

## Assistententoets kindergeneeskunde: een beschrijving van de psychometrische eigenschappen

L.W.T. Schuwirth, J.P.P. Schrande, C.P.M. van der Vleuten

### Inleiding

Sinds 1989 vindt op initiatief van de Nederlandse Vereniging voor Kindergeneeskunde jaarlijks een landelijke experimentele assistententoets kindergeneeskunde plaats. De assistenten kindergeneeskunde in Nederland hebben naast hun klinische opleiding de mogelijkheid themabijeenkomsten te bezoeken, waarvan het doel is de theoretische kennis op te frissen en waar nodig te vergroten. Het doel van de experimentele assistententoets was om een instrument te verkrijgen waarmee het effect van de themabijeenkomsten vast te stellen zou zijn. Daarbij werd echter niet gestreefd naar een vervanging van de evaluatievormen die op dit moment gehanteerd worden in de diverse opleidingsplaatsen, maar meer naar een mogelijke aanvulling hierop.

Om de bruikbaarheid van een dergelijke vorm van toetsing te beoordelen, werden, naast de uitvoerbaarheid, het discriminerend vermogen en de relevantie van belang geacht. Daarbij de eerste twee toetsafnames (1989, 1990) gebleken was dat de uitvoerbaarheid alleszins redelijk was, zijn de laatste twee aspecten meer op de voorgrond komen te staan bij de laatste twee toetsafnames (1991, 1992). Bespreking van discriminerend vermogen en relevantiebepaling zal onderwerp zijn van dit artikel.

### Methode

**Toetsvorm:** De toetsvorm die gebruikt is, is afgeleid van de Maastrichtse Voortgangstoets.<sup>1</sup> Hierbij wordt gebruikgemaakt van juist-onjuist-vraagteken-vragen. Dit zijn stellingen waarvan de deelnemer dient aan te geven of deze juist of onjuist zijn. De deelnemer krijgt een punt voor ieder correct antwoord. Voor een

foutief antwoord wordt een punt afgetrokken en invullen van een vraagteken levert geen punten op. De totale score van een deelnemer is dus het aantal vragen goed min het aantal vragen fout (goed-min-fout score).

Om de toetsinhoud zoveel mogelijk aan te laten sluiten bij het domein van de kindergeneeskunde is de toets onderverdeeld in categorieën overeenkomend met de verschillende deelspecialismen binnen de kindergeneeskunde. Voorbeelden van dergelijke categorieën zijn: metabole ziekten, neonatologie, oncologie, neurologie, klinische genetica.

Verschillende deelspecialisten hebben op verzoek vragen gemaakt. Deze vragen zijn gescreend op vorm en inhoud, en waar nodig in overleg met de auteurs aangepast. Uit deze vragen is de toets samengesteld. In 1991 en 1992 bestond de toets uit 175 items. De deelnemers zijn na de toetsafname uitgenodigd kritiek te leveren op de vragen, hetgeen in beide jaren heeft geresulteerd in het laten vervallen van vijf vragen.

**Deelnemers:** In 1991 en 1992 is de toets voorgelegd aan alle assistenten kindergeneeskunde, zowel assistenten-in-opleiding als assistenten-niet-in-opleiding (AGNIO's) van de verschillende opleidingscentra in Nederland. De betrokken assistenten zijn verdeeld in zes jaargroepen (AGNIO's vormen jaargroep 0). Het aantal deelnemende assistenten bedroeg respectievelijk 131 en 145.

**Relevantie:** In 1992 is de toets toegestuurd aan tien algemeen kinderartsen met als doel hen een relevantiebeoordeling te laten geven van iedere vraag op een schaal van 1 tot 5 (1 = zeer irrelevant, 3 = neutraal, 5 = zeer relevant).

