Standardized assessment of ill-defined clinical problems : the script concordance test

Summary

The dissertation is built around a general research question:
Is it possible to conceive a clinical reasoning assessment tool that may be congruent with recent developments in cognitive psychology and assessment sciences?

In cognitive research on medical expertise, there has been a shift from the search for a generic problem-solving skill toward a focus on memory organization, knowledge use, problem representation, and how they change with experience. In the testing and evaluation domain, this change of focus did not have yet many applications. The script concordance test (SCT) is conceived in the perspective of the recommendation of Elstein *et al* (1990). These authors suggested that evaluation should concentrate on judging the quality of a set of cognitive operations or knowledge structures by comparing a student's problem representation, judgments, and choices to those of the experienced group.

Theoretical validity of the tool is the theme of the first research question. In what way can we corroborate the choices that were made in the design of the tool with theoretical considerations and research outcomes in assessment?

Chapter 1 presents an adaptation of cognitive psychology script theory to the characteristics of clinical reasoning. We explain that during the data collection process, physicians are systematically comparing incoming information to the values that are acceptable or unacceptable for each of the script attributes. For each attribute, if unacceptable values are found, the script is rejected, and other scripts that accept that value are activated or reinforced. Among acceptable values for an attribute, some bring more weight to a hypothesis than others. According to theory, clinical reasoning is therefore made of a series of qualitative judgments. Each of these judgments can be measured. This provides a method of clinical reasoning process assessment.

Chapter 2 describes how the script concordance test (SCT) is conceived to achieve this. The principle is to describe rich authentic clinical contexts in which there is

not enough data to make a diagnostic or management decision. Therefore the situation represents a problem, even for an expert of the domain. Items are made from the questions experienced physicians ask and the actions they take in these contexts. The SCT is innovative in the three components any assessment tool has: the stimulus format (examinees' task), the response format (how examinees' performances are recorded), and the scoring system (the transformation of performances into scores). The stimulus consists of a challenging clinical task. The response format is in accordance with what is known from clinical reasoning processes: a Likert scale, measuring the judgments that are constantly made within the clinical reasoning process, capturing examinees' answers. The scoring method, an adaptation of the aggregate scoring method, takes into account variation of answers among a criterion group. With these characteristics, stemming from the described theoretical frameworks, the SCT allows in the context of a series of clinical tasks, to compare examinees' scripts to those of a panel of experienced clinicians (hence the name of the test).

The second research question concerns the construct validity of the SCT. One of the limitations of current clinical assessment tools is the intermediate effect i.e., the puzzling fact that experienced clinicians score little better and sometimes worse than less experienced clinicians or students. The SCT probes script development i.e., the development of specialized knowledge structures that contain the clinically relevant information that clinicians use in their clinical activities. It also explores the capacity for data interpretation when making clinical decisions, clearly a skill that belongs more to clinical competence than the simple recall of factual data. Four studies presented in the dissertation are applications for different goals of the SCT in different disciplines. We found a disappearance of the intermediate effect in all four applications / contexts. We conclude that these results represent an argument in favor of the SCT construct validity.

The third research question explores the reliability and feasibility of the SCT. In three studies, Cronbach alpha coefficient value was in the 0.79 to 0.82 range, with relatively small numbers of items (29 to 80). These values indicate a good reliability of the test, especially when interpreted in comparison with the time required by other examination formats to reach the 0.80 value. In terms of feasibility, in light of the experienced acquired with the studies described in the dissertation, SCT appears relatively easy to build, to administer and to score.

The fourth research question concerns the validity of the scoring method used in the SCT, the aggregate method. The fourth chapter of the dissertation is a comparison of that method with the usual consensus method in which jury members (criterion group) are required to decide, by group discussion, on a single right answer. The study showed a change of answers on 59 % of items. This magnitude of change, if confirmed by other studies, has important consequence for the assessment of ill-defined problems. It suggests that criterion group responses should be obtained in contexts that are as similar as possible to those examinees will have.

