De toekomst van ROM: computergestuurd adaptief testen

G. FLENS, E. DE BEURS

ACHTERGROND Meetinstrumenten die worden gebruikt voor het monitoren van de behandeluitkomst bij patiënten in de ggz zijn ontwikkeld op basis van de principes uit de klassieke testtheorie. Omdat de aannames achter deze theorie achterhaald zijn, is het tijd om toe te werken naar een nieuwe meetmethode die is gebaseerd op de itemresponstheorie: computergestuurd adaptief testen (CAT).

DOEL De CAT-methodiek introduceren bij de Nederlandse ggz, en een overzicht geven van de huidige en toekomstige ontwikkelingen.

METHODE We lichten toe wat CAT is, wat de voordelen zijn voor de ggz, wat de beperkingen zijn, welke nationale en internationale ontwikkelingen er al zijn geweest, en welke ontwikkelingen nog wenselijk zijn.

RESULTATEN De patient-reported outcomes measurement information system(PROMIS)-itembanken voor angst en depressie voor volwassenen illustreren dat het mogelijk is om met CAT efficiënter te meten dan met het klassieke meetinstrumentarium, met zeer precieze uitkomsten.

CONCLUSIE Eind 2017, begin 2018 komen de eerste gevalideerde CAT's voor angst en depressie beschikbaar voor volwassenen in de ggz. De huidige resultaten zijn veelbelovend en de CAT-technologie brengt meten in de ggz op een hoger plan.

TIJDSCHRIFT VOOR PSYCHIATRIE 59(2017)12, 767-774

TREFWOORDEN angst, computergestuurd adaptief testen, depressie, itemresponstheorie, routine outcome monitoring

ARTIKEL

Tijdens de behandeling van patiënten in de geestelijke gezondheidzorg (ggz) worden meetinstrumenten periodiek afgenomen om de voortgang te monitoren (ROM; de Beurs e.a. 2011). Tot nog toe werden deze meetinstrumenten in Nederland ontwikkeld volgens de klassieke testtheorie (KTT; Lord & Novick 1968). De KTT berust erop dat een geobserveerde testscore uit twee delen bestaat: de ware score en een algemene meetfout. De geobserveerde score komt eenvoudig tot stand, veelal door een vast aantal ruwe itemscores bij elkaar op te tellen. Deze score kan vervolgens samen met de meetfout worden gebruikt om de ware score te schatten. Bij de KTT beschouwt men de meetfout van de geobserveerde score als een algemene eigenschap van een meetinstrument die voor alle respondenten hetzelfde is. Als de meetfout van een meetinstrument dus bijvoorbeeld 2 punten is, dan ligt bij een geobserveerde score van 100 de ware score ergens tussen 98 en 102, en bij een geobserveerde van score van 120 ergens tussen 118 en 122. Dit alles betekent dat een meetinstrument betrouwbare scores oplevert voor alle respondenten voor wie het meetinstrument ontwikkeld is, zolang de vastgestelde meetprecisie maar voldoet aan de minimale eisen voor betrouwbaarheid (Lord & Novick 1968).

Klassieke testtheorie

Aan de KTT liggen de aannames ten grondslag dat alle items dezelfde informatiewaarde hebben en dat deze informatiewaarde ook hetzelfde is voor alle ernstniveaus. Deze aannames blijken in de praktijk niet goed houdbaar te zijn.
In de eerste plaats kunnen items wel degelijk een verschillende informatiewaarde hebben. Neem bijvoorbeeld de items '*ik voelde me angstig*' en '*mijn spieren trilden of vertrokken zich onwillekeurig*'. Respondenten met een hoger angstniveau zijn over het algemeen angstiger, maar heb-

ben niet meteen meer trillende spieren of onwillekeurig vertrokken spieren. Omdat het eerste item dus beter het construct angst meet (d.w.z. de schaal die het item meet), zou de respons een groter gewicht moeten krijgen in het bepalen van het angstniveau en de meetprecisie.

In de tweede plaats kan de informatiewaarde van een item ook verschillen tussen ernstniveaus. Voor ggz-meetinstrumenten geldt over het algemeen dat items informatiever zijn voor respondenten met een gemiddeld ernstniveau dan voor respondenten met een lager of hoger ernstniveau (Reise & Waller 2009). Dit blijkt bijvoorbeeld uit het item '*ik voelde me angstig*'. Dit item is voornamelijk informatief voor respondenten met een gemiddeld angstniveau omdat je slechts een beperkte aanname kan doen over de respons die ze zullen geven. Met behulp van de respons leer je dus wat nieuws over de respondent. Voor respondenten met een laag of hoog angstniveau leer je daarentegen weinig nieuws omdat je redelijkerwijs kan aannemen dat ze respectievelijk bijna nooit en bijna voortdurend angstig zullen zijn. Zodra er door het voorleggen van een of meerdere items dus een vermoeden bestaat dat een respondent een laag of hoog angstniveau heeft, zal het item 'ik voelde me angstig' nog maar weinig informatie toevoegen aan de meetprecisie en het angstniveau, terwijl dit item daar juist wel een relevante bijdrage aan kan leveren als het vermoeden bestaat dat de respondent een gemiddeld angstniveau heeft.

