Kennistoetsing bij huisartsen

SAMENVATTING

Inleiding

De ontwikkeling van toetsingsinstrumenten ten dienste van de nascholing is een kernpunt van het huisartsgeneeskundig kwaliteitsbeleid. Toetsgegevens moeten de individuele huisarts helpen bij de inrichting van diens persoonlijke nascholingstraject. Als artsen vrij zijn hun nascholingsprogramma zonder hulp van dergelijke gegevens vast te stellen, bestaat namelijk het risico dat niet wordt gekozen voor onderwerpen waarvoor nascholing, objectief gezien, het meest nodig is. Zelf ervaren leerbehoeften neigen ertoe aan te sluiten op persoonlijke interesses en specifieke ervaringen. Dit zou ertoe leiden dat vooral wordt gekozen voor nascholingsthema's waar men al veel vanaf weet. Ook op collectief niveau wordt verwacht dat objectieve toetsingsgegevens de nascholing goede diensten kunnen bewijzen. Als immers bekend is wat de 'werkelijke' leerbehoeften en de lacunes van huisartsen zijn, kan dat de planning en de ontwikkeling van nascholingsprogramma's mede sturen. Tenslotte wordt van objectieve toetsingsgegevens verwacht dat ze informatie opleveren die de samenleving voldoende garantie biedt dat de vakbekwaamheid van huisartsen op peil blijft.
Dit onderzoek beperkt zich tot de mogelijkheden die toetsing van huisartsgeneeskundige kennis in deze opzichten te bieden heeft. Een adequaat kennisniveau wordt daarbij beschouwd als één van de voorwaarden om zorg van goede kwaliteit te kunnen leveren.

In het onderzoek stond de vraag centraal naar de waarde en toepasbaarheid van kennistoetsen voor de deskundigheidsbevordering van huisartsen. Om deze vraag te kunnen beantwoorden, moeten de doelen van de toetsing geëxpliciteerd zijn. Van gegevens die worden verzameld met behulp van kennistoetsen op het gebied van de huisartsgeneeskunde, wordt verwacht dat ze het inzicht in sterke en zwakke kanten in het huisartsgeneeskundig kennisniveau vergroten. Daarnaast wordt er een ondersteunende en sturende rol van verwacht bij de planning van (nascholings-)onderwijsactiviteiten. Op basis van deze doelen is een aantal gebruiksfuncties onderscheiden.

- een *screeningsfunctie*: screening van het algemene kennisniveau van huisartsen moet verschillen in kennisniveau tussen diverse segmenten van de huisartsenpopulatie zichtbaar maken, teneinde een gerichtere planning en sturing van de nascholing mogelijk te maken;
- een *educatieve functie* voor de individuele huisarts: dat wil zeggen dat toetsgegevens moeten bijdragen aan het inzicht in sterke kanten en lacunes. Deze functie moet de individuele huisarts helpen de juiste keuzes te maken in het nascholingsaanbod;
- een *selectieve functie* in het kader van de waarde van de bijdrage van kennistoetsgegevens aan registratie-, c.q. herregistratiebeslissingen; en tenslotte
- een *evaluatieve functie* in de zin van de bruikbaarheid van toetsen om effecten van (nascholings-)onderwijs vast te stellen.

Om de toepasbaarheid van kennistoetsen voor deze functies te kunnen onderzoeken, zijn twee soorten kennistoetsen onderscheiden: *algemene* kennistoetsen, die het gehele huisartsgeneeskundige kennisdomein dekken, en *onderwerpgebonden* kennistoetsen.
De algemene kennistoets is erop gericht informatie op te leveren over de kennis over het gehele huisartsgeneeskundige vakgebied teneinde verschillen in kennis tussen individuele huisartsen of groepen van huisartsen te kunnen achterhalen. De algemene toets is daarmee vooral gericht op de screeningsfunctie. Door adequate feedback kan de toets tevens een educatieve functie hebben omdat de toetsresultaten aangeven waar individuele kennislacunes bestaan. Ook voor de selectieve functie geldt dat de toets het kennisdomein moet dekken.
De onderwerpgebonden toetsvorm is erop gericht informatie op te leveren over de kennis over bepaalde huisartsgeneeskundig relevante onderwerpen, c.q. verandering in kennis onder invloed van onderwijs over deze onderwerpen. Ook deze toetsvorm heeft een individuele educatieve functie. Daarnaast is deze toetsvorm op groepsniveau vooral gericht op de evaluatieve functie van kennistoetsing. Voor beide toetsvormen zijn instrumenten ontwikkeld en onderzocht op hun kwaliteit als meetinstrumenten in termen van validiteit, betrouwbaarheid en praktische toepasbaarheid.

