

Toetsing van professionele gespreksvaardigheden in het onderwijs

G.N. Smit

Samenvatting

Voor toetsing van gespreksvaardigheden in het onderwijs wordt de gesprekssimulatie veel gebruikt. Met deze methode kan direct worden geobserveerd of een student de gespreksvaardigheden beheerst. De interbeoordelaarsovereenstemming tussen beoordelaars is bij dit soort toetsen hoog mits de beoordelaars terdege op hun taak zijn voorbereid. Er zijn echter zo'n tien simulaties nodig om met enige zekerheid uitspraken te kunnen doen hoe goed een student de gespreksvaardigheden beheerst. Ten aanzien van de validiteit geven onderzoeksresultaten aanwijzingen dat werkelijk gespreksvaardigheden worden beoordeeld. Schriftelijke toetsvormen of toetsvormen waarin gebruik wordt gemaakt van audiovisuele media zijn bruikbaar voor toetsing van inzicht in een professioneel gesprek en voor toetsing van het kunnen toepassen van deelvaardigheden zoals vragen stellen. De betrouwbaarheid van dit soort toetsvormen is hoog. Daarnaast bieden onderzoeksresultaten een ondersteuning voor hun validiteit. Aanbevolen wordt om simulaties aan te vullen met audiovisuele toetsen om zo een goede en hanteerbare toetsingsprocedure te verkrijgen.

Inleiding

Van oudsher worden in het onderwijs kennis en praktisch-technische vaardigheden bijgebracht om studenten voor te bereiden op de beroepspraktijk. Het inzicht dat ook gespreksvaardigheden getraind kunnen worden, is van veel recenter datum.¹ Vanaf de jaren zestig verschijnen regelmatig boeken over gespreks-

voering in het algemeen en arts-patiëntcommunicatie in het bijzonder.²⁻⁴ Van Dalen omschrijft gespreksvaardigheden in de medische situatie als die vaardigheden waarmee

- een arts-patiëntrelatie wordt opgebouwd en onderhouden;
- verbale informatie wordt ingewonnen die relevant is voor het verhelderen en oplossen van het probleem waarmee de patiënt komt;
- informatie wordt verstrekt over het hulpaanbod.⁵

Van Dalen gebruikt de term vaardigheden, terwijl het gaat om vaardigheid in het voeren van gesprekken. Het inzicht dat een gesprek in kleine eenheden kan worden opgedeeld is een grote doorbraak geweest. Onder andere Ivey heeft in de jaren zeventig veel hulpverleningsgesprekken bijgewoond.⁶ Hij heeft toen onderkend dat in een goed gesprek deelvaardigheden zoals vragen stellen en samenvatten zijn te onderscheiden. Door deze vaardigheden precies te omschrijven kon training preciezer en doelgerichter verlopen en werd ook beoordeling mogelijk. In plaats van de vage observatie van 'het was wel een prettig gesprek' kan nu preciezer worden gezegd welke gespreksvaardigheden goed zijn toegepast en welke niet.

In bijna alle medische opleidingen zijn trainingen in gespreksvaardigheden nu een vast curriculumonderdeel. Naast een training in basisvaardigheden, waarin aandacht wordt geschonken aan zaken zoals non-verbaal gedrag en vragen stellen, zijn er ook trainingen voor gevorderden in bijvoorbeeld anamnesevaardigheden, consultvaardigheden of psychiatrische interviewtechnieken. Om studenten in concreto te laten oefenen, wordt vaak voor een rollenspelvorm gekozen, waarin medestuden-

ten of simulatiepatiënten de rol van patiënt op zich nemen. Meta-analyses tonen aan dat gesprekst rainingen over het algemeen effectief zijn.^{7 8} Dit geldt met name voor duidelijk gestructureerde trainingen waarin wordt gewerkt met het nu gangbare onderscheid in gespreksvaardigheden. Gespreksvaardigheden zijn dus te leren. In hoeverre is het ook mogelijk om op verantwoorde wijze de beheersing van gespreksvaardigheden te toetsen?