*Analyse:* Om een indruk te krijgen van het discriminerend vermogen van de toets zijn gemiddelde jaargroepsscores, standaarddeviaties, standard errors en de betrouwbaarheden berekend. De resultaten zijn gegroepeerd per jaargroep over de opleidingscentra heen, waarbij steeds gebruik is gemaakt van de goed-min-fout scores. Van de relevantieoordelen van de algemeen kinderartsen zijn per item de mediaan berekend. Per mediaan is daarna bepaald bij hoeveel items deze voorkwam.

*Selectie op vraaginhoud:* Bij de toets van 1991 rees het vermoeden dat op inhoudelijke gronden twee verschillende vraagsoorten te onderscheiden waren. De eerste vraagsoort had betrekking op een specifiek vaststaand feit, zoals dat in het algemeen eenvoudig in een bepaalde paragraaf van een leerboek terug te vinden is. De vraag is gericht op herkenning of herinnering van een feit; combinatie van gegevens of afweging ervan werd niet gevraagd. Deze vraagsoort zal verder 'feitenvraag' genoemd worden. Een voorbeeld van een 'feitenvraag' is:

*Het gemiddelde verschil tussen de lengte van mannen en vrouwen in het Nederlandse groei-onderzoek van 1980 was 14 cm. Juist!/Onjuist*

De tweede soort was veeleer gericht op het toepassen van kennis. De vraag richtte zich op de combinatie van verschillende in de stam genoemde feiten en de inschatting van waar-

schijnlijkheden en hun afzonderlijk belang.<sup>2</sup> Deze vraagsoort zal verder 'toepassingsvraag' genoemd worden. Een voorbeeld van een 'toepassingsvraag' is:

*Een Grieks jongetje dat vaak luchtweginfecties heeft en alleen witbrood met jam eet, blijkt een bloedarmoede te hebben. De arts heeft in haar differentiaaldiagnose thalassemie, ijzertekort en recidiverende infecties als oorzaken staan.*

*De gegevens van de ijzerstatus (serumijzer, ijzerbindingscapaciteit en ferritine) en de MCV (mean corpuscular volume) van dit jongetje geven voldoende informatie om van deze drie de juiste diagnose te kiezen. Juist!/Onjuist*

Hierbij is slechts geprobeerd een beschrijving van de vraagstimulus te geven en niet om een uitspraak te doen over de feitelijke mentale processen van de deelnemers.

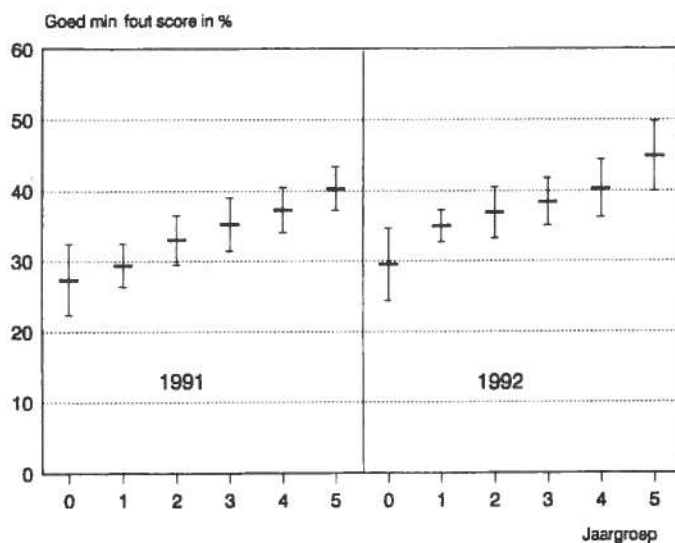
In beide toetsen zijn een cluster van 'feitenvragen' en een cluster van 'toepassingsvragen' gemaakt. In 1991 waren deze clusters ieder twintig vragen groot; in 1992 bevatten beide clusters ieder dertig vragen.