Itemresponstheorie

Omdat de aannames van KTT niet goed houdbaar zijn in de praktijk is de behoefte ontstaan om scores en meetprecisie te bepalen op basis van zowel persoonseigenschappen als itemeigenschappen. Dit heeft geleid tot de itemresponstheorie (IRT; Embretson & Reise 2000). IRT is een theorie waarbij de ware score op een meetinstrument niet wordt geschat op basis van een geobserveerde score en een meetinstrumentgerelateerde meetprecisie, maar op basis van een geschatte score en een respondentgerelateerde meetprecisie. Voor het bepalen van deze uitkomsten wordt behalve van de respons op een item tevens gebruikgemaakt van de eigenschappen die een item heeft voor een construct. De belangrijkste itemeigenschappen van ggz-constructen zijn *het onderscheidend vermogen* en *de drempelwaarde*.

Het onderscheidend vermogen is de mate waarin we met een item onderscheid kunnen maken tussen respondenten met verschillende ernstniveaus. Hierbij geldt: hoe beter het onderscheidend vermogen, hoe beter het item onderscheid kan aangeven tussen respondenten die maar weinig verschillen in ernstniveau. Met het item '*ik voelde me angstig*' kan men bijvoorbeeld het angstniveau van respondenten beter onderscheiden dan met het item '*mijn spieren trilden of vertrokken zich onwillekeurig*'.

AUTEURS

GERARD FLENS, onderzoeker Stichting Benchmark GGZ.
EDWIN DE BEURS, hoofd wetenschappelijk onderzoek Stichting Benchmark GGZ en hoogleraar ROM en Benchmarken, Universiteit Leiden.

CORRESPONDENTIEADRES
Gerard Flens
E-mail: gerard.flens@sbggz.nl

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 6-6-2017.

De drempelwaarde van een item is het ernstniveau vanaf waar het waarschijnlijker wordt dat een respondent een hogere responscategorie verkiest boven een lagere bij twee opvolgende responscategorieën. Het aantal drempelwaarden van een item is daarmee gelijk aan het aantal responscategorieën minus 1, en geven de ernstniveaus aan waarvoor het item het informatiefst is. Het item 'ik maakte me zorgen' heeft bijvoorbeeld vooral drempelwaarden bij de lagere en gemiddelde ernstniveaus, terwijl het item 'ik voelde me doodsbang' juist vooral drempelwaarden heeft bij de gemiddelde en hogere ernstniveaus.

Met deze itemeigenschappen is het niet meer nodig dat alle items van een meetinstrument worden afgenomen om het ernstniveau te bepalen, maar kan deze al (samen met de meetprecisie) ruwweg worden geschat met de respons op een enkel item. Omdat de meetprecisie bij een enkel item over het algemeen echter nog te laag is om betrouwbare gevolgtrekkingen te kunnen maken (Gliem & Gliem 2003), worden aanvullende items afgenomen. Op deze manier zal de meetprecisie steeds wat meer stijgen waardoor de schatting van het ernstniveau gaandeweg beter wordt. De uiteindelijke meetprecisie kan hierbij verschillend zijn voor respondenten afhankelijk van hun geschatte ernstniveau.

Computergestuurd adaptief testen

Het bepalen van het ernstniveau op basis van een vast aantal items lijkt bij IRT in eerste instantie niet heel anders te zijn dan bij KTT. Deze moderne meettheorie maakt het echter ook mogelijk om dynamische meetmethoden toe te passen die eerder niet mogelijk waren.