De aandacht voor toetsing van gespreksvaardigheden heeft de afgelopen jaren een enorme vlucht genomen in Nederland en daarbuiten, en dan vooral in de medische opleidingen.⁹⁻¹³ Op veel beperktere schaal worden gespreksvaardigheden bij de opleiding psychologie getoetst.^{14 15} In Nederland wordt verder in het Middelbaar Dienstverlenings- en Gezondheidszorg Onderwijs en in het Hoger Economisch en Administratief Onderwijs op het niveau van gespreksvaardigheden getoetst.¹⁶

Methodes voor toetsing van gespreksvaardigheden

Bij de bespreking van toetsvormen voor gespreksvaardigheden is het van belang twee nuances in het achterhoofd te houden. Ten eerste is de scheidslijn tussen gespreksvaardigheden en vakinhoudelijke kennis in de praktijk niet altijd even scherp te trekken. De vragen die een arts bijvoorbeeld stelt om een diagnose te kunnen stellen, komen voort uit een medisch referentiekader. Medische kennis en gespreksvaardigheden zijn dan met elkaar verweven: in hoeverre een medisch student vlot vragen kan stellen, heeft in ieder geval deels te maken met de mate waarin zijn medische kennis toereikend is. Bij toetsing van gespreksvaardigheden ligt echter de nadruk op de gesprekstechnische kant. Bij vragen stellen gaat het er dan om of de vragen duidelijk zijn en of de patiënt wordt gestimuleerd in eigen woorden over zijn of haar klachten te praten.

Ten tweede heeft de benadering die bij

toetsconstructie wordt gebruikt, de taak- of de constructbenadering, consequenties voor de vorm van de toets.¹⁷ Bij de taakbenadering vormen de taken die een student moet kunnen uitvoeren het uitgangspunt. In de toets krijgt de student dan de opdracht om een gesprek met een patiënt te voeren. Voor de beoordeling van de kwaliteit van de uitvoering worden criteria opgesteld. Bij de constructbenadering wordt eerst het domein, het geheel aan kennis, deelvaardigheden en strategieën, gespecificeerd dat een voorwaarde vormt voor het voeren van een gesprek. De toets en de beoordelingscriteria sluiten daarbij aan. Een toets uitgaande van de constructbenadering bestaat vaak uit beperktere opdrachten dan een toets uitgaande van de taakbenadering. Een voorbeeld van zo'n opdracht is studenten te laten kijken naar een op video opgenomen lastige gesprekssituatie en ze te vragen hoe zij de situatie zouden aanpakken. Beide benaderingen kennen hun sterke punten en valkuilen, maar de ene benadering is niet bij voorbaat beter dan de andere.

Welke toetsvormen zijn er nu gebruikt om gespreksvaardigheden te beoordelen? De oudste methode, die wereldwijd wordt gebruikt, is dat een clinicus een arts in spe beoordeelt naar aanleiding van een gesprek met een patiënt. Deze methode wordt vaak toegepast als afsluitend examen bij de co-assistentenschappen, waarbij het oordeel dan zowel de toepassing van medische kennis als gespreksvaardigheden betreft. Het oordeel wordt meestal geveld door één beoordelaar op basis van één enkele observatie, waarbij deze beoordelaar zelf probeert te corrigeren voor de moeilijkheidsgraad van het gesprek. Als de beoordelaar de student vaker aan het werk heeft gezien, wordt soms ook een algemene indruk bij de beoordeling meegenomen. Het voordeel van deze beoordelingsmethode is dat de student beoordeeld wordt in een werksituatie waarvoor zij wordt opgeleid. Er kleven echter verschillende nadelen aan deze methode. Ten eerste moeten er op het juiste moment een patiënt en een beoorde-

laar aanwezig zijn. Ten tweede is de methode belastend voor de patiënt, die een extra gesprek moet ondergaan. Ten derde kan de toetssituatie niet worden gestandaardiseerd, zodat elke student een andere toets krijgt qua inhoud en moeilijkheidsgraad. Ook kan de student moeilijk verhaal halen als deze het niet eens is met het cijfer. Vanwege de nadelen verbonden aan deze wijze van beoordelen, is in het onderwijs gezocht naar alternatieven.