## Resultaten

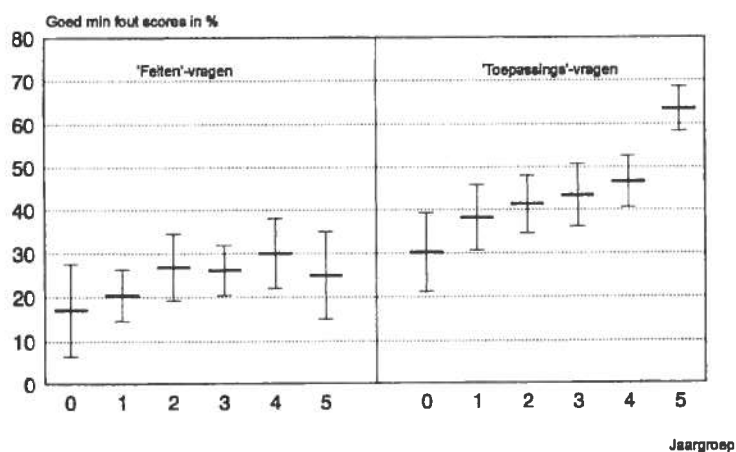
Tabel 1 toont de gemiddelde procentuele goed-min-fout scores en de daarbij behorende standaarddeviaties. Ook zijn de betrouwbaarheden (Cronbach's alfa) van beide toetsen uitgesplitst per jaargroep. Voor beide toetsen geldt dat de gemiddelde goed-min-fout score toeneemt per

**Tabel 1.** Beschrijvende statistiek van de procentuele goed minus fout scores over beide toetsen

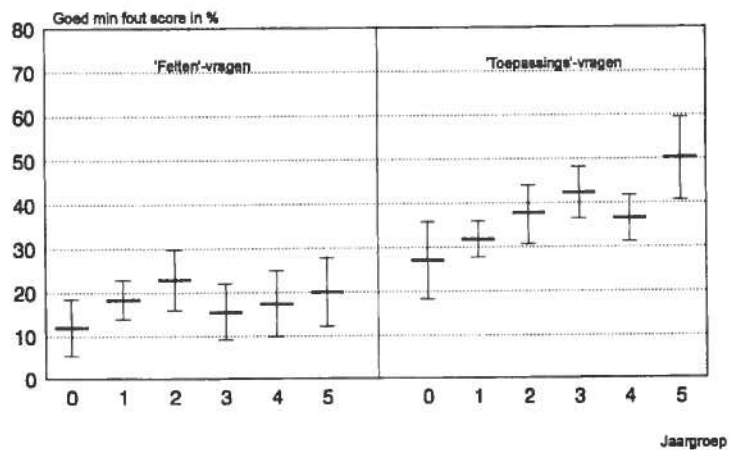
Jaargroep	Toets 1991				Toets 1992			
	N	gem.	sd	alfa	N	gem.	sd	alfa
0	17	27.14	10.6	0.71	20	29.47	11.7	0.76
1	34	29.40	8.9	0.58	38	34.97	7.1	0.31
2	19	33.00	7.9	0.52	26	36.88	9.4	0.64
3	26	35.20	9.9	0.67	30	38.39	9.4	0.62
4	19	37.31	7.0	0.34	19	40.25	9.0	0.59
5	16	40.33	6.7	0.28	12	44.90	8.7	0.58
Totaal	131	33.30	9.6	0.64	145	36.77	9.8	0.65



**Figuur 1.** Gemiddelde toetsscores per jaargroep en 95%-betrouwbaarheidsintervallen van beide toetsen



**Figuur 2.** Gemiddelde toetsscores per jaargroep en 95%-betrouwbaarheidsintervallen van beide clusters in 1991



**Figuur 3.** Gemiddelde toetsscores per jaargroep en 95%-betrouwbaarheidsintervallen van beide clusters in 1992

jaargroep. De alfa's zijn met 0.64 en 0.65 niet onredelijk te noemen, echter gebruikmakend van de Spearman-Brown correctieformule voor toetsverlenging, blijkt dat een verlengingsfactor van 2.18 nodig zou zijn om een betrouwbaarheid van 0.80 te krijgen. Dit komt neer op een vergelijkbare toets van 370 items.