Een veelbelovende en opkomende meetmethode is computergestuurd adaptief testen (CAT; Reeve e.a. 2007). CAT is een via de computer uitgevoerde test waarbij de computer items selecteert en voorlegt op basis van de respons op voorgaande items. Wanneer een respondent bijvoorbeeld

aangeeft dat hij of zij zich vaak angstig voelde, legt CAT automatisch een item voor dat past bij een hoger angstniveau, bijvoorbeeld '*ik voelde me doodsbang*', en legt geen items meer voor die passen bij een lager angstniveau, bijvoorbeeld '*ik maakte me zorgen*'. CAT blijft vervolgens nieuwe items aanbieden op deze manier, maar slechts zo lang als nodig is om het ernstniveau met een vooraf ingestelde meetprecisie te schatten. Het aantal aangeboden items kan hierbij verschillen tussen respondenten omdat de meetprecisie afhankelijk is van het ernstniveau. De definitieve scores van respondenten blijven echter vergelijkbaar doordat ze worden geschat op een gestandaardiseerde schaal met een gemiddelde van 0 en een standaarddeviatie van 1 (Smits e.a. 2012; Flens e.a. 2016; 2017). Om de interpreteerbaarheid van deze schaal te verhogen wordt er in de praktijk vaak nog een normering toegepast op basis van patiënten en/of de algemene bevolking naar de welbekende T-scoreschaal met een gemiddelde van 50 en een standaarddeviatie van 10 (Cella e.a. 2010).

Met CAT kan de belasting voor patiënten worden geminimaliseerd met een kortere test terwijl de meetprecisie hetzelfde blijft (Flens e.a. 2017). IRT wordt daarom gaandeweg de geprefereerde testtheorie voor de ontwikkeling van nieuwe meetinstrumenten. In de VS worden CAT's al volop gebruikt in de klinische praktijk (zie ook http://www.healthmeasures.net), en ook in Nederland zien we de ontwikkeling van CAT's langzaam maar zeker op gang komen. Dit is daarom het moment om CAT te introduceren aan de Nederlandse ggz, en een overzicht te geven van de huidige ontwikkelingen.

We beginnen dit overzichtsartikel met een nadere toelichting op de CAT-methodologie om de lezer bekend te maken met de basisprincipes. Vervolgens gaan we in op de wenselijkheid van deze nieuwe methodologie in het ggz-veld en exploreren de mogelijke beperkingen aan CAT. Ten slotte geven we een overzicht van de huidige CAT-ontwikkelingen in de internationale en de Nederlandse ggz, en worden er wenselijke CAT-ontwikkelingen voor de toekomst geschetst.

Hoe werkt CAT?

CAT bestaat uit drie componenten:

– een itembank;
– de bijbehorende itemparameters;
– de specificaties voor de werking van de CAT.

In de volgende paragrafen lichten we deze drie componenten toe.

Itembanken bevatten een willekeurig aantal items om een construct te meten. Bij voorkeur is het aantal items aanzienlijk zodat het construct gedekt wordt over de volle breedte, van licht tot zwaar, van makkelijk tot moeilijk. De informatiewaarde van de itembank als geheel is dan optimaal.

Voordat een itembank ingezet kan worden als CAT is het noodzakelijk dat deze geëvalueerd wordt op psychometrische kwaliteit. Deze evaluatie wordt uitgevoerd op een steekproef die representatief is voor de doelgroep waarvoor het meetinstrument ontwikkeld wordt, bijvoorbeeld een specifieke patiëntenpopulatie, de algemene Nederlandse bevolking, of een combinatie daarvan. De itembank wordt uiteindelijk samengesteld uit de items die (samen) de beste psychometrische eigenschappen hebben.

Voor een uitgebreide beschrijving van de psychometrische eigenschappen die onderzocht worden voor een IRT-itembank verwijzen we naar het artikel van Reeve e.a. (2007). Voor een praktische uitwerking daarvan, zie Smits e.a. (2012). Wij gaan voor het doel van dit artikel alleen in op het meest kenmerkende gedeelte van de psychometrische evaluatie van IRT-itembanken: de itemparameters.

Itemparameters representeren de eigenschappen van items ten opzichte van het construct dat ze meten, en worden gebruikt bij het bepalen van de meetprecisie en het ernstniveau. De parameters worden voor elk item afzonderlijk geschat, en geven enerzijds de verschillen aan tussen items, en anderzijds – binnen een item – de verschillen tussen ernstniveaus.