Op dit moment wordt de gesprekssimulatie het meest gebruikt. Deze methode komt voort uit de taakbenadering. In een simulatie neemt een getrainde acteur de rol van patiënt op zich en moet een student in de rol van arts het gesprek leiden. Een observator beoordeelt de prestaties van de student aan de hand van een lijst met beoordelingscriteria. Het voordeel van deze toetsvorm is dat het - tot op zekere hoogte - mogelijk is de gesprekssituatie te standaardiseren. Een praktisch nadeel is dat het toepassen van deze toets veel tijd en menskracht kost.

In de toetsvormen die uitgaan van de constructbenadering wordt vaak gebruik gemaakt van apparatuur zoals audiovisuele media, computers, of cd-i- of cd-rom-apparatuur. Alhoewel dit soort apparatuur als hulpmiddel in het onderwijs al ingeburgerd begint te raken, wordt het nog niet veel gebruikt voor toetsingsdoeleinden. Technieken met beeldmateriaal zijn het meest geschikt, omdat daarmee informatie wordt overgebracht over gezichtsuitdrukkingen, zithouding en intonatie. Dit zijn uiteraard belangrijke elementen van communicatie. De videotoets zoals die bij psychologie in Groningen wordt gebruikt, is hier illustratief.¹⁴ Studenten krijgen hierin korte videovignetten te zien naar aanleiding waarvan ze gespreksvaardigheden moeten toepassen, zoals het geven van een samenvatting. Het voordeel van dit soort toetsen is dat binnen een redelijk tijdsbestek verschillende gesprekssituaties aan de student zijn voor te leggen. Een nadeel is dat het non-verbale gedrag van de student zelf

niet is te beoordelen. Ook kan niet worden nagegaan hoe een student een geheel gesprek leidt en structureert.

Tot slot kunnen gesprekssituaties ook schriftelijk aan studenten worden voorgelegd, of per computer worden aangeboden. Dit is een efficiënte toetsmethode die dezelfde nadelen kent als de methodes met audiovisuele apparatuur. Een bijkomend nadeel is dat niet valt na te gaan of een student non-verbale signalen kan opvangen en interpreteren.

Psychometrische gegevens

Er is veel onderzoek verricht naar de betrouwbaarheid en validiteit van methodes voor toetsing van gespreksvaardigheden. Met betrouwbaarheid wordt hier de nauwkeurigheid bedoeld waarmee de toets meet. Validiteit wil zeggen dat de toets een adequate representant moet zijn van het te meten begrip, oftewel meet de toets wel echt gespreksvaardigheden en niet wat anders? We geven een overzicht van de belangrijkste resultaten.

Oordeel begeleider

Wat betreft de beoordelingsmethode waarin een arts of stagebegeleider een student in actie beoordeelt, leert de psychometrie ons dat één beoordelaar te weinig betrouwbaar is. Hofstee heeft overtuigend beargumenteerd dat een panel van beoordelaars in zo'n beoordelingssituatie nodig is voor een betrouwbare en eerlijke beoordeling.¹⁸ Hij komt op basis van eigen onderzoek tot de schatting dat twee onafhankelijke beoordelaars gemiddeld voor zo'n 15% overeenstemmen en voor 85% van mening verschillen. Ook vanwege de eerder genoemde standaardisatieproblemen is deze methode moeilijk te verdedigen als summatieve beoordelingsvorm waaraan zak- en slaagconsequenties voor de student verbonden zijn. De methode is echter wel voor feedback- en opleidingsdoeleinden bruikbaar.

Tabel 1. Interbeoordelaarsbetrouwbaarheid (*r* of ICC)

dataset	interbeoordelaars- overeenstemming 2 beoordelaars
CFPC *	.63 .60
UMass data set 1 *	.77
Bögels	.64** .51** .35-.39 (M=.36)
Kraan & Crijnen	.44 (basic interviewing skills) .32 (structuring the interview)
Pieters	.24-.77 (M=.46)
Smit	.81-.87 (M=.84)
	.85-.93 (M=.88)
Vermeulen	.69-.88 (M=.83)

* Uit: Van der Vleuten & Swanson¹⁹

** Met Spearman-Brown formule berekend

Simulaties

Voor een overzicht van de kwaliteit van toetsing met behulp van simulatiepatiënten, in het Engels vaak 'Standardized Patients' genoemd, is gebruik gemaakt van een 'state of the art'-artikel en enkele proefschriften. Het betreft:

- Geestelijke gezondheidskunde: Bögels;⁹
- Huisartsgeneeskunde, eerstelijnszorg geestelijke gezondheidsproblemen: Kraan & Crijnen;¹⁰
- Huisartsgeneeskunde: Pieters;¹¹
- Psychologie: Smit;¹⁴
- Middelbaar Dienstverlenings- en Gezondheidszorgonderwijs: Vermeulen;¹⁶
- Geneeskunde: Van der Vleuten & Swanson.¹⁹

Ook enkele studies buiten het medisch vakgebied zijn geraadpleegd voor dit artikel.