De standaarddeviaties suggereren dat de scoreverschillen tussen de opeenvolgende jaargroepen niet groot genoeg zijn om de verschillen significant te laten zijn. Figuur 1 toont een grafische representatie van de gemiddelde jaargroepsscores met hun 95%-betrouwbaarheidsintervallen.

De grote overlap van de betrouwbaarheidsintervallen tussen opeenvolgende jaargroepen geeft aan dat de verschillen niet statistisch significant zijn. Wel zijn er statistisch significante verschillen in gemiddelde scores tussen begin (jaargroepen 0 en 1) en einde (jaargroepen 4 en 5) van de opleiding.

In figuur 2 zijn de gemiddelde goed-min-fout scores op de clusters van 1991 uitgezet met hun 95%-betrouwbaarheidsintervallen. Uit figuur 2 wordt duidelijk dat er verschillen bestaan tussen de 'feiten'- en de 'toepassingsvragen': de gemiddelde scores op de 'toepassingsvragen' zijn fors hoger dan die van de 'feitenvragen'. Er is een verschil in tendens tussen beide clusters: daar waar de gemiddelde scores op de 'toepassingsvragen' een continu stijgende tendens vertonen, treedt bij de 'feitenvragen' na jaar 2 een stagnatie en zelfs een daling van de scores op. Figuur 3 toont eveneens de gemiddelde scores en 95%-betrouwbaarheidsintervallen, maar nu van de toets van 1992.

Ook bij de toets van 1992 blijkt het mogelijk op grond van de vraaginhoud twee clusters samen te stellen, waarbij dezelfde verschillen gezien worden als bij de toets van 1991. De 'toepassingsvragen' discrimineren beter dan de 'feitenvragen'. Correlaties tussen de 'feitenvragen' en de 'toepassingsvragen' blijken in beide afnamejaren laag te zijn. De geobserveerde correlaties bedragen .18 in 1991 en .19

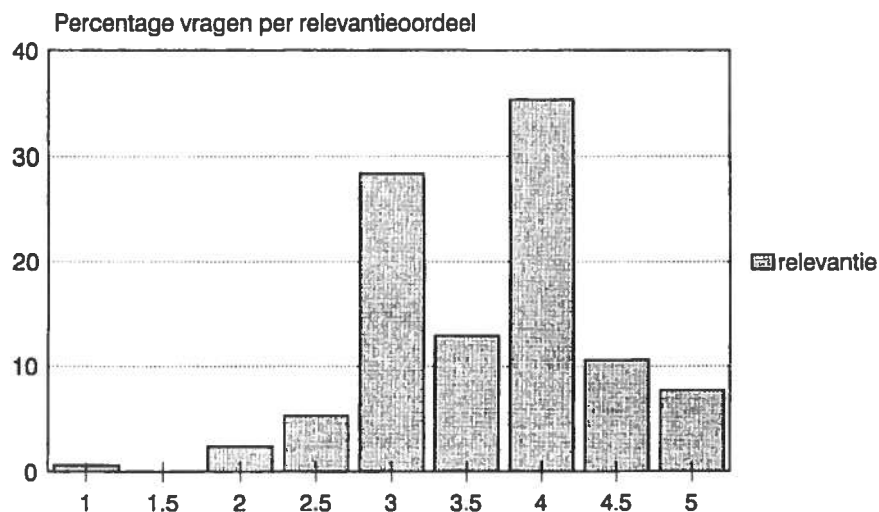
in 1992. Beide correlaties zijn statistisch niet significant. Correctie voor attenuatie, waarbij een schatting gemaakt wordt van de correlaties indien beide clusters een ideale betrouwbaarheid zouden hebben, levert hogere waarden op .78 en .91. De verschillen tussen de geobserveerde en de ware correlaties zijn groot. Dit feit in combinatie met de lage betrouwbaarheden, maakt dat weinig te zeggen valt over de sterkte van het verband tussen beide clusters.