In *Hoofdstuk 1* wordt een overzicht gegeven van de achtergronden van het onderzoek en van de literatuur over het gebruik van kennistoetsen in de huisartsgeneeskundige nascholing. Dit mondt uit in de probleemstelling en in de beschrijving van de onderzoeksdoelen. Het hoofdstuk wordt afgesloten met een overzicht van de onderzoeksvragen.
De beschrijving van het onderzoek valt uiteen in vier onderdelen. Deel I beschrijft de algemene huisartsgeneeskundige kennistoetsing, waarvoor is aangesloten op de landelijke kennistoetsontwikkeling binnen de huisartsopleiding. In Deel II wordt beschreven welke samenhangen zijn gevonden tussen kenmerken van huisartsen en hun scores op een algemene kennistoets. Deel III betreft de onderwerpgebonden kennistoetsing. Als exemplarisch voorbeeld zijn kennistoetsen ontwikkeld bij deskundigheidsbevorderingspakketten behorend bij NHG-Standaarden. Daarbij is aansluiting gezocht bij lopende nascholings-activiteiten op het gebied van 'vaginaal bloedverlies' en 'cholesterol'. In Deel IV tenslotte wordt nader ingegaan op de invloed van de vraagvorm op de huisartsgeneeskundige kennistoetsen.

Deel I: Algemene huisartsgeneeskundige kennistoetsing

De *validiteitsstudie* wordt beschreven in *Hoofdstuk 2*. Deze is gericht op de inhouds- en de constructvaliditeit van de toets. De inhoudsvaliditeit van de toets wordt ondersteund door de motivatie en verantwoording van de inhoudsstructuur van de toets. Daartoe is een blauwdruk vastgesteld. Deze moet ertoe leiden dat de toets, vanuit verschillende perspectieven, een evenwichtige afspiegeling vormt van het huisartsgeneeskundig kennisdomein. De blauwdruk is vastgesteld in een stapsgewijze consensusprocedure met experts op het terrein van de huisartsgeneeskundige scholing en toetsing. In de blauwdruk fungeren de huisartsgeneeskundig-inhoudelijke klachten- en aandoeningencomponenten van de International Classification of Primary Care (ICPC) als hoofdindeling. Deze zijn aangevuld met een huisartsgeneeskundig-theoretisch onderdeel. Daarnaast zorgen vier nevenindelingen ervoor dat belangrijke huisartsgeneeskundige invalshoeken, die niet worden gedefinieerd door de ICPC, in voldoende mate in de toets zijn vertegenwoordigd. Dit zijn leeftijdscategorieën van patiënten, aspecten van het consult, spoedeisende gevallen en chronische aandoeningen.

De totale toetslengte is bepaald op 160 vragen, overeenkomend met een geschatte toetsduur van twee uur. De omvang van de blauwdrukcategorieën is vastgesteld op praktische en inhoudelijke gronden. In het onderzoek is gecontroleerd of veranderingen in de relatieve omvang van de hoofdstukken van de blauwdruk van invloed was op de totale toetsbetrouwbaarheid. Dit bleek niet zo te zijn. Dit is beschouwd als een empirische ondersteuning van de inhoudsvaliditeit van de toets.
Om de constructvaliditeit te onderzoeken, is een op basis van de blauwdruk samengestelde kennistoets voorgelegd aan groepen co-assistenten huisartsgeneeskunde, huisartsen-in-opleiding in diverse opleidingsfasen, en aan ervaren huisartsen in verschillende fasen van hun beroepscarrière. Om het kennisbegrip waarover een praktizerend huisarts moet beschikken, goed te kunnen omschrijven, is aansluiting gezocht bij de theorievorming over de ontwikkeling van medische expertise. Deze theorievorming is voortgekomen uit cognitief-psychologische inzichten over de samenhang tussen kennisontwikkeling, het oplossen van problemen en het opdoen van praktijkervaring. Op basis daarvan is het kennisbegrip omschreven als *gegeneraliseerde huisartsgeneeskundige ervaringskennis*. Het is de kennis waarop de huisartsen in beginsel aanspreekbaar zijn. De toetsvragen moeten een beroep doen op huisartsgeneeskundig relevante kenniselementen, ook wel aangeduid als '*key features*'. De vragen moeten geloofwaardig zijn en gaan over relevante medische aspecten in de context van patiëntenproblematiek. Als de kennistoets een adequate operationalisatie vormt van dit theoretische kennisbegrip, kon worden verwacht dat het kennisniveau, uitgedrukt in gemiddelde toetsscore, van groepen artsen in verschillende fasen van opleiding en ervaring, geleidelijk hoger zou worden. Ervan uitgaande dat de ontwikkeling van gerichte kennis over huisartsgeneeskundig-inhoudelijke concepten het meest geconcentreerd plaatsvindt tijdens de huisartsopleiding, werd verwacht dat het kennisniveau het hoogst zou zijn aan het eind van de huisartsopleiding. Ervaren huisartsen bleken als groep inderdaad gemiddeld significant hogere scores te behalen dan de groep huisartsen-in-opleiding, die op hun beurt weer significant hogere gemiddelde scores behaalden dan de co-assistenten huisartsgeneeskunde. De ervaren huisartsen bereikten een gemiddelde toetsscore die ongeveer op het niveau ligt dat door huisartsen-in-opleiding vlak voor het einde van de huisartsopleiding wordt bereikt. Deze bevindingen worden beschouwd als ondersteuning van de constructvaliditeit van de algemene huisartsgeneeskundige kennistoets.