The study provides another argument in favor of the aggregate scoring method. The method appears more sensitive to detection of expertise than the typical consensus method. When experts are judged on a consensus-based scoring scheme built by other experts, their own expertise is difficult to detect. The aggregate method appears superior to the more conventional consensus method. It offers an objective scoring process for the assessment of complex performance in ill-defined situations.

The fifth research question examines the applicability of the SCT in different clinical contexts. Studies presented in chapters 3 to 6 show that the SCT can be used in disciplines as diverse as gynecology, radiology or urology and that it is able to discriminate examinees across their level of experience in the three disciplines. They also indicate that same tools can be used for students, residents and physicians in practice. The tool is well accepted and found interesting to pass. In the study presented in chapter 5, the same test was administered in English to the Canadian subjects and in French to the French participants. The test was able to discriminate among participants of the two countries according to their level of clinical experience, despite linguistic, cultural, educational and institutional differences. This was true even when examinees were judged with the criterion group of the other culture, although there was a slight difference in that examinees obtained higher scores when criterion groups from their own culture assessed them. The study described in chapter 6 was held to verify if the SCT could be used as a test of X-ray interpretation skills, with actual films as stimuli. The perception skills of radiological signs were assessed with a specific perception test. The study shows that SCT is able to detect a dimension of competence that progresses along training at a different speed than the perception skill.

In the light of the presented studies, the SCT appears as a simple and direct approach to test organization and use of knowledge. It has the strong advantage for a testing method of being relatively easy to construct and use and to be machine-scorable. It can be either paper or computer-based and can be used in under-graduate, post-graduate, or continuing medical education. The dissertation ends with a description of the research and development avenues SCT offers, and with a proposition for the place SCT may occupy in a strategy of clinical competence assessment.

Samenvatting

De onderzoeksvraag die centraal staat in dit proefschrift heeft betrekking op de mogelijkheid om een toetsinstrument voor klinische redeneervaardigheid te ontwikkelen dat aansluit bij de actuele ontwikkelingen in de cognitieve psychologie en de wetenschappelijke inzichten over toetsing.

In het cognitief onderzoek naar medische expertise is de aandacht verschoven van algemene probleemoplossingsvaardigheden naar de organisatie van het geheugen, de toepassing van kennis, probleemrepresentatie en de veranderingen die zich hierin voordoen met het toenemen van de ervaring. Deze verplaatsing van het zwaartepunt is nog niet tot uiting gekomen in veel toepassingen op het gebied van toetsing en evaluatie. De script concordance test (SCT) is ontwikkeld vanuit het perspectief van de aanbeveling gedaan door Elstein et al. (1990). Deze auteurs stelden voor om evaluatie te richten op het beoordelen van de kwaliteit van een aantal samenhangende cognitieve handelingen of kennisstructuren door een vergelijking te maken tussen de probleemrepresentatie, oordeelsvorming en keuzen van studenten en die van een groep ervaren deskundigen.

De eerste onderzoeksvraag gaat in op de theoretische validiteit van het toetsinstrument. Gekeken is naar de theoretische onderbouwing van de keuzes die gemaakt zijn bij het ontwerpen van het instrument en naar de mate waarin het instrument aansluit bij de resultaten van onderzoek over toetsing.

In hoofdstuk 1 wordt beschreven hoe de scripttheorie uit de cognitieve psychologie aangepast is aan de kenmerken van het klinisch redeneren. Beschreven wordt hoe artsen tijdens het proces van gegevensverzameling nieuwe informatie voor elk scriptkenmerk systematisch toetsen aan de waarden die voor dat kenmerk aanvaardbaar of niet aanvaardbaar zijn. Als de gevonden waarden niet bij een bepaald kenmerk passen, wordt het script verworpen en worden andere scripts waar de nieuwe waarde wel bij past geactiveerd of ondersteund. Sommige waar-

den die in overeenstemming zijn met een bepaald kenmerk, vormen een krachtiger ondersteuning voor de hypothese dan andere. Volgens de theorie bestaat klinisch redeneren uit een reeks kwalitatieve oordelen. Elk van deze oordelen kan gemeten worden. Met deze overwegingen als uitgangspunt kan een toetsmethode voor klinische redeneervaardigheid ontwikkeld worden.