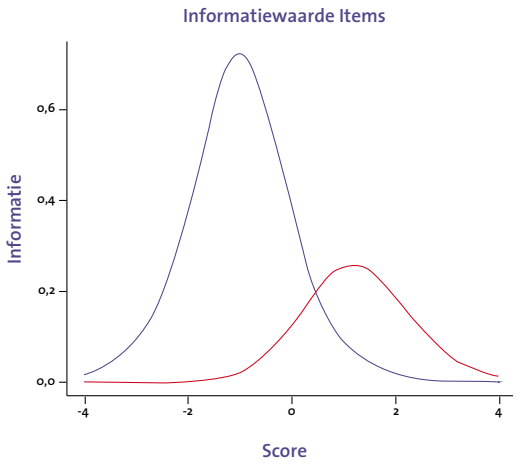
Om dit te illustreren staat in **FIGUUR 1** de informatiewaarde van twee fictieve items weergegeven op een standaard IRT-schaal (d.w.z. gemiddelde = 0, standaarddeviatie = 1). De informatiewaarde die een item voor een construct heeft, wordt bepaald op basis van de itemparameters, en geeft voor alle ernstniveaus aan hoe precies dat construct gemeten kan worden. Met de illustratie wordt duidelijk dat het eerste item (paarse lijn) informatiever is (d.w.z meer meetprecisie toevoegt) voor de lage en gemiddelde ernstniveaus, terwijl het tweede item (rode lijn) juist informatiever is voor de hoge ernstniveaus. Binnen een CAT-afname zal daarom het eerste item eerder geselecteerd worden als de schatting van het ernstniveau laag of gemiddeld is, en het tweede item als deze schatting hoog is. Daarnaast is duidelijk te zien dat de informatiewaarde voor beide items verschilt tussen alle ernstniveaus. Afhankelijk van het geschatte ernstniveau wordt er dus door elk item een verschillende mate van meetprecisie toegevoegd.

Specificaties voor de werking van de CAT: na het vaststellen van de definitieve itembank met bijbehorende itemparameters wordt de CAT ingericht. Een CAT-afname bestaat uit vier onderdelen:

1. een startitem;
2. een scoreschatter;
3. een itemselectieprocedure;
4. een stopregel.

CAT begint met het voorleggen van een eerste item. Dit item kan willekeurig gekozen worden, maar meestal wordt het item genomen dat de grootste informatiewaarde heeft

FIGUUR 1 **Informatiewaarde van twee fictieve items**



Item 1 = paarse lijn, item 2 = rode lijn.

voor respondenten met een gemiddeld ernstniveau. Nadat er een respons is gegeven, schat de software een voorlopige score en wordt de bijbehorende meetprecisie bepaald. Vervolgens evalueert de software de resultaten tegen een stopregel. Dit kunnen een of meer voorwaarden zijn waar de resultaten aan moeten voldoen om het aanbieden van items stop te zetten. Veelgebruikte voorwaarden zijn een vooraf ingestelde meetprecisie, een maximaal aantal aangeboden items, of een combinatie daarvan (Smits e.a. 2012). Zolang de stopregel niet is bereikt, wordt er een nieuw item geselecteerd en aangeboden. Ook dit item wordt weer gekozen op basis van de grootste informatiewaarde, alleen nu waar deze het grootst is gezien het tijdelijk geschatte ernstniveau. De procedure wordt herhaald totdat de stopregel is bereikt en het definitieve ernstniveau kan worden vastgesteld.

Waarom CAT introduceren in de ggz?

De ggz bevindt zich in een uitstekende uitgangspositie voor de invoering van de CAT-technologie: ROM wordt al op grote schaal toegepast met ondersteuning van ICT-oplossingen, en nagenoeg iedereen heeft een computer en/of smartphone met internetaansluiting om CAT-meetinstrumenten in te kunnen vullen. Dat het mogelijk is om een nieuwe technologie toe te passen, betekent echter nog niet dat het wenselijk is. Waarom hebben we CAT nodig voor het monitoren van patiënten in de ggz?

EFFICIËNTIE

ROM wordt gebruikt voor verschillende doelstellingen. Meetinstrumenten worden gebruikt ter ondersteuning van de behandeling, voor verbetering van de dienstverlening, wetenschappelijk onderzoek en benchmarking. Deze verscheidenheid aan doelstellingen en de wens om op meerdere meetdomeinen uitkomsten vast te stellen (klachten, functioneren, kwaliteit van leven, persoonlijk herstel, generiek en stoornisspecifiek) heeft als consequentie dat de patiënt steeds meer meetinstrumenten moet invullen. Hiermee neemt de belasting toe en is het haast onvermijdelijk dat er verlies aan gegevens optreedt. Dit verlies is ten dele terug te voeren op een begrijpelijk gebrek aan bereidheid van een deel van de patiënten om veel meetinstrumenten in te vullen, waarin items voorgelegd kunnen worden die in de beleving van de respondent irrelevant of overbodig zijn. Zo hoeft het item '*ik maakte me zorgen*' niet relevant te zijn voor een patiënt met hevige angstaanvallen omdat het vanzelfsprekend is dat hij/zij zich zorgen maakt.