Ten aanzien van de betrouwbaarheid is in de eerste plaats de mate van overeenstemming tussen beoordelaars van belang. Aangezien in de meeste studies een correlatiecoëfficiënt of een intraclass-coëfficiënt wordt gerapporteerd, zijn deze waarden in tabel 1 opgenomen om zo een indruk te krijgen van de gemiddelde

in- terbeoordelaarsovereenstemming.²⁰ Tabel 1 geeft dus een indicatie van de range waarin de overeenstemmingsmaten liggen. In sommige gevallen worden in één publicatie meer zelfstandige deelstudies gerapporteerd; vandaar dat er dan meerdere overeenstemmingsgegevens in de tabel worden vermeld. Het overzicht in tabel 1 laat zien dat de overeenstemming tussen twee getrainde beoordelaars uiteenloopt van .24 tot .93. In verschillende studies wordt voldoende overeenstemming gevonden (waarden boven de .80) maar er zijn ook uitschieters naar beneden. Deze gegevens duiden erop dat een goede training van de beoordelaars noodzakelijk is. De gegevens tonen ook aan dat bevredigende waarden zijn te bereiken.

Behalve door goede training kan de beoordelaarsbetrouwbaarheid ook worden bevorderd door meer beoordelaars in te zetten. Hier geldt het principe 'twee zien meer dan één'. Er zijn dan twee opties: de prestaties van een student in één gesprekssimulatie worden door meer dan één beoordelaar beoordeeld, of een student neemt deel aan verschillende gespreks-simulaties en wordt daarbij steeds door een

Tabel 2. Aantal casus benodigd voor een generaliseerbaarheidscoëfficiënt van minimaal .80

dataset	aantal simulaties	tijdsduur simulatie
CFPC*	>10	15 min.
UMass*	8/9	15 min.
Bögels	8	30 min.
	10	
	13	5 min.
	10	10 min.
	>25	20 min.
Kraan & Crijnen	>40	15 min.
	>20	
Smit	>20	20 min.
	>30	25 min.
Vermeulen	5	20 min.

* Uit: Van der Vleuten & Swanson¹⁹

andere beoordelaar beoordeeld. Deze laatste variant wordt in het medisch onderwijs het meest toegepast.

Een andere vraag in het kader van de betrouwbaarheid is in hoeverre een score op een simulatie iets zegt over hoe goed een student in andere gesprekssimulaties zal scoren. Dit is de vraag naar de generaliseerbaarheid. In bijna alle studies is nagegaan hoeveel simulaties en soms ook beoordelaars nodig zijn voor een score met een generaliseerbaarheid van .80. In het algemeen geldt: hoe meer simulaties en hoe meer beoordelaars hoe betrouwbaarder de uitspraken zijn over de beheersing van gespreksvaardigheden. In tabel 2 is te zien hoeveel simulaties er nodig zijn voor een generaliseerbaarheidscoëfficiënt van .80, uitgaande van één beoordelaar. De waarden gaan uit van het domeingeoriënteerd perspectief, waarbij niet alleen verstoring in rangordening van studenten maar ook in algemeen niveau worden verdisconteerd.²¹ Indien deze domeingeoriënteerde waarden niet werden gerapporteerd, zijn ze voor tabel 2 berekend.

Uit tabel 2 blijkt dat minimaal vijf, maar meestal zo'n tien, simulaties nodig zijn voor

een generaliseerbare score. Wat uit de literatuur verder blijkt, is dat het toevoegen van beoordelaars veelal ook leidt tot een verhoging van de generaliseerbaarheid, maar bij goed getrainde beoordelaars heeft het toevoegen van simulaties veel meer effect.

