijdens de co-schappen en direct na het afstuderen. Op het vaststellen van deze niveaus is vervolgonderzoek gericht.

5.4 Praktische uitvoerbaarheid

Ten slotte is onderzocht of de test uitvoerbaar is in de normale onderwijspraktijk, dus met grote aantallen studenten. In de onderzochte situatie is de test geen onderdeel van het reguliere programma maar afgenomen op vrijwillige basis. In de praktijk bleken bijna alle studenten uit de populatie te participeren, en het grootste deel participeerde in beide afnamen. Niet-deelname had overwegend triviale oorzaken (ziekte, studieonderbreking). Daaruit kan worden verondersteld dat studenten de test niet als belastend ervaren. Studenten vonden deelname interessant omdat het hun informatie gaf over hun relatieve niveau. Wij verkregen in deze studie geen enkele aanwijzing dat dit type test op dit moment van de opleiding ongeschikt is, bijvoorbeeld vanwege onderontwikkelde ziektescripts. Ook voor de onderwijsorganisatie bleek testafname en testanalyse geen probleem. Hierin komt de kracht tot uitdrukking van een 'papieren' test die met een computer (optisch leesbare scoreformulieren) is na te kijken. Het is met name de testconstructie die tijdrovend is, vooral ook omdat er nog weinig routine is wat betreft het formuleren van SCT-casus en vraagitems in de onderwijsorganisatie. Opgemerkt werd dat docenten die betrokken werden bij de constructie de neiging hadden om de items steeds meer gesloten te maken: hiervoor moet expliciet gewaakt worden. Belangrijk is verder dat de antwoordsleutels niet beschikbaar komen voor studenten zodat de test inderdaad meerdere malen tijdens de studie afgenomen kan worden. Uit de analyses en de evaluaties bleek overigens geen effect van de eerste afname op de tweede afname. De tussenliggende periode (een half jaar), de hoeveelheid casusinformatie en de duurzame onbekendheid met 'het goede antwoord' garanderen (kennelijk) dat studenten ook in de tweede afname het gewenste cognitieve proces van klinisch redeneren volgen als strategie om tot een antwoord te komen en dat zij niet in hun geheugen gaan graven.

Samenvattend concluderen wij dat het SCT-format zoals onderzocht in deze studie, robuust genoeg is om een test te construeren waarmee voortgang in de ontwikkeling van competentie in het oordelen en beslissen bij klinische problemen, in deze context kan worden bepaald. Noch de breedte van het terrein, noch de beperkte klinische ervaring van studenten in de fase van de opleiding waarin de test werd afgenomen, heeft geleid tot waarneembaar negatieve effecten op kwaliteit van het instrument. De test omvat een substantieel aantal casus en vragen, representatief voor de problemen en omstandigheden in de professionele praktijk. Deze vergen het combineren van informatie, het toetsen van hypothesen en nemen van beslissingen over vervolgstappen in het onderzoek c.q. behandeling of preventie. Waar toetsvormen gebaseerd op echte patiënten meer mogelijkheden bieden om enkele specifieke aspecten van het klinisch probleemoplossen te testen, kan een SCT relatief gemakkelijk met een groot aantal casus en bij vele studenten worden afgenomen, zonder excessieve belasting van patiënten. Het geheel aan sterke kanten van het Script Concordance Test-format weegt, naar ons oordeel, dan ook ruimschoots op tegen de beperkingen ervan.

Dat een in deze context valide, betrouwbare en sensitieve SCT kon worden ontwikkeld waarmee de vorderingen van deze studenten in hun ontwikkeling van het klinisch redeneren en beslissen kon worden bepaald, roept de vraag op in hoeverre de vooronderstellingen over de ontwikkeling van scripts en de noodzaak van eigen praktijkervaring in relatie tot het gebruik van dit type tests, houdbaar zijn. De bevindingen in deze studie wijzen vooralsnog niet op beperkingen met betrekking tot het gebruik in een aan de praktijk voorafgaande fase en voor een breder domein dan de afgebakende specialismen waarin zij tot op heden werden ingezet. De aangetoonde kwaliteit van het instrument

biedt naar ons oordeel bovendien voldoende mogelijkheid om deze test in te zetten als onderzoeksinstrument bij verdere studies in deze context.

Deze studie bleef beperkt tot de progressie in het 'overgangsjaar' binnen de opleiding, van de preklinische fase naar co-schappen. In hoeverre met deze test verdere progressie in de laatste, klinische fase kan worden vastgesteld, vergt vervolgonderzoek. Wij raden verder onderzoek van de SCT aan om meer zicht te krijgen op zijn specifieke karakteristieken bij gebruik voor werkelijke toetsdoeleinden (formatief-summatief) en binnen andere domeinen.

Literatuur

Balla, J. I., & Edwards, H. M. (1986). Some problems in teaching clinical decision-making. *Medical Education, 20*, 487-491.

Berg, M. (1997). Problems and promises of the protocol. *Social Science & Medicine, 44*, 1081-1088.