Door alleen items voor te leggen die het best passen bij het ernstniveau gezien de antwoorden op voorgaande items, duurt de meetinstrumentafname korter, waarmee verlies aan gegevens gedeeltelijk kan worden tegengegaan. Hierdoor ontstaat er ook meer ruimte om aanvullende concepten te meten die relevante informatie voor de behandeling kunnen opleveren.

PRECISIE (BETROUWBAARHEID)

In de ggz worden generieke en stoornisspecifieke meetinstrumenten gebruikt. Sommige instrumenten bieden zowel een generieke totaalscore als stoornisspecifieke subschaalscores. De bruikbaarheid van deze subschaalscores kan echter twijfelachtig zijn. Subschalen die zijn ontwikkeld met de KTT worden geëvalueerd op hun algemene meetprecisie, maar niet op hun specifieke meetprecisie voor verschillende ernstniveaus. Deze specifieke meetprecisie kan voor veel ernstniveaus beperkt zijn, omdat het aantal items van subschalen vaststaat en vaak laag is om de afnameduur van het gehele meetinstrument binnen de perken te houden. Bij CAT daarentegen maakt men gebruik van itembanken die een veel grotere hoeveelheid items kunnen omvatten. Door de beschikbaarheid van een grotere selectie items kan CAT aan alle respondenten de voor hen meest informatieve items aanbieden. Het gevolg is dat de meetprecisie (en daarmee de bruikbaarheid) van subschaalscores geoptimaliseerd kan worden terwijl het aantal aangeboden items toch laag blijft.

Beperkingen van CAT

Naast de voordelen van CAT zijn er ook enkele beperkingen waar we rekening mee moeten houden. Deze beperkingen betreffen complexiteit van de CAT-methodologie, representativiteit van de itemparameters, en afname van verschillende items over tijd.

In de eerste plaats is de CAT-methodologie moeilijker te begrijpen dan de KTT-methodologie door het gebruik van

complexe schattingsprocedures en het feit dat scores vergelijkbaar blijven ondanks het afnemen van verschillende items (Smits e.a. 2012; Flens e.a. 2017). Deze complexiteit kan als gevolg hebben dat CAT minder snel omarmd wordt, omdat behandelaars niet willen werken met iets wat ze nog onvoldoende begrijpen. Om deze acceptatie te vergroten, is het relevant dat er toegankelijke informatie wordt gepubliceerd over de CAT-methodologie en de resultaten van empirisch onderzoek. Daarnaast is het wenselijk om bijeenkomsten voor ggz-instellingen te organiseren voor het opbouwen van gewenning en vertrouwen bij behandelaren.

In de tweede plaats is het noodzakelijk om de volledige itembank periodiek af te nemen bij een steekproef om representatieve itemparameters te behouden. Door maatschappelijke veranderingen kunnen er namelijk wijzigingen optreden in het klachtenverloop van patiënten, de ernst van de klachten, of de betekenis van items (Flens e.a. 2017). Hierdoor kunnen de itemparameters ook veranderen. Deze periodieke dataverzameling is in principe ook noodzakelijk bij het klassieke meetinstrumentarium voor het bepalen van normscores, maar is vaak al beschikbaar doordat het meetinstrument wordt gebruikt in de dagelijkse praktijk. Omdat CAT echter niet de gehele itembank afneemt, maar deze wel nodig is om de itemparameters te kunnen schatten, is het noodzakelijk om gegevens van de volledige itembank afzonderlijk te verzamelen.

Ten slotte beschouwen sommige behandelaars als bezwaar van CAT dat individuele itemscores niet meer gebruikt kunnen worden om de voortgang van een patiënt te monitoren, omdat CAT verschillende items afneemt bij verschillende ernstniveaus. Dit bezwaar berust echter op de misvatting dat individuele itemscores voldoende betrouwbaar zijn voor dit doel. De meetprecisie van een enkel item is over het algemeen te klein voor betrouwbare gevolgtrekkingen over (verandering in) symptomen, waardoor individuele items alleen met grote terughoudendheid gebruikt kunnen worden bij het monitoren van de behandeling (Gliem & Gliem 2003). Dat CAT het dus moeilijk maakt om individuele itemscores van patiënten over de tijd te vergelijken kan daarom ook juist als voordeel worden gezien. CAT plaatst namelijk automatisch de focus op de enige score die wel voldoende meetprecisie heeft: de geschatte eindscore.

Huidige status van CAT in de ggz

De ambitie om tot een state-of-the-artmeetinstrumentarium te komen op basis van IRT blijkt haalbaar, wat is aangetoond met het *patient-reported outcomes measurement information system*(PROMIS)-initiatief in de Verenigde Staten. Het Amerikaanse PROMIS-initiatief ging in 2002 van start met een inventarisatie van domeinen voor het meten van zelfgerapporteerde gezondheid en welbevinden. Meetdomeinen kunnen een klachtgebied zijn (pijn of depressie), of een vermogen of capaciteit (mobiliteit of fysieke functionaliteit). Inmiddels zijn er voor diverse meetdomeinen een groot aantal itembanken ontwikkeld, gevalideerd en genormeerd (Cella e.a. 2010).