Tevens geeft het onderzoek van Bögels aanwijzingen dat het loont om gesprekken korter te maken. Twaalf gesprekken van vijf minuten leveren een veel hogere generaliseerbaarheid op dan drie gesprekken van twintig minuten, terwijl het in beide gevallen gaat om een uur aan testtijd. Een natuurlijke ondergrens is de tijd die is benodigd om een bepaalde gespreksituatie of een gedeelte daaruit realistisch te simuleren.

Naast de betrouwbaarheid is - in mindere mate - ook de begripsvaliditeit van gespreks-simulaties onderzocht: meet de toetsvorm echt gespreksvaardigheden? Eén manier om daar zicht op te krijgen is door na te gaan of studenten na een training hoger op de simulatie scoren dan ervoor, en door na te gaan of de toets kan differentiëren tussen groepen die naar verwachting in niveau verschillen.

Bögels en Smit tonen aan dat de door hen geconstrueerde simulaties in staat zijn een toename in vaardigheidsniveau als gevolg van training te registreren. In hun studies, de studies van the University of Massachusetts en die van Vermeulen bleek dat de simulaties differentiëren tussen studenten die in het begin en die verder in hun studie zijn. Overeenkomstig de verwachting bleek dat studenten die verder in hun studie waren gemiddeld hoger scoorden dan studenten die minder ver waren. Professionals blijken ongeveer even goed of beter te scoren dan studenten die zich in de laatste fase van hun studie bevinden.

Deze resultaten zeggen echter niets over het type vaardigheden dat met een simulatie wordt gedekt. De studies waarin correlaties met andere studiematen zijn berekend, zoals scores op een tentamen over bepaalde (medische) kennis, zijn wat dat betreft informatiever. Dit

Tabel 3. Correlaties van scores op simulaties met scores op andere beoordelingen van gespreksvaardigheden (vetgedrukt) en (studie)maten

dataset	toetsen	correlaties
Umass data set 2	-mc-test (NBME deel 1)	.19
	-mc-test (NBME deel 2)	.27
	-klinische beoordelingen	.44
	-follow-up met mc-en korte-antwoordvragen	.26
Umass data set 3	-mc-test (NBME deel 1)	.10
	-mc-test (NBME deel 2)	.22
	-klinische beoordelingen	.25.
Kraan & Crijnen, interpersoonlijke vaardigheden en communicatie	-globaal expert oordeel	MAAS-GP/MAAS-PMHC .53/ .29 .10/-.15
	-medische kennis	-.24
Smit		basistraining/training gevorderden
	-videotoets gespreksvaardigheden	.14 / .08
	-schriftelijke toets gespreksvaardigheden	.06 / .20
	-gemiddeld oordeel 2 trainers	.29 / .17
Vermeulen	-stagebeoordeling	.52 (1e jaars), .26 (3e jaars)
	-examencijfer regels, voorzieningen en procedures	-.02
	-examencijfer rapportage	-.09
	-examencijfer sociale vaardigheden	-.01
	-examencijfer organisatie en beleid	-.08

geldt met name als in de studie andersoortige gesprekstoeetsen en toetsen die een beroep doen op kennis en andersoortige vaardigheden zijn betrokken. De correlaties tussen toetsen die betrekking hebben op gespreksvaardigheden zouden hoger moeten zijn dan de correlaties tussen gesprekstoeetsen met studiematen (tabel 3).

De gemiddelde correlatie tussen scores op simulaties met scores op gespreksmaten, zoals beoordelingen door begeleiders of toetsen met audiovisuele media, is .23. De correlaties tussen scores op simulaties met scores op studiematen zijn gemiddeld lager, namelijk .06. Deze bevinding geeft steun aan de verwachting dat met simulaties gespreksvaardigheden worden gemeten.

Een andere vraag is in hoeverre scores op simulaties kunnen voorspellen hoe adequaat een student met een patiënt of cliënt in werkelijkheid communiceert. Dit is de vraag naar de predictieve validiteit. De correlaties die Vermeulen noemt ten aanzien van stagebeoordelingen zijn hier van belang.¹⁶ Deze correlaties zijn .52 voor eerstejaars- en .26 voor derdejaarsstudenten. Deze resultaten vormen een aanwijzing dat uit gedrag in een simulatie te voorspellen is hoe goed een student in werkelijkheid met een patiënt communiceert.