Bland, A. C., Kreiter, C. D., & Gordon, J. A. (2005). The Psychometric properties of five scoring methods applied to the Script Concordance Test. *Academic Medicine, 80*, 395-399.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. v. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Boshuizen, H. P. A. (2003, XX). *Expert development; The transition between school and work*. Paper presented at the Expert development: How to bridge the gap between school and work. Plaatsnaam.

Boshuizen, H. P. A., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science, 16*(2), 153-184.

Brailovsky, C. A., Charlin, B., Beausoleil, S., Cote, S., & Vleuten, C. P. M. van der. (2001). Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. *Medical Education, 35*, 430-436.

Brazeau-Lamontagne, L., Charlin, B., Gagnon, R., Samson, L., & Vleuten, C. P. M. van der. (2004). Measurement of perception and interpretation skills along radiology training: utility of the script concordance approach *Medical Teacher, 26*, 326-332.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Caire, F., Sol, J.-C., Charlin, B., Isidori, P., & Moreau, J.-J. (2004). Le test de concordance de script (TCS) comme outil d'évaluation formative des internes en neurochirurgie: implantation du test sur Internet à l'échelle nationale. *Pédagogie Médicale, 5*, 87-94.

Charlin, B., Boshuizen, H. P. A., Custers, E. J., & Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical Education, 41*, 1178-1184.

Charlin, B., Brailovsky, C., Leduc, C., & Blouin, D. (1998). The Diagnosis Script Questionnaire: A new tool to assess a specific dimension of clinical competence. *Advances in Health Sciences Education, 3*(1), 51-58.

Charlin, B., Brailovsky, C. A., Brazeau-Lamontagne, L., Samson, L., Leduc, C., & Vleuten, C. P. M.van der. (1998). Script questionnaires: Their use for assessment of diagnostic knowledge in radiology. *Medical Teacher, 20*, 567-571.

Charlin, B., Desaulniers, M., Gagnon, R., Blouin, D., & Vleuten, C. P. M. van der. (2002). Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and Learning in Medicine, 14*, 150-156.

Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. P. M. (2000). The Script Concordance Test, a tool to assess the reflective clinician. *Teaching and Learning in Medicine, 12*, 189-195.

Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Academic Medicine, 75*, 182-190.

Charlin, B., & Vleuten, C. P. M. van der. (2004). Standardized assessment of reasoning in contexts of uncertainty: The Script Concordance Approach. *Evaluation & the Health Professions, 27*, 304-319.

Custers, E. J. F. M., Boshuizen, H. P. A., & Schmidt, H. G. (1996). The influence of medical expertise, case typicality, and illness script component on case processing and disease probability estimates *Memory & Cognition, 24*, 384-399.

Downing, S. M. (2003). Validity: on the meaningful

interpretation of assessment data. *Medical Education, 37*, 830-837.

Eddy, D. M. (1990). Clinical decision making: from theory to practice. Anatomy of a decision. *Journal Of The American Medical Association, 263*, 441-443.

Elstein, A. S. (2004). On the origins and development of evidence-based medicine and medical decision making. *Inflammation Research, 53*, S184-S189.

Elstein, A. S., Schulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Elstein, A. S., & Schwarz, A. (2002). Evidence base of clinical diagnosis: Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *British Medical Journal, 324*, 729-732.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *Journal of the American Medical Association, 287*, 226-235.

Eraut, M. (2004). *Developing professional knowledge and competence*. London: Routledge-Falmer.

Forde, R. (1998). Competing conceptions of diagnostic reasoning; Is there a way out? *Theoretical Medicine and Bioethics, 19*(1), 59-72.

Gagnon, R., Charlin, B., Coletti, M., Sauve, E., & Vleuten, C. P. M. van der. (2005). Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Medical Education, 39*, 284-291.

Gagnon, R., Charlin, B., Roy, L., St-Martin, M., Sauvé, E., Boshuizen, H., et al. (2006). The cognitive validity of the Script Concordance Test: A processing time study. *Teaching and Learning in Medicine, 2006, 18*(1), 22-27.

Grant, J., & Marsden, P. (1988). Primary knowledge, medical education and consultant expertise. *Medical Education, 22*, 173-179.

Hunink, M. G. M. (2001). In search of tools to aid logical thinking and communicating about medical decision making. *Medical Decision Making, 21*, 267-277.

Jonassen, D. H. (2004). *Learning to solve problems: An instructional design guide*. San Francisco: Pfeiffer.

Kirschner, P. A. (2002). Cognitive load theory: implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*, 1-10.

Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis. *Science, 130*, 9-21.

Lilford, R. J., Pauker, S. G., Braunholtz, D. A., & Chard, J. (1998). Getting research findings into practice: Decision analysis and the implementation of research findings. *British Medical Journal, 317*, 405-409.

Linn, R. L., Baker, E., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 23*(16), 1-21.