Figuur 4 toont de percentages van de items met hun medianen. De mediane relevantieoordelen van de items zijn over het algemeen laag. Indien we aanhouden dat slechts 46% van de items een relevantiebeoordeling krijgt die overeenkomt met 'relevant' en slechts 8% een beoordeling 'zeer relevant', dan is dit samen iets meer dan de helft, namelijk 54%. Een relevantieoordeel overeenkomend met 'neutraal' wordt aan 41% van de items gegeven, terwijl de rest (8%) als oordeel 'irrelevant' tot 'zeer irrelevant' krijgt.

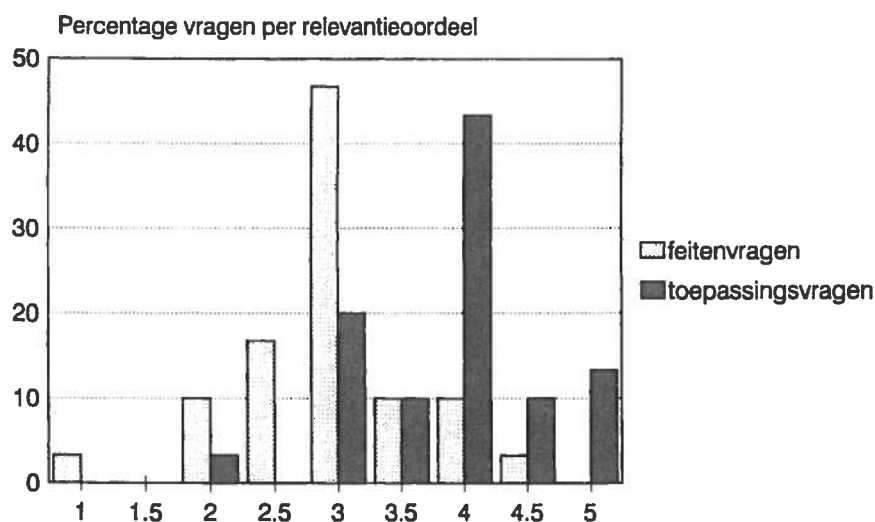
Eenzelfde berekening is uitgevoerd voor de vragen in beide clusters. De resultaten hiervan worden getoond in figuur 5. Ook hier zijn de aantallen omgerekend in percentages. Dit is gedaan om de vergelijkbaarheid tussen figuur 4 en 5 groter te maken. Uit figuur 5 blijkt dat de 'toepassingsvragen' over het algemeen een beduidend hoger relevantieoordeel krijgen dan de 'feitenvragen'.

## Discussie

Het discriminerend vermogen van de landelijke experimentele assistententoets kindergeneeskunde in zijn huidige vorm is voor verbetering vatbaar. De stijging die gezien wordt over de jaargroepen heen is te klein, de gevonden verschillen zijn niet significant. De betrouwbaarheden binnen de jaargroepen zijn matig. Indien de betrouwbaarheden van de toetsen bepaald worden voor de gehele deelnemersgroep, worden iets hogere waarden gevonden. Echter ook over de gehele deelnemersgroep zou een toetsverlenging tot meer dan



**Figuur 4.** Percentage items met hun bijbehorende mediane relevantieoordelen over alle items (n=170)



**Figuur 5.** Percentage items van de toets in 1992 met hun bijbehorende mediane relevantieoordelen per cluster van vraagsoorten (n=30)

twee maal het huidige aantal items nodig zijn om betrouwbaarheden van 0.80 of hoger te bereiken. Verlenging van de toets is echter moeilijk haalbaar, daar het vereiste aantal vragen (ongeveer 370) een lange toetstijd (naar schatting vijf uur) vereist. De huidige mogelijkheden laten het ook niet toe om meer dan één toetsmoment per jaar te organiseren.