Samenvattend kunnen we stellen dat de overeenstemming tussen beoordelaars geen problemen hoeft op te leveren, mits de beoordelaars goed op hun taak zijn voorbereid. De generaliseerbaarheid van de scores is wel pro-

Tabel 4. Interbeoordelaarsbetrouwbaarheden (correlaties) voor toetsen uitgaande van de constructbenadering

dataset	soort toets	r
Sharf & Lucas ²³	gecomputeriseerde simulatie counselingsvaardigheden	80.5% overeenstemming*
Smit ¹⁴	videotoets	m=.93 m=.90
	mc-toets	1
	toets met essayvragen	m=.91
Stricker ²⁴	videotoets	m= .84 (effectiviteit)

*Fleiss's K^m statistic

blematisch. Er zijn rond de tien simulaties nodig om met enige zekerheid uitspraken te kunnen doen hoe goed een student de gespreksvaardigheden beheerst. Tot slot geven de resultaten ten aanzien van de validiteit aanwijzingen dat met een simulatie werkelijk gespreksvaardigheden worden beoordeeld en bijvoorbeeld niet alleen medische kennis.

Toetsen uitgaande van de constructbenadering

De onderzoeksresultaten gevonden met toetsvormen waarin gebruik is gemaakt van de computer, multimedia-applicaties of schriftelijk materiaal worden samengenomen omdat het in alle gevallen om dezelfde soort vragen en opdrachten gaat die voortkomen uit de constructbenadering.²² In tabel 4 worden de interbeoordelaarsbetrouwbaarheden op een rij gezet van studies waarin dit soort toetsen zijn gebruikt voor beoordelingsdoeleinden.

Uit de tabel blijkt dat de interbeoordelaarsovereenstemming hoog is, wat erop duidt dat met één goed ingewerkte beoordelaar kan worden volstaan. Ten aanzien van de generaliseerbaarheid blijkt uit het overzichtsartikel van Swanson en anderen dat de correlaties tussen scores op verschillende opdrachten die gespreksvaardigheden betreffen gemiddeld

.40 is.²² Er valt dan te berekenen dat zo'n acht opdrachten voldoende zijn om generaliseerbare scores te verkrijgen. Aangezien de opdrachten vaak niet langer dan vijf tot tien minuten duren, zullen de toetsen meestal meer dan acht opdrachten bevatten en dus aan deze eis voldoen.

Ten aanzien van de validiteit zijn de volgende resultaten gevonden.^{14 22-25} In de eerste plaats is nagegaan of deze toetsen kunnen differentiëren tussen personen die naar verwachting in beheersingsniveau verschillen. Het blijkt inderdaad dat personen met meer relevante ervaring hoger op de toetsen scoren dan personen met minder ervaring.

In tabel 5 worden de correlaties met andere maten vermeld. De verwachting is weer dat de toetsscores meer samenhang zullen vertonen met beoordelingen van gespreksvaardigheden dan met cognitieve studiematen.

De correlaties tussen de toetsen met maten die betrekking hebben op gespreksvaardigheden zijn matig, gemiddeld .14, en in het algemeen hoger dan de correlaties met cognitieve studiematen. Hierbij moet worden aangetekend dat er maar twee correlaties werden gerapporteerd waarin cognitieve maten waren betrokken.

Ten aanzien van de predictieve validiteit worden correlaties tussen computersimulaties

Tabel 5. Correlaties van scores op gesprekstoetsen met scores op andere beoordelingen gespreksvaardigheden (vetgedrukt) en cognitieve (studie)maten

dataset	maten	geobserveerde correlaties
Smit ¹³	docentenoordeel van 2 docenten over prestaties in de training	videotoets basistraining/ training gevorderden .24 / .23
		schriftelijke toets basistraining/ training gevorderden .15 / .23
Stricker ^{26 33}	wiskundetoets	.01
	vocabulairetest	-.01
	peer ratings sociale vaardigheden	.15*
	test herkennen non-verbaal gedrag	-.06 .07

* gecorrigeerd voor onbetrouwbaarheid

Tabel 6. Overzicht resultaten

	overeenstemming beoordelaars	generaliseerbaarheid	begripsvaliditeit	predictieve validiteit
simulaties met acteurs	+	-	+	+
toetsen uitgaande van de constructbenadering	++	+	±	+

en praktijkbeoordelingen gerapporteerd tussen .30 en .40.²⁶ Een kanttekening bij dit resultaat is dat de computersimulaties primair zijn gemaakt voor de beoordeling van de wijze waarop een student medische vraagstukken oplost.