De betrouwbaarheid maakt duidelijk hoeveel 'ruis' of meetfouten in de score zijn verdisconteerd. In *Hoofdstuk 3* wordt ingegaan op de *betrouwbaarheid* van de scores. De betrouwbaarheid is vanuit drie perspectieven geïnterpreteerd: relatief, absoluut en beslissings-georiënteerd. Het relatieve perspectief is vooral van belang bij gebruik van de toets voor screening en educatie. Het absolute perspectief is van belang als het gaat om absolute beheersing van de getoetste stof, dat wil zeggen als het gaat om de vraag of aan de scores een absolute betekenis kan worden toegekend. Het beslissingsperspectief speelt een rol bij de betrouwbaarheid van zak-slaag beslissingen, dat wil zegen bij selectie of (her)registratie. Omdat de generaliseerbaarheidstheorie het mogelijk maakt verschillende betrouwbaarheidsindices te berekenen voor alle onderscheiden perspectieven, is het onderzoek gebaseerd op deze theorie. Alle berekeningen zijn gebaseerd op goed-min-foutscores, uitgedrukt in percentages van de maximaal te bereiken score.
De gemiddelde toetsscore van de huisartsen is 45% met een standaarddeviatie van 11%. De betrouwbaarheid van individuele toetsscores vanuit relatief perspectief bedraagt .70 met een standaardmeetfout van 6. Dat betekent een foutenmarge (95%-betrouwbaarheidsinterval) voor

de individuele toetsscores van 12 punten die moeten worden opgeteld of afgetrokken van de individuele score. Vanuit absoluut perspectief bedraagt deze marge plus of min 14 punten. Verdubbeling van de toets tot 320 vragen (vier uur toetstijd) levert een betrouwbaarheid op van .80 voor beide perspectieven, maar de marge (95%-betrouwbaarheidsinterval) bedraagt dan nog steeds plus of min 8, respectievelijk 10. Betrouwbare zak-slaag beslissingen (met een betrouwbaarheidscoëfficiënt van .80) kunnen worden genomen op basis van een cesuur die ligt onder een goed-min-foutscore van 35% of boven die van 55%. De betrouwbaarheid van gemiddelde scores liet een veel gunstiger beeld zien. Het 95%-betrouwbaarheidsinterval bedroeg 2, resp. 6 punten voor het relatieve en het absolute perspectief, bij een groepsgrootte van ongeveer 30 huisartsen.
Op basis van de betrouwbaarheidsschattingen wordt geconcludeerd dat de kennistoets voor screeningsdoelen voldoende betrouwbare gegevens oplevert. Ook voor educatieve doelen is de toets goed bruikbaar. Bij de interpretatie van individuele scores moet rekening worden gehouden met een te grote foutenmarge om zak-slaagconsequenties aan het toetsresultaat te kunnen verbinden. De betrouwbaarheid van beslissingen varieert met de plaats van de cesuur en is hoger naarmate deze verder van de gemiddelde score af ligt. Omdat onduidelijk is welk minimumniveau aan kennis nog verenigbaar is met het goed functioneren als huisarts, is op grond van dit onderzoek geen goede uitspraak te doen over de beste plaats van de cesuur voldoende/onvoldoende.

Op de *praktische toepasbaarheid* van algemene huisartsgeneeskundige kennistoetsing wordt ingegaan in *Hoofdstuk 4*. Een op de blauwdruk gebaseerde toets is voorgelegd aan ruim 350 ervaren huisartsen en hen is gevraagd naar hun opinie over de kwaliteit en de toepasbaarheid van de toets. De huisartsen oordeelden in het algemeen positief over de toets en achtten de toets huisartsgeneeskundig relevant. De huisartsen vonden de toets het beste bruikbaar als instrument voor zelfbeoordeling. Het invullen van een toets van 160 vragen kost gemiddeld ongeveer anderhalf uur, hetgeen vanwege het feit dat bijna eenderde van de deelnemende huisartsen de toets te lang vond, het maximaal haalbare lijkt.
Wat betreft de toetsconstructie en organisatie van grootschalige toetsafnames, wordt geconcludeerd dat de benodigde constructietijd (gemiddeld 70 minuten per item, uitgevoerd door een ervaren huisarts-toetsvragenschrijver) en de voorwaarden die een de gegevensverwerking en feedbackverzorging vergen, vereisen dat deze worden uitgevoerd door een centrale organisatie met een adequate outillage.

Deel II: Kennis in relatie tot achtergrondkenmerken van huisartsen

Het verband tussen de algemene kennis van huisartsen en hun achtergrondkenmerken wordt weergegeven in *Hoofdstuk 5*. Dit verband is onderzocht door de gemiddelde toetsscores van ervaren huisartsen te relateren aan een aantal persoonskenmerken, demografische en professionele kenmerken.
Naarmate huisartsen meer ervaring hebben, is hun gemiddelde kennistoetsscore lager. Alleen lidmaatschap van het NHG en opleiderservaring vertonen een significant positief verband met de gemiddelde kennistoetsscore. Er kon geen relatie worden vastgesteld tussen werkverband of praktijksituering en kennisniveau.
De scores voor met name medisch-inhoudelijke thema's doen vermoeden dat de kennis daarvan nauw samenhangt met de mate waarin huisartsen er in de praktijk mee worden geconfronteerd. Dit geldt bijvoorbeeld voor aandoeningen van het bewegingsapparaat of

huidaandoeningen. Hetzelfde wordt gevonden voor 'medicamenteus beleid'. Het vermoeden dat kennis over bepaalde onderwerpen en daadwerkelijke praktijkervaring met deze thema's samenhangen, wordt ondersteund door de bevinding dat huisartsen met praktijken met relatief veel ouderen, gemiddeld significant hogere scores behalen op de casuïstiek, die betrekking heeft op 'ouderen (boven 75 jaar)' en 'chronische aandoeningen'. Huisartsen met veel jonge gezinnen in de praktijk scoren gemiddeld wel hoger dan hun collega's op de categorie 'jongeren beneden 15 jaar', maar dit verschil is niet significant.