Voor de ggz zijn PROMIS-itembanken ontwikkeld voor angst, depressie en boosheid, zowel voor volwassenen (Pilkonis e.a. 2011) als voor jongeren (Irwin e.a. 2010). Onderzoek naar deze itembanken heeft aangetoond dat CAT een efficientere, accuratere, preciezere en responsievere methodologie is voor het meten van psychische problemen dan het klassieke Amerikaanse meetinstrumentarium (Pilkonis e.a. 2014; Schalet e.a. 2016).

Het internationale antwoord hierop mag er zijn: de itembanken zijn al in veel talen vertaald (http://www.nihpromis.org/measures/translations). De Nederlands-Vlaamse PROMIS-groep heeft een groot aantal itembanken vertaald in het Nederlands-Vlaams (http://www.dutchflemishpromis.nl; Terwee e.a. 2014; Haverman e.a. 2016). Deze itembanken worden ook wel aangeduid als 'Dutch-Flemish'.

Van de zes beschikbare PROMIS-itembanken voor de Nederlandse ggz zijn de angst-itembank (29 items) en depressie-itembank (28 items) voor volwassenen onderzocht op de bruikbaarheid voor CAT (Flens e.a. 2017). Op basis van een grote steekproef waarin zowel patiënten met veelvoorkomende psychische stoornissen in ambulante behandeling zijn opgenomen (n = 1008) als respondenten uit de Nederlandse bevolking (n = 1002), is vastgesteld dat beide itembanken uitstekende psychometrische eigenschappen bezitten.

Daarnaast is met een *CAT-simulatie* op deze data vastgesteld dat CAT resulteert in efficiënt en zeer precies meten, en nagenoeg net zo nauwkeurig is als de afname van de volledige itembanken. Een CAT-simulatie is geen echte CAT-afname, maar gebruikt de data van een volledige itembankafname en evalueert deze data alsof ze als CAT zijn afgenomen. Voor de simulatie is een combinatiestopregel gebruikt, bestaande uit een hoge meetprecisie (standaardmeetfout < 0,22; Bernstein & Nunnally 1994) en een maximaal aantal items (angst = 12; depressie = 9).

Met deze stopregel werden gemiddeld 8 items afgenomen om angst te meten en 6 items om depressie te meten. De CAT-scores waren daarnaast hoog gecorreleerd met de volledige itembankscores (r = 0,98), lieten een verwaarloosbaar verschil in gemiddelde score zien met de volledige itembankscores (Cohens d = 0,02), en waren gelijkwaardig in staat patiënten van respondenten zonder stoornis te onderscheiden.

Een ander noemenswaardig resultaat is dat het verlagen van de meetprecisie van hoog naar voldoende (standaardmeetfout < 0,32; Bernstein & Nunnally 1994) ervoor zorgde

dat angst gemeten kon worden met gemiddeld 4 items en depressie met 3 items. Een wat lagere meetprecisie kan prima gebruikt worden bij een eenmalige CAT-afname (bijvoorbeeld ter screening), maar voor ROM gaat de voorkeur uit naar een hoge meetprecisie (Flens e.a. 2017). Dit is namelijk een voorwaarde om klinisch significante verandering bij patiënten goed aan te kunnen tonen, waardoor de behandeling eerder bijgestuurd of afgerond kan worden. De voordelen die CAT biedt voor de efficiëntie van meetinstrumenten en de meetprecisie van scores zijn ook aangetoond in tal van andere onderzoeken (bijvoorbeeld Smits e.a. 2012; Pilkonis e.a. 2014; Flens e.a. 2016; Schalet e.a. 2016). Voordat de Nederlandse CAT's echter gebruikt kunnen worden in de praktijk is het noodzakelijk om echte CAT-afnames te realiseren. Vanaf eind 2016 heeft de Nederlands-Vlaamse PROMIS-groep software beschikbaar waarmee dit mogelijk is geworden. De groep heeft hierin samengewerkt met IRT-deskundigen van Universiteit Twente en Rijksuniversiteit Groningen.