Op basis van de resultaten kunnen we concluderen dat bij toetsen uitgaande van de constructbenadering de overeenstemming tussen beoordelaars hoog is. Ook de generaliseerbaarheid van de scores levert bij toetsen met minimaal acht opdrachten geen problemen op. Er wordt enige steun gevonden voor de validiteit, maar deze steun is minder overtuigend dan bij de simulaties.

Conclusies

In tabel 6 zijn de resultaten in één overzicht samengevat. Op basis van dit overzicht van

onderzoeksbevindingen is de vraag te beantwoorden hoe een verantwoorde toetsingsprocedure ten aanzien van gespreksvaardigheden er uit zou kunnen zien. Daarbij maken we een onderscheid tussen trainingen in gespreksvoering voor beginners en trainingen voor gevorderden. De les die de bekende onderwijspsycholoog De Groot ons leert, is dat een toets moet aansluiten bij de trainingsdoelen.²⁷ Dit impliceert dat voor een eerste training met als doel inzicht in de beginselen van gespreksvoering vragen naar aanleiding van video-opnamen of vragen naar aanleiding van schriftelijke of per computer aangeboden casus te verdedigen zijn. Bij een gevorderde training waarin de doelen beheersing van een gesprekstype betreffen, ligt het voor de hand om ook simulaties te gebruiken. Een onderwijskundig argument voor gebruik van de simulatie is de motiveren-

de en richtinggevende invloed op het studiegedrag van studenten. Met name deze toetsvorm blijkt in staat om studenten te stimuleren om de gespreksvaardigheden te oefenen.¹⁴

Een praktisch probleem is echter dat het gebruik van dit soort simulaties een aanzienlijke tijdsinvestering vraagt van de staf en een groot beroep doet op middelen en ruimtes. Een complicerende factor daarbij is dat minimaal tien simulaties nodig zijn voor een generaliseerbare score. Een mogelijkheid om de toetsingsprocedure te optimaliseren en toch praktisch haalbaar te houden, is door één of twee kortere simulaties aan te vullen met toetsen die vanuit de constructbenadering zijn geconstrueerd. Een *combinatie van toetsmethoden* wordt ook elders voorgesteld.¹² Het voordeel hiervan is dat een eindoordeel gebaseerd op meer toetsvormen betrouwbaarder is dan een oordeel op basis van één toetsvorm.²⁸ Daarnaast wordt het generaliseerbaarheidsprobleem tegengegaan door bijvoorbeeld een videotoets af te nemen, aangezien daarin een aantal *verschillende* gesprekssituaties aan een student kan worden voorgelegd. Deze procedure is aanzienlijk efficiënter dan het afnemen van verscheidene rollenspeltoetsen. Tevens is zo een goede toetsingsprocedure beter haalbaar. Goede toetsing is belangrijk: daardoor wordt namelijk een bijdrage geleverd aan een behartigenswaardig doel, te weten studenten afleveren die in staat zijn een professioneel gesprek te voeren.