Hoofdstuk 6 betreft het verband tussen de zelfkennis van huisartsen en hun scores op de algemene kennistoets. De toetsscores van de huisartsen zijn daartoe vergeleken met het oordeel dat ze over hun eigen expertise gaven op de onderwerpen uit de kennistoets.
Gemiddeld gaven de huisartsen ruim twee expertisegebieden aan. Samenhang tussen het aantal expertisegebieden en de gemiddelde toetsscore kon niet worden vastgesteld. Ruim 40% van de huisartsen (en zelfs ruim 70% van de vrouwelijke huisartsen) gaf aan zich meer deskundig te achten dan collega's op het gebied van psychosociale aandoeningen. Deze groep huisartsen behaalde op dit terrein echter geen significant hogere scores dan de overige huisartsen. Ook de vrouwelijke huisartsen behaalden geen hogere scores dan hun mannelijke collega's op dit gebied. De conclusie is dat er geen eenduidige relatie is gevonden tussen kennisscore en inschatting van eigen expertise: de totale gemiddelde scores van de huisartsen hangen niet significant samen met het aantal gebieden waarop de huisartsen zich deskundig achten. Als wordt gekeken naar specifieke onderwerpen, scoren de huisartsen die zich expert noemen op die onderwerpen meestal op die gebieden wél iets hoger dan hun collega's die zich geen expert voelen. De verschillen zijn echter meestal niet significant en het verband ligt ook wel eens andersom.

Deel III: Onderwerpgebonden huisartsgeneeskundige kennistoetsen

De *validiteit* van onderwerpgebonden kennistoetsen die zijn bedoeld voor gebruik in de nascholing van huisartsen, wordt beschreven in *Hoofdstuk 7*. De validiteitsstudie was gericht op het aantonen van de inhoudsvaliditeit en de constructvaliditeit van deze toetsen. Inhoudsvaliditeit is nagestreefd door de inhoud van de toetsen zo goed mogelijk af te stemmen op de kenniselementen die door deskundigen op het betreffende gebied wordt beoordeeld als de parate kennis waarover de praktizerend huisarts moet beschikken over het betreffende onderwerp. Voor de kennistoets over cholesterol werd een speciale procedure toegepast waarbij de fundamentele kenniselementen werden vastgesteld in een consensusprocedure, waarbij de samenstellers van de NHG-Standaard Cholesterol werden betrokken.
De constructvaliditeit van de toetsen is onderzocht door paralleltoetsen voor te leggen aan huisartsen, direct voorafgaand aan en meteen na afloop van nascholing over het betreffende thema. Voor het onderwerp 'cholesterol' is daarnaast een retentiemeting verricht ongeveer acht maanden na het nascholingsprogramma. In de constructvalidering van de toets over cholesterol was ook een vergelijkingsgroep betrokken. Bovendien deden niet alleen huisartsen maar ook huisartsen-in-opleiding mee aan de toetsing en nascholing over cholesterol.
De gemiddelde voortoetsscore van huisartsen is, zoals verwacht, significant hoger dan die van de huisartsen-in-opleiding. Direct na de nascholing bereiken beide groepen een even hoge gemiddelde score. De gemiddelde score op de retentietoets van de groepen die de nascholing hadden gevolgd, is significant hoger dan hun aanvankelijke score, al is de score in vergelijking met de score onmiddellijk na de nascholing wel gedaald. De gemiddelde score

van de vergelijkingsgroepen verandert niet. Uit een stapsgewijze multipele regressie-analyse blijkt dat alleen de score op de voortoets en het al dan niet bijgewoond hebben van het nascholingsprogramma, een significante invloed hebben op de bereikte retentiescore.
De conclusie uit de validiteitsstudie is dat het goed mogelijk is inhoudsvalide onderwerpgebonden kennistoetsen te construeren waarmee op groepsniveau verschillen kunnen worden gemeten die samenhangen met deelname aan een nascholingsprogramma. Deze laatste bevinding ondersteunt de constructvaliditeit van de toetsen.

Hoofdstuk 8 handelt over de *betrouwbaarheid* van de onderwerpgebonden toetsen. Deze is onderzocht op basis van dezelfde theoretische uitgangspunten als de algemene kennistoets. Conform de verwachting blijkt de betrouwbaarheid zowel vanuit het relatieve als het absolute perspectief te laag om consequenties aan individuele scores te kunnen verbinden. Op individueel niveau bieden de onderwerpgebonden kennistoetsen geen betrouwbare informatie over de kennisbeheersing. Dit staat overigens een zinvol educatief gebruik van de toetsen niet in de weg. De gemiddelde scores zijn betrouwbaar genoeg om uitspraken te kunnen doen over de evaluatieve functie van de onderwerpgebonden toetsen. Met de toetsen is het op groepsniveau goed mogelijk om kennisverschillen in samenhang met het volgen van nascholingsprogramma's vast te stellen.