In eerste instantie zijn CAT's beschikbaar voor onderzoeksprojecten, maar naar verwachting wordt in 2017 een Nederlands-Vlaams Assessment Center opgericht waarmee Nederlands-Vlaamse PROMIS-CAT's aangeboden kunnen worden aan het veld. Op termijn kan dit via een website die door het Nederlands-Vlaamse Assessment Center beschikbaar zal worden gesteld. Het is echter nu al mogelijk om de software van de Nederlands-Vlaamse PROMIS-groep te koppelen aan de software van ICT-leveranciers voor ROM. VitalHealth is de eerste leverancier die deze koppeling heeft gerealiseerd.

In de eerste helft van 2017 zal in samenwerking met enkele ggz-instellingen een onderzoek starten naar de toepasbaarheid van de Nederlands-Vlaamse PROMIS-CAT's voor angst en depressie in de Nederlandse ggz. Dit onderzoek heeft twee doelen: ten eerste demonstreren dat CAT de belofte waarmaakt van efficiënt en precies meten, en ten tweede vergelijken van de responsiviteit van de CAT's met de angst- en depressiesubschalen van de *Symptom Questionnaire* (SQ48) en de *Brief Symptom Inventory* (BSI).

Ook de ICT-leverancier Roqua (van het Universitair Centrum Psychiatrie; UMCG) heeft een web-gebaseerde toepassing van CAT gerealiseerd. De alfaversie van deze applicatie omvat de itembanken angst en depressie (PROMIS), distress (uit de *VierDimensionale KlachtenLijst*; 4DKL), en positieve en negatieve symptomen van psychose (uit de *Prodromal Questionnaire*). Daarnaast bevat deze testbatterij de PROMIS-itembanken voor emotionele steun, vriendschap en satisfactie met sociale rollen en activiteiten. Deze aanvullende itembanken worden vooralsnog niet adaptief gemeten omdat de itemparameters nog niet voor gebruik in Nederland zijn gevalideerd. De applicatie heet 'CATja' en zal huisartsen en praktijkondersteuners huisartsen ggz faciliteren bij het inschatten van de zorgbehoefte van hun cliënten. Vanaf februari 2017 wordt de applicatie in de praktijk getest, en naar verwachting is de bètaversie van de applicatie in de tweede helft van 2017 beschikbaar voor de drie noordelijke provincies van Nederland.

Toekomstige ontwikkelingen

Terwijl CAT al volop wordt gebruikt in de Amerikaanse ggz staat deze moderne meettechniek in de Nederlandse ggz nog in de kinderschoenen. We verwachten dat de eerste gevalideerde CAT's eind 2017, begin 2018 beschikbaar zullen zijn voor het veld, maar er zal nog veel moeten gebeuren om de CAT-methodiek tot volle potentie te brengen. Welke ontwikkelingen zijn hier in ieder geval wenselijk voor?

In de eerste plaats is het nodig dat – naast VitalHealth – andere ICT-leveranciers van ROM gaan koppelen op het Nederlands-Vlaamse assessmentcentrum. Dit maakt het voor de meeste ggz-instellingen mogelijk om CAT's geïntegreerd via hun eigen ROM-applicatie aan te bieden aan hun patiënten. Ggz-instellingen die geen gebruik maken van de diensten van een ICT-leverancier voor ROM kunnen zelf koppelen op het assessmentcentrum. Voor het realiseren van de koppeling met het assessmentcentrum kan contact worden opgenomen met de Nederlands-Vlaamse PROMIS-groep (http://www.dutchflemishpromis.nl).

In de tweede plaats is het belangrijk dat er CAT's ontwikkeld worden voor meer constructen. De ggz-itembanken van PROMIS die bijvoorbeeld wel zijn vertaald, maar nog niet worden gevalideerd, zijn boosheid voor volwassenen, en angst, depressie en boosheid voor de kinder- en jeugdpsychiatrie. Daarnaast kan er initiatief genomen worden voor de ontwikkeling en validatie van CAT's voor andere constructen. Hierbij kunnen we denken aan generieke constructen (bijvoorbeeld functioneren en gezondheidgerelateerde kwaliteit van leven), stoornisspecifieke constructen (bijvoorbeeld piekeren of sociale angst), constructen uit andere zorgdomeinen (bijvoorbeeld ouderenpsychiatrie of middelenmisbruik), of constructen die worden gemeten met beoordelingsschalen (in plaats van zelfrapportage).

Omdat de ontwikkeling van deze nieuwe itembanken veel middelen kan vragen (Cella 2010), raden we aan om ook items van bestaande meetinstrumenten te overwegen voor itembanken. Het grote voordeel hiervan is dat de benodigde data voor de psychometrische evaluatie van deze itembanken over het algemeen al beschikbaar zijn. Voor de *Mood and Anxiety Symptom Questionnaire* (MASQ) is op deze manier bijvoorbeeld aangetoond via een grote steekproef bestaande uit patiënten met veelvoorkomende psychische stoornissen (n = 3597) dat de drie subschalen 'positief affect', 'negatief affect' en 'somatische angst' zeer geschikt zijn als itembanken voor CAT (Flens e.a. 2016).