De mate van discriminerend vermogen lijkt samen te hangen met de vraaghoud. Er wordt consistent in beide toetsen een uitgesprokener

stijging van gemiddelde jaargroepsscores gezien op de 'toepassingsvragen' dan op de 'feitenvragen'. Geobserveerde correlaties tussen beide vraagsoorten blijken ook bij herhaling laag te zijn, hetgeen een indicatie zou kunnen zijn voor de idee dat beide soorten zich richten op verschillende kennisaspecten. Het spreekt echter voor zich dat het aantal items in beide clusters (ieder twintig in 1991 en ieder dertig in 1992) te laag is om op grond hiervan verdere conclusies te trekken. Nader onderzoek hier-

naar is zeker gewenst. Niettemin zou mogelijk een beter discriminerend vermogen bereikt kunnen worden door minder feitenvragen en meer toepassingsvragen in de toets op te nemen.

De relevantieoordelen van de algemeen kinderartsen zijn over het algemeen laag. Mogelijk is hieraan de gevolgde procedure bij de itemproductie debet: de som van de grootste deelspecialismen binnen het terrein van de kindergeneeskunde is waarschijnlijk niet in overeenstemming met de doelstellingen van de opleiding tot algemeen kinderarts. Met name de 'feitenvragen' worden in het algemeen als niet relevant beoordeeld. Ook hier dient echter opgemerkt te worden dat de data summier zijn, en hardere conclusies pas gerechtvaardigd zijn na verder onderzoek.

Op grond van bovenstaande is voorlopig besloten de toets samen te laten stellen door algemeen kinderartsen. Hun is gevraagd de items op een zodanige wijze samen te stellen dat deze voldoen aan de operationalisering voor toepassingsvragen die in dit artikel beschreven zijn.

In het komende jaar zal de toetsing geen consequenties voor de deelnemers hebben, hoewel het de bedoeling is dat dit in het jaar daarop wel het geval zal zijn. Zak-slaagbeslissingen zullen echter niet genomen worden. Het gewicht van de bestaande beoordelingen in de diverse opleidingsplaatsen zal niet aangetast

worden. Indien consequenties aan deze toetsvorm verbonden zullen worden, zullen zij niet verder reiken dan dringende nascholingsadviezen bij herhaaldelijk gebleken lacunes.

## Literatuur

1. Verwijnen GM, et al. The evaluation system at the medical school of Maastricht. *Assessment and Evaluation in Higher Education* 1982; 7: 235-44.
2. Van Leeuwen YD, Van Hessen PAW. Clinical competence and objective questions: tic-tacs to realize a true/false format assessing competence. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP (eds.). *Teaching and assessing clinical competence*. Groningen: BoekWerk Publications, 1989.

## DE AUTEURS

*L.W.T. Schuwirth, arts, is als universitair docent verbonden aan de vakgroep Onderwijs- ontwikkeling en Onderwijsresearch, lid van het Projekt Evaluatie van Studieresultaten van de Rijksuniversiteit Limburg.*

*J.P.P. Schrandt is als kinderarts verbonden aan het Academisch Ziekenhuis Maastricht.*

*C.P.M. van der Vleuten, psycholoog, is universitair hoofddocent bij de vakgroep Onderwijsontwikkeling en Onderwijsresearch en projectleider van het Projekt Evaluatie van Studieresultaten van de Rijksuniversiteit Limburg te Maastricht*

## Correspondentie-adres:

*L.W.T. Schuwirth, Vakgroep Onderwijsontwikkeling en Onderwijsresearch, Rijksuniversiteit Limburg, Postbus 616, 6200 MD Maastricht*