De *praktische toepasbaarheid* van onderwerpgebonden kennistoetsing wordt beschreven in *Hoofdstuk 9*. Deze is onderzocht door het oordeel van de doelgroep, praktizerende huisartsen, in te winnen over de waarde van het gebruik van kennistoetsen bij nascholingsprogramma's. Daarnaast is evenals bij de algemene huisartsgeneeskundige kennistoetsing, de uitvoerbaarheid van de toetsing nagegaan door de tijd en menskracht te inventariseren die nodig is om toetsen te construeren, door de huisartsen te vragen hoeveel tijd het invullen van een toets kost en door hun oordeel over de zin en bruikbaarheid van dergelijke toetsen in te winnen, en tenslotte door na te gaan in hoeverre analyse en verwerking van de toetsgegevens en feedbacksamenstelling praktisch uitvoerbaar zijn. De beschikbaarheid van de NHG-Standaarden, vooral als daarnaast kan worden beschikt over een lijst met essentiële kenniselementen uit de Standaarden, blijkt een efficiënt hulpmiddel bij de vraagconstructie. De huisartsen beoordelen kennistoetsing bij nascholing vrijwel unaniem als een zinvolle activiteit, vooral als smaakmaker of 'warming up' voor de nascholing. Onderwerpgebonden kennistoetsen zijn volgens de gebruikers niet zo geschikt als hulpmiddel bij de keuze van bepaalde nascholingsprogramma's. Wat betreft de verwerking en analyse van de toetsgegevens en het samenstellen van feedback geldt, evenals is geconcludeerd bij de algemene kennistoets, dat dit onder de voorwaarde dat adequate gegevensverwerkende apparatuur en programmatuur beschikbaar is, geen problemen oplevert.

Deel IV: Vraagvorm in huisartsgeneeskundige kennistoetsen

Het laatste onderzoeksdeel betreft een studie die is uitgevoerd naar de invloed van verschillende vraagvormen die in breed toepasbare huisartsgeneeskundige kennistoetsen kunnen worden gehanteerd, op de kwaliteit van het meetinstrument. Het resultaat wordt beschreven in *Hoofdstuk 10*. De invloed is nagegaan van het gebruik van de juist-onjuist-vraagtekenvraagvorm op de validiteit, de betrouwbaarheid en de praktische toepasbaarheid van kennistoetsen, in vergelijking met het gebruik van traditionele meerkeuzevragen. Daartoe is voor de juist-onjuist-vraagteken-kennistoets over cholesterol een parallelversie samengesteld

met meerkeuzevragen. Beide toetsen zijn voorgelegd aan ervaren huisartsen. De indeling van huisartsen aan de groep die hetzij de juist-onjuist-vraagtekenvorm, hetzij de meerkeuzevorm voorgelegd kreeg, vond plaats volgens een random procedure. Achtergrondkenmerken van de huisartsen zijn verzameld om na te gaan in hoeverre beide groepen vergelijkbaar waren. Daarnaast is nagegaan of de acceptabiliteit van beide vraagvormen verschillend is.
De resultaten wijzen er niet op dat een keuze voor een van beide vormen evidente voordelen biedt. Voor het toetsen van bepaalde kenniselementen lijkt om inhoudelijke redenen de meerkeuzevorm geschikter. Voor andere elementen moet echter naar kunstmatige oplossingen worden gezocht om ze in de vorm van een vierkeuzevraag te kunnen formuleren. De betrouwbaarheidsschattingen laten voor de meerkeuzevorm een iets gunstiger beeld zien dan voor de juist-onjuist-vraagtekenvorm. Een toets bestaande uit meerkeuzevragen vergt echter een langere invultijd dan een toets in de juist-onjuist-vraagtekenvorm. De acceptabiliteit van beide vormen is ongeveer gelijk, al was de non-respons voor de juist-onjuist-vraagtekenvorm groter dan voor de meerkeuzevorm. Geconcludeerd wordt dat er geen argumenten zijn om, als dat op inhoudelijke gronden wenselijk is, niet te variëren met de vraagvorm. Deze conclusie is in overeenstemming met de literatuur die aangeeft dat met betrekking tot de inhoudelijke en psychometrische kwaliteit van schriftelijke kennistoetsen de vorm ondergeschikt is aan de inhoud.

Conclusie

In *Hoofdstuk 11* worden de onderzoeksresultaten aan een nadere beschouwing onderworpen. Daarbij wordt vooral ingegaan op de bruikbaarheid van de kennistoetsgegevens voor de verschillende gebruiksfuncties die in de inleiding zijn geformuleerd. Geconcludeerd wordt dat de algemene huisartsgeneeskundige kennistoets een goed instrument lijkt voor screening op groepsniveau. Wat betreft de educatieve functie is de toets als zelfevaluatie-instrument en als leermiddel zeer bruikbaar voor huisartsen. De betrouwbaarheid van de individuele toetsscores is bij de huidige toetslengte te laag om de toets zonder meer toe te passen in het kader van selectie- of herregistratiebeslissingen. Wat betreft de onderwerpgebonden toetsen wordt geconcludeerd dat de kennisdomeinen, die de NHG-Standaarden en bijbehorende deskundigheidsbevorderingspakketten bestrijken, te beperkt zijn om er kennistoetsen op te baseren die voldoende betrouwbare resultaten opleveren om er op individueel niveau conclusies aan te kunnen verbinden. De toetsen kunnen echter een nuttige rol vervullen in de voorbereiding op nascholingsprogramma's. De onderwerpgebonden toetsen laten zien dat effecten van nascholing in termen van kennisverandering na verloop van tijd lijken weg te ebben. Dit gegeven leidt tot de aanbeveling dergelijk onderzoek voor verschillende onderwerpen te herhalen. Daarnaast moet worden nagegaan in hoeverre verschillende onderwijsvormen verschillende effecten bewerkstelligen. Onderwerpgebonden kennistoetsen vormen een goed instrument om dergelijk onderzoek mee uit te voeren.