Meterissian, S., Zabolotny, B., Gagnon, R., & Charlin, B. (2007). Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *American Journal of Surgery, 193*, 248-251.

Nendaz, M. R., Gut, A. M., Perrier, A., Reuille, O., Louis-Simonet, M., Junod, A. F., et al. (2004). Degree of concurrency among experts in data collection and diagnostic hypothesis generation during clinical encounters. *Medical Education, 38*(1), 25-31.

Neufeld, V. R., Norman, G. R., Feightner, J. W., & Barrows, H. S. (1981). Clinical problem-solving by medical students: a Longitudinal and Cross-sectional Analysis. *Medical Education, 15*, 315-322.

Newmann, F. M., & Archbald, D. A. (1992). The nature of authentic academic achievement. In H. Berlak, F.M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven & T.A. Romberg (Eds.), *Toward a new science of educational testing and assessment*. Albany, NY: State University of New York Press.

Norman, G. R. (2005). Research in clinical reasoning: past history and current trends. *Medical Education, 39*, 418-427.

Norman, G. R., & Brooks, L. R. (1997). The non-analytical basis of clinical reasoning. *Advances in Health Sciences Education, 2*, 173-184.

Norman, G. R., & Schmidt, H. G. (1992). The psychological basis of problem-based learning: a review of the evidence. *Academic Medicine, 67*, 557-565.

Norman, G. R., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: the role of experience. *Medical Education, 41*, 1140-1145.

Patel, V. L., Arocha, J. F., & Zhang, J. (2005).

Thinking and reasoning in Medicine. In K.J. Holyoak & R.G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 727-751). New York: Cambridge University Press.

Rikers, R. M. J. P., Schmidt, H. G., & Moulaert, V. (2005). Biomedical knowledge: encapsulated or two worlds apart? *Applied Cognitive Psychology, 19*, 223-231.

Rimoldi, H. J. A. (1961). The test of diagnostic skills. *Journal of Medical Education, 36*, 73-79.

Sarasin, F. P. (2001). Decision analysis and its application in clinical medicine. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 94*, 172-179.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory & Cognition, 21*, 338-351.

Segers, M., Dochy, F., & De Corte, E. (1999). Assessment practices and students knowledge profiles in a problem-based curriculum. *Learning Environments Research, 2*, 191-213.

Sibert, L., Charlin, B., Corcos, J., Gagnon, R., Grise, P., & Vleuten, C. P. M. van der. (2002). Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. *Medical Teacher, 24*, 522-527.

Sibert, L., Charlin, B., Corcos, J., Gagnon, R., Lechevallier, J., & Grise, P. (2002). Assessment of clinical reasoning competence in urology with the Script Concordance Test: An exploratory study across two sites from different countries. *European Urology, 41*, 227-233.

Sibert, L., Darmoni, S. J., Dahamna, B., Hellot, M. F., Weber, J., & Charlin, B. (2006). On line clinical reasoning assessment with Script Concordance test in urology: Results of a French pilot study. *BMC Medical Education, 6, XX-XX*.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: lessons from the health professions. *Educational Researcher, 24* (5), 5-11.

Thornton, T. (2006). Tacit knowledge as the unifying factor in evidence based medicine and clinical judgement. *Philosophy, ethics and humanities in medicine, 1* (March, 17), 2.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, XX-XX.

Vleuten, C. P. M. van der. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education, 1* (1), 41-67.

Vleuten, C. P. M. van der, & Schuwirth, L. W. T. (2005). Assessing professional competence: from methods to programmes. *Medical Education, 39*, 309-317.

Manuscript aanvaard: 12 maart 2009

Auteurs

Stephan Ramaekers, Wim Kremer, Albert Pilot, Peter van Beukelen en **Hanno van Keulen** zijn werkzaam aan de Universiteit Utrecht. De eerste, derde en vijfde auteur bij het IVLOS, en de tweede en vierde auteur bij de Faculteit der Diergeneeskunde.

Correspondentieadres: Stephan Ramaekers, IVLOS, Universiteit Utrecht, Postbus 80127, 3508 TC Utrecht. E-mail: s.p.j.ramaekers@uu.nl.

Abstract

Determination of the power to decide in uncertainty with the method of Script Concordance Test

Real-life professional problems may require making decisions, although the available information is limited or ambiguous. Incorporating such authentic uncertainties into an assessment, however, poses problems in analysing results and establishing its psychometric qualities. Based on the Script Concordance Test format, we developed a test on clinical decision making in veterinary medicine, incorporating realistic uncertainties. This test was administered twice to 148 students, near the beginning and the end of a course in clinical reasoning. Their answers were compared to the judgements and decisions of 28 experienced practitioners. Student scores on the pre- and post-test show a mean improvement of 4.59 points. Individual scores correlate positively (r = 0.65) and the effect size is large (d = 0.89). From the analysis and substantive appraisal of the cases and items, it is concluded that this SCT can be used for large student groups to assess competence in clinical decision making, without creating a burden on patients.