Het onderwerp van hoofdstuk 2 is de ontwikkeling van de script concordance test (SCT), een instrument om klinische redeneervaardigheid te toetsen. In de SCT worden rijke authentieke klinische situaties beschreven die onvoldoende aanknopingspunten bevatten om beslissingen over diagnose en beleid te kunnen nemen. Dit betekent dat de situatie ook een expert op het betreffende terrein voor een echt probleem stelt. De toetsitems bestaan uit de vragen en handelingen van een ervaren arts in een vergelijkbare situatie. De SCT bevat vernieuwingen op de drie belangrijke onderdelen van toetsinstrumenten: de stimulus (de opdracht die de kandidaat moet uitvoeren), het antwoord (de manier waarop de prestatie van de kandidaat wordt geregistreerd) en de scoring (het omzetten van de prestatie in een waardering). De stimulus wordt gevormd door een uitdagende klinische opdracht. De antwoordvorm sluit aan bij wat we weten over klinische redeneerprocessen. De kandidaat kan de antwoorden aangeven op een Likertschaal, die de oordelen meet die tijdens het klinisch redeneerproces voortdurend worden gevormd. De toets wordt gescoord met behulp van een aangepaste versie van de aggregatiemethode. Deze methode maakt het mogelijk om verschillende antwoorden die in een criteriumgroep naar voren komen, in de waardering te betrekken. Bovengenoemde kenmerken, die voortvloeien uit de beschreven theoretische kaders, maken de SCT tot een instrument waarmee de scripts die kandidaten ontwikkelen naar aanleiding van een reeks klinische opdrachten vergeleken kunnen worden met de scripts van een panel ervaren clinici (vandaar de naam van de toets).

De tweede onderzoeksvraag gaat in op de constructvaliditeit van de SCT. Een van de beperkingen van de gangbare toetsmethoden voor klinische vaardigheid is het intermediërende effect, dat wil zeggen het verwarrende verschijnsel dat ervaren clinici niet veel beter en soms zelfs slechter scoren dan minder ervaren clinici of studenten. De SCT onderzoekt hoe de kandidaat een script ontwikkelt, met andere woorden de ontwikkeling van specialistische kennisstructuren waarbinnen de klinisch relevante informatie gevonden kan worden die door clinici bij het klinisch handelen wordt toegepast. De SCT onderzoekt ook het vermogen om gegevens te interpreteren bij het nemen van klinische beslissingen, een vaardigheid die duidelijk eerder thuishoort bij klinische competentie dan bij het alleen maar reproduceren van feitenkennis. In dit proefschrift worden vier onderzoeken beschreven waarin de SCT voor verschillende doeleinden en in verschillende disciplines is toegepast. Het intermediërende effect bleek te verdwijnen bij alle vier de toepassingen/contexten. De conclusie lijkt gerechtvaardigd dat deze resultaten de constructvaliditeit van de SCT ondersteunen.

De derde onderzoeksvraag betreft de betrouwbaarheid en de haalbaarheid van de SCT. In drie onderzoeken, waarin het ging om toetsen met een betrekkelijk gering aantal items (29 tot 80), bleek Cronbachs alfa te variëren van .79 tot .82. Deze resultaten vormen een aanwijzing dat de SCT een betrouwbare toets is, vooral in vergelijking met andere examenvormen die meer toetstijd vereisen om een betrouwbaarheid van .80 te bereiken. Wat haalbaarheid betreft lijken de ervaringen uit de onderzoeken die in dit proefschrift beschreven worden erop te wijzen dat de SCT een toets is die eenvoudig is samen te stellen, af te nemen en te scoren.