SUMMARY

Introduction

Quality assurance in general practice has many facets. One of these is the development of assessment instruments aimed at enhancing the efficiency and efficacy of continuing education in general practice. One of the goals of assessment procedures is to help the individual general practitioner in making the right choices in continuing medical education. If doctors chose topics for continuing medical education of their own accord, there is a considerable risk of their choosing topics that are unnecessary from an objective point of view. Self-perceived learning needs tend to match personal interests and specific experiences. Another application for objective tests lies in the field of planning and developing of continuing medical education. The use of collective test results can enhance the quality of this process. If the 'true' learning needs and gaps in knowledge of general practitioners are known, planning and development of educational programmes can be better guided. Finally, objective test results can be used to show society as a whole that the professional ability of general practitioners is being maintained at a high level.
We have studied the various goals of assessment for one area of the competence of general practitioners, i.e. that of knowledge. An adequate level of knowledge is considered an important prerequisite for quality of care in general practice.

Our main research question concerns the suitability and applicability of knowledge tests in continuing medical education for general practitioners. To answer this question testing purposes must be clear. Objectively collected information should increase the insight to strengths and weaknesses in general practice knowledge. Besides, it should support and guide the planning of educational activities in continuing education. Based on these purposes, we distinguished the following areas of application of knowledge tests:

- *screening*: obtaining information about the general knowledge level in various segments of the population of general practitioners to enhance planning and guiding of continuing education;
- *education*: giving individual general practitioners feedback on strengths and weaknesses to help them make the right choices from the educational programmes offered;
- *selection*: knowledge tests as an element in certification and recertification decisions; and
- *evaluation*: assessing the usefulness of knowledge tests in establishing the effectiveness of continuing education programmes.

To meet these goals two types of knowledge tests have been developed: a *general* test covering the whole field of general practice, and *specific* knowledge tests covering certain specific topics. The general test is meant both for screening of the whole population and for giving feedback to individual general practitioners. To serve selective purposes the test should also cover the whole domain of general practice.

The specific tests can be used for individual feedback about one's knowledge of specific topics. It can also be used at group level to assess the effectiveness of a particular continuing medical education course. The validity, reliability and applicability of both the general and the specific test have been assessed.

Chapter 1 gives an overview of the background of the study and of the literature on knowledge testing in continuing education for general practice, culminating in the problem definition and research questions.

The study is made of four parts. Part I is a description of the construction and validation of the general knowledge test. This part of the study has been carried out in close cooperation with a study on the usefulness and applicability of the general knowledge test in vocational training for general practice. In Part II the relation between characteristics of general practitioners and their scores on the general knowledge test is described. The specific tests are presented in Part III. Two subjects were chosen in order to study the validity, reliability and applicability of these tests, 'vaginal bloodloss' and 'cholesterol'. These knowledge tests are related to national standards of the Dutch College of General Practitioners. In Part IV the effects of varying the item format in knowledge tests for general practice is reported.

Part I: General knowledge test

Chapter 2 covers the content and construct *validity* of the test. To guarantee high content validity a blueprint was developed for the construction of the test. This blueprint ensures a balanced representation of the domain of general practice knowledge from different perspectives. The blueprint was developed in a stepwise consensus procedure by experts in the field of continuing education and assessment in general practice. The chapters of the International Classification of Primary Care (ICPC) concerning symptoms and complaints, and diagnoses and diseases, were adopted as the main classification in the blueprint. A chapter on theoretical topics of relevance to general practice, covering non-clinical aspects, was added. Furthermore, four additional classifications ensure the representation of important issues in general practice which are not defined in the ICPC. These are 'age of patients', 'aspects of consultation' (diagnostics, management), 'emergency medicine' and 'chronic diseases'. The total test consists of 160 items, resulting in an average testing time of two hours. The size of the blueprint chapters was set using criteria of practicality as well as content. Reliability studies, in which the relative sizes of all chapters were varied, showed no influence of chapter size on reliability, which supports the high content validity of the test in yet another way.

A general knowledge test, constructed according to the blueprint, was taken by medical students in their clerkships, trainees during vocational training for general practice at various levels, and experienced general practitioners in the various phases of their professional careers. Theories from the field of cognitive psychology on changes in knowledge and problem solving skills on gaining experience in daily practice, were used to specify the knowledge concept in the test. We have described it as *generalized expert knowledge on general practice*. This is the knowledge a general practitioner is assumed to master at the end of vocational training. The items in the knowledge test should focus on 'key features', knowledge elements that are crucial to general practice care. Furthermore, the items should cover relevant aspects of realistic patient problems.