Literatuur

1. L'Abate L, Milan MA. Handbook of social skills training and research. New York [etc.]: John Wiley & Sons, 1985.
2. Balint M. The doctor, his patient and the illness. London: Pitman Medical, 1956.
3. Schouten JAM. Anamnese en advies. Alphen aan den Rijn: Samsom, 1985.
4. Wouda J, Wiel H van der, Vliet K van. Medische communicatie: gespreksvaardigheden voor de arts. Utrecht: Lemma, 1996.
5. Dalen J van. Attitude-ontwikkeling en sociale vaardigheden: een kwestie van complementariteit? In: Dochy FJ, Luyk SJ van, redactie. Handboek voor vaardigheds-onderwijs. Lisse: Swets & Zeitlinger, 1987:105-13.
6. Ivey AE. Microcounseling. Springfield, IL: Charles C. Thomas, 1971.
7. Molen HT van der, Smit GN, Hommes MA, Lang G. Two decades of cumulative microtraining in the Netherlands: an overview. Educational Research and Evaluation 1995;14:347-78.
8. Baker SB, Daniels TG, Greeley AT. Systematic training of graduate-level counselors: narrative and meta-analytic reviews of three major programs. Counseling Psychologist 1990;18:355-421.
9. Bögels SM. Teaching and assessing diagnostic interviewing skills. An application to the mental health field [dissertatie]. Maastricht: Rijksuniversiteit Limburg, 1994.
10. Kraan HF, Crijnen A. The Maastricht history taking and advice checklist: studies of instrumental validity [dissertatie]. Amsterdam: Lundbeck, 1987.
11. Pieters HM. De Utrechtse consult evaluatie methode. Vaardigheden in consultvoering van huisartsen in opleiding getoetst [dissertatie]. Lelystad: Meditekst, 1991.
12. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. Educational Researcher 1995;24(5):5-11.
13. Hart IR, Harden RM, eds. Further developments in assessing clinical competence. Proceedings of the second Ottawa Conference. Montreal: Can-Heal Publications, Inc, 1987.
14. Smit GN. De beoordeling van professionele gespreksvaardigheden. Constructie en evaluatie van rollenspel-, video- en schriftelijke toetsen [dissertatie]. Baarn: Nelissen, 1995.
15. Gallagher MS, Hargie ODW. An investigation into the validity of role play as a procedure for counsellor skill assessment. British Journal of Guidance and Counselling 1989;17:155-65.
16. Vermeulen W. Toetsing van communicatieve vaardigheden. Constructie en evaluatie van gedragstoetsen voor professionele gespreksvaardigheden [dissertatie]. Arnhem: Cito, 1993.
17. Messick S. The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher 1994;23(2):13-23.
18. Hofstee WKB. Beoordeling: wetenschap of kunst. Voordracht gehouden in de Verenigde Vergadering van de Afdelingen van de Koninklijke Nederlandse Akademie van Wetenschappen, 1995.

19. Vleuten CPM van der, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teaching and Learning in Medicine* 1990;2(2):58-76.
20. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979;86:420-8.
21. Vleuten CPM van der, Wijnen WHFW. Niets praktischer dan een goede theorie: generaliseerbaarheidstheorie als instrument voor betrouwbaarheidsstudies. *Bulletin Medisch Onderwijs* 1991;10:2-14.
22. Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987;12(3):220-46.
23. Sharf RS, Lucas M. An assessment of a computerized simulation of counseling skills. *Counselor Education and Supervision* 1993;32:254-66.
24. Stricker LJ. Interpersonal competence instrument: development and primary findings. *Applied Psychological Measurement* 1982;6(1):69-91.
25. Stricker LJ, Rock DA. Interpersonal competence, social intelligence, and general ability. *Personality and Individual Differences* 1989;11(8):833-9.
26. Norcini JJ, Meskauskas JA, Langdon LO, Webster GD. An evaluation of a computer simulation in the assessment of physician competence. *Evaluation of Health Professionals* 1986;9:286-304.
27. Groot AD de. 'Dekkingsproblemen' bij de constructie van examenprogramma's en curricula. In: *Handboek voor de onderwijspraktijk*: afl. 14. Deventer: Van Loghum Slaterus, 1982:2.1.Gro.1. -2.1.Gro.25.
28. Starren J, Bakker SJ, Wissel A van der. *Inleiding in de onderwijspsychologie*. Muiderberg: Dick Coutinho, 1988.

DE AUTEUR

Dr. G.N. Smit was ten tijde van deze studie als docent/onderzoeker verbonden aan de afdeling Persoonlijkeits- en Onderwijspsychologie van de Rijksuniversiteit Groningen. Momenteel werkt ze als adviseur Bedrijfspsychologie bij GITP.

Correspondentieadres:

G.N. Smit, GITP Bedrijfspsychologie, Berg en Dalseweg 127, 6522 BE Nijmegen.

Deze bijdrage is een bewerking van de plenaire voordracht die gehouden is tijdens het Gezond Onderwijs Congres 1996.