De vierde onderzoeksvraag betreft de validiteit van de scoringsmethode van de SCT, de aggregatiemethode. In het vierde hoofdstuk van dit proefschrift wordt een vergelijking gemaakt tussen deze methode en de gebruikelijke consensusmethode waarbij panelleden (criteriumgroep) door middel van discussie consensus over het enig juiste antwoord moeten bereiken. In het onderzoek bleek dat het antwoord bij 59% van de toetsitems gewijzigd werd. Als wijzigingen van een vergelijkbare omvang ook in ander onderzoek gevonden worden, heeft dat ingrijpende gevolgen voor de toetsing van onduidelijk omschreven problemen. Deze bevinding zou erop wijzen dat de criteriumgroep de vragen moet beantwoorden in omstandigheden die vergelijkbaar zijn met de omstandigheden waarin het examen wordt afgenomen. Het onderzoek levert nog een argument op dat de aggregatiemethode ondersteunt. De aggregatiemethode blijkt een gevoeliger methode te zijn voor het identificeren van expertise dan de conventionele consensusmethode. Het is moeilijk om de specifieke expertise van een expert te meten met een beoordelingsnorm die door andere experts met behulp van een consensusprocedure is vastgesteld. De aggregatiemethode blijkt superieur te zijn aan de conventionele consensusmethode. Hij biedt een objectieve methode om een complexe prestatie in een onduidelijke situatie te beoordelen.

De vijfde onderzoeksvraag heeft betrekking op de bruikbaarheid van de SCT in verschillende klinische contexten. De onderzoeken die in hoofdstuk drie tot zes beschreven worden, laten zien dat de SCT geschikt is voor uiteenlopende disciplines, zoals gynaecologie, radiologie of urologie en dat de methode in de drie disciplines verschillen in ervaring tussen kandidaten kan meten. De beschreven onderzoeken wijzen ook uit dat dezelfde instrumenten gebruikt kunnen worden voor toetsing van studenten, assistenten en praktiserende artsen. Kandidaten vinden de SCT een acceptabele toets. Ook vinden zij het een interessante toets om te doen. In het onderzoek dat beschreven wordt in hoofdstuk 5 werd een Engelstalige toets afgenomen bij Canadese proefpersonen en een Franstalige versie van dezelfde toets bij Franse proefpersonen. Ondanks de verschillen tussen de deelnemers in taal, cultuur, opleidingssysteem en instelling, bleek de toets onderscheid te kunnen maken tussen de deelnemers in het niveau van klinische expertise. Dit was zelfs het geval als de kandidaten beoordeeld werden met behulp van de criteriumgroep uit de andere cultuur. Wel was er dan een gering

verschil: de kandidaten behaalden namelijk hogere scores als zij beoordeeld werden aan de hand van de criteriumgroep uit hun eigen cultuur. Het onderzoek dat in hoofdstuk 6 wordt beschreven had tot doel te onderzoeken of de SCT ook toegepast kan worden om het beoordelen van röntgenfoto's te toetsen, waarbij echte röntgenfoto's als stimulus gebruikt werden. De vaardigheid om radiologische verschijnselen waar te nemen werd beoordeeld door middel van een speciale waarnemingstoets. Het onderzoek wees uit dat de SCT een dimensie van competentie kan meten die tijdens de opleiding in een ander tempo toeneemt dan de waarnemingsvaardigheid.

De beschreven onderzoeken laten zien dat de SCT een eenvoudige en directe benadering biedt voor het toetsen van de organisatie en toepassing van kennis. Het grote voordeel van de methode is dat toetsconstructie en -afname betrekkelijk eenvoudig zijn en dat automatische scoring mogelijk is. De SCT kan schriftelijk of via de computer afgenomen worden en is zowel in de basisopleiding, de specialistenopleiding als voor nascholing bruikbaar. Het proefschrift wordt afgesloten met aanbevelingen voor verder onderzoek naar en ontwikkeling van de SCT. Ook wordt een voorstel gedaan betreffende de plaats van de SCT als onderdeel van een methode voor het toetsen van klinische competentie.