If indeed the knowledge test is an adequate representation of this theoretical knowledge concept, it was expected that the mean test scores would gradually increase with training and/or experience. As the acquisition of specific knowledge of relevant concepts of general practice takes place in a more intensive manner during general practice training than thereafter, we postulated that the scores would be highest at the end of training. Indeed, students during their clerkships had the lowest scores and the mean scores of all experienced general practitioners surpassed the mean scores of all trainees in general practice. The general practitioners had an average score at the level of the group of trainees just before the end of vocational training. These findings support the construct validity of the general knowledge test.

The reliability of a test elucidates the amount of measurement error which has been taken into account in the scores. *Chapter 3* covers the *reliability* of the test scores of the general knowledge test from three perspectives: relative (norm), absolute (domain) and decision (mastery)oriented. For screening and educational purposes the relative-oriented perspective is the most relevant. When one is interested in how well the scores actually represent mastery of the knowledge, the absolute-oriented perspective is paramount. Lastly, for pass-fail decisions, such as those for (re)certification, the decision-oriented perspective is used. Because generalizability theory provides a framework to estimate reliability indices according to all three perspectives, the reliability study is based on this theory. All estimates are based on correct-minus-incorrect scores, expressed as percentages of the maximum score.
The average test score for general practitioners is 45% with a standard deviation of 11%. The reliability of individual test scores from a relative perspective is .70 with a standard error of measurement (SEM) of 6. This means that the 95%-confidence interval of an individual score lies between 12 points added to or subtracted from the individual score. From the absolute perspective this interval is plus or minus 14 points. Doubling the test size to to 320 items (four hours testing time) increases the estimated reliability coefficient to .80 from both perspectives but still leaves one with a 95%-confidence interval of plus or minus 8 to 10 points. Reliable pass-fail decisions (with a generalizability coefficient of .80) can be taken with either a fail score below 35% or a pass score over 55% (correct-minus-incorrect).
The reliability of group mean scores is far better, with a 95%-confidence interval of 2 and 6 points for the relative and absolute perspective respectively, assuming a group size of about 30 general practitioners.
In conclusion, the reliability of test scores for screening purposes is sufficient. This is also the case when the scores are used for educational purposes. However, the reliability of the individual scores is not high enough for pass-fail decisions. The reliability of pass-fail decisions varies according to where the cut-off score is placed and is higher, the further the cut-off score is from the average score. As there is no evidence so far as to which level of knowledge is a prerequisite for the good performance of a general practitioner in daily practice, it is impossible to pronounce upon the best place for a cut-off score for pass-fail decisions.

Chapter 4 describes the *applicability* of the general knowledge test. The knowledge test based on the blueprint was taken by more than 350 general practitioners. They were asked for their opinion of the quality and the applicability of the test. On the whole, they were positive about the test, and considered it relevant to daily practice. They thought the test most useful as an instrument for self-evaluation. One-third of the participants considered the time it took to take

the test (1½ to 2 hours) too long. This limits the feasibility of increasing the number of items per test. As to the organization of test construction and test taking, the following conclusions can be drawn. With an average time of 70 minutes for the construction of one test item, when done by an experienced general practitioner-test writer, and considering the ample resources needed for data processing and compiling and returning of feedback, the use of this test is only feasible on a large scale when done by a central organization with extensive resources and equipment.

Part II: Knowledge related to background characteristics of general practitioners

The relationship between the knowledge of general practitioners and their background characteristics is described in *Chapter 5*. Personal, demographic and professional characteristics of general practitioners were related to their knowledge as measured with the general test. As general practitioners gain more experience their average score gradually decreases. Membership of the Dutch College of General Practitioners and experience as a trainer were the only two variables that had a significant positive correlation with the test score. No relation was found for type of practice or geographical situation and test score. Especially the scores on patient-bound themes suggest that this kind of knowledge is closely related to the degree with which general practitioners are confronted with the themes in their daily practice, for example musculoskeletal or skin problems. The same was found for the part of the test that concerned 'medication'. The assumption of a relation between knowledge on certain issues and actual practice experience with these issues, is supported by the finding that general practitioners with relatively many elderly patients, achieve significantly higher mean scores on items on 'elderly patients (age over 75)'. The mean score of general practitioners with relatively more young families in their practice is higher than the mean score of their colleagues on items on 'young people (age under 15)', but the difference is not statistically significant.

Chapter 6 describes the relationship between self-knowledge of general practitioners and their scores on the general knowledge test. The test scores of the general practitioners are compared with their judgment on their own expertise on topics of the test. On average the general practitioners reported specific expertise on more than two subjects. No relation was found for the number of 'expertise subjects' and mean test score. More than 40% of the general practitioners (even more than 70% of the female general practitioners) reported being more expert than their colleagues on the subject of psychosocial problems. This group of general practitioners, however, did not achieve significantly higher scores on this subject than the other general practitioners. Moreover, the female general practitioners did not achieve higher scores on this subject than their male colleagues.
The conclusion is that no unambiguous relationship was found between test scores and the assessment of own expertise. The average score of 'experts' is mostly slightly higher than the score of 'non-experts' on a certain subject. However, the differences were seldom significant and opposite relationships were found as well.

Part III: Specific general practice knowledge tests

The *validity* of specific knowledge tests meant for use in the continuing education of general practitioners is described in *Chapter 7*. The validity study was aimed at establishing content

validity and construct validity. Content validity has been sought by attuning the content of the tests as closely as possible to the knowledge elements, which experts in the area judged as the necessary knowledge relevant to the day-to-day practice of a general practitioner. A special procedure was adopted for the development of the knowledge test on cholesterol. The essential knowledge elements were determined in a consensus procedure which included the composers of the standard on cholesterol. The construct validity of the tests has been investigated by submitting parallel tests to general practitioners, immediately before and directly after continuing education on the theme of the test. On top of this a follow-up post test eight months after the continuing education, was done for the theme 'cholesterol'. The design of the study on construct validity of the cholesterol test involved the use of comparison groups. In addition, not only general practitioners but also trainees in general practice were involved in the education and testing on cholesterol. The mean test score of the general practitioners before the education programme took place was, as expected, significantly higher than the mean score of the trainees. Immediately after following the educational programme both groups achieved the same mean test score. The mean follow-up test score of the groups that followed the educational programme was significantly higher than their mean score at the onset. The mean score after eight months showed a decrease in comparison to the score after following the programme. The mean scores of the comparison groups did not change. A stepwise multiple regression analysis showed that only the initial scores and the participation in the educational programme had a significant influence on the score after eight months.
The conclusion from the validity study is that the construction of a content-valid specific knowledge test on a particular topic is feasible. Also group differences related to the participation in continuing education programmes can be measured with this test. This last finding supports the construct validity of the tests.

Chapter 8 concerns the *reliability* of the specific tests. The investigation of the reliability of the tests was based on the same theoretical assumptions used in the overall knowledge test. As expected, reliability turns out to be too low to attach consequences to individual test scores, from the relative as well as from the absolute perspective. On an individual level the specific knowledge tests do not give reliable information. This does, however, not hinder meaningful educational use of the tests. The mean group scores are reliable enough to warrant the use of the specific tests in evaluation studies. The tests make it readily possible to bring about changes in knowledge related to continuing education on group level.

The *applicability* of specific knowledge tests is described in *Chapter 9*. This aspect was investigated by obtaining the judgment of the practising general practitioners on the practical value of knowledge tests in continuing education. Furthermore, the feasibility of testing was verified by surveying time and manpower necessary for the construction of tests, by asking the general practitioners how much time was needed to fill in a test, and finally, by investigating the feasibility of data processing and the production of feedback.
The availability of the national standards, and the list with essential knowledge elements distilled from the standards proved to be an efficient aid to the construction of test items. Nearly all general practitioners found the testing of knowledge in combination with continuing education worthwhile, especially as a 'warming-up' to the education programme. They did not find the specific knowledge tests very useful in helping to chose between continuing educational programmes. As to data processing and production of feedback, the same

conclusions as those of the overall knowledge test apply. Provided adequate data processing hardware and software are available, there are no important impediments.

Part IV: Question format of general practice knowledge test

The final part of our investigation concerns a study of the influence of two different question formats that can both be used in broadly applicable general practice knowledge tests, on the quality of the instruments. The results are described in *Chapter 10*. The influence of the application of the true-false-question mark item format, compared to the application of traditional multiple choice questions, on validity, reliability and practical applicability of knowledge tests is examined. A parallel version to the true-false-question mark version of the cholesterol test was constructed in the multiple choice format. Both test versions were taken by experienced general practitioners. General practitioners were randomly allocated to groups that got the true-false-question mark version or the multiple choice version. Differences in the acceptability of both question formats were investigated.
The results do not indicate that a choice for one of both formats yields obvious advantages. For content related reasons the multiple choice format seems more appropriate for the testing of certain knowledge elements. However, for some other elements the multiple choice format requires artificial alternative choices. Reliability estimates show a slightly better picture for the multiple choice format compared to the true-false-question mark format. However, it takes longer to do a test that consists of multiple choice items than a test in the true-false-question mark format. The acceptability of both formats is similar, but the non-response for the true-false-question mark format was higher than for the multiple choice format. The conclusion is that no proof has been found that variation in question format is unwise, if content-related arguments make variation advisable. This conclusion is in line with the literature which indicates that regarding the psychometric quality of tests, the format of written knowledge tests is of secondary importance to the content.

Conclusion

In *Chapter 11* the overall results of the research are examined more closely. Attention focuses on the usefulness of the knowledge test results for the different purposes that have been distinguished. The conclusion is that the overall general practice knowledge test seems an adequate instrument for the screening of groups. The test is very useful for general practitioners as an instrument for self-evaluation and as an educational tool. The reliability of individual test scores using the current test is too low to apply the test as part of selection or recertification decisions. Concerning the specific tests, the conclusion is that the knowledge areas as covered by the national standards and their corresponding educational programmes are too limited to function as a basis for knowledge tests large enough to be sufficiently reliable to allow conclusions on an individual level. However, the tests may be very useful in the preparation of continuing education. The specific tests indicate that the effects of continuing education in terms of knowledge changes seem to fade away with the lapse of time. This fact leads to the recommendation to repeat this type of evaluation research for other themes. Besides that, differences in impact of different educational methods should be investigated. Specific knowledge tests form an adequate instrument to carry out this type of research.