Daarnaast kan men ook overwegen om schalen met een gelijkwaardige meetpretentie uit verschillende meetinstrumenten samen te voegen voor een nog informatievere itembank. Het valideren van de CAT's ten slotte zal patiënten over het algemeen minder belasten dan bij de validatie van KTT-meetinstrumenten gebeurt, omdat het aantal aangeboden items onder CAT relatief laag is.

Besluit

We verwachten dat het monitoren van ggz-patiënten een sterke stimulans zal krijgen als deze ontwikkelingen worden opgepakt. Met CAT is een win-winsituatie te creëren: patiënten worden minder belast, terwijl de meetprecisie behouden blijft. Hierdoor kunnen behandelaars patiënten vaker verzoeken een meting in te vullen en is er meer ruimte om aanvullende constructen te meten. Uiteindelijk zal er daardoor meer informatie aanwezig zijn om de patiënt uitgebreid te karakteriseren, de behandeling te evalueren en deze waar nodig bij te sturen, of op of af te schalen. Uit dit alles moge blijken: CAT is de toekomst van ROM.

LITERATUUR

- Bernstein IH, Nunnally JC. Psychometric theory (3de ed.). New York; McGraw-Hill: 1994.
- Beurs E de, den Hollander-Gijsman M, van Rood Y, van der Wee N, Giltay E, van Noorden M, e.a. Routine outcome monitoring in the Netherlands: practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. Clin Psych Psychother 2011; 18: 1-12.
- Cella D, Riley W, Stone A, Rothrock N, Reeve B, Youn, S, e.a. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010; 63: 1179-94.
- Embretson S, Reise S P. Item response theory for psychologists. Mahwah: Erlbaum; 2000.
- Flens G, Smits N, Carlier I, van Hemert A M, de Beurs E. Simulating computer adaptive testing with the Mood and Anxiety Symptom Questionnaire. Psychol Assess 2016; 28: 953-62.
- Flens G, Smits N, Terwee CB, Dekker J, Huijbrechts I, de Beurs E. Development of a Computer Adaptive Test for Depression based on the Dutch-Flemish version of the PROMIS Item Bank. Eval Health Prof 2017; 40: 79-105.
- Gliem J, Gliem R. Calculating, interpreting, and reporting Cronbach's Alpha reliability coefficient for Likert-type scales. In 2003 Midwest Research to Practice Conference in Adult, Continuing and Community Education 2003. Columbus, OH.
- Haverman L, Grootenhuis MA, Raat H, van Rossum MA, van Dulmen-den Broeder E, Hoppenbrouwers K, e.a. Dutch-Flemish translation of nine pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)®. Qual Life Res 2016; 25: 761-5.
- Irwin DE, Stucky B, Langer MM, Thissen D, DeWitt EM, Lai S, e.a. An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. Qual Life Res 2010; 19: 595-607.
- Lord FM, Novick MR. Statistical theories of mental lest scores. Reading: Addison-Wesley; 1968.
- Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment 2011; 18: 263-83.
- Pilkonis PA, Yu L, Dodds NE, Johnston K L, Maihoefer CC, Lawrence SM. Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. J Psychiatr Res 2014; 56: 112-9.
- Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, e.a. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007; 45: S22-31.
- Reise SP, Waller NG. Item response theory and clinical measurement. Rev Clin Psych 2009; 5: 27-48.
- Schalet BD, Pilkonis PA, Yu L, Dodds N, Johnston KL, Yount S, e.a. Clinical validity of PROMIS Depression, Anxiety, and Anger across diverse clinical samples. J Clin Epidemiol 2016; 73: 119-27.
- Smits N, Zitman FG, Cuijpers P, den Hollander-Gijsman ME, Carlier IVE. A proof of principle for using adaptive testing for routines outcome monitoring: the efficiency of the Mood and Anxiety Symptoms Questionnaire – Anhedonic Depression CAT. BMC Med Res Methodol 2012; 12: 4.

SUMMARY

The future of ROM: computerised adaptive testing

G. FLENS, E. DE BEURS

BACKGROUND Measurement instruments that are used for monitoring patients in mental health care are developed according to the principles of classical test theory. Because the assumptions underlying this theory are outdated, this is a good time to work towards a new method of measurement known as computerised adaptive testing (CAT), the method being based on item response theory.

AIM To introduce the CAT-methodology into Dutch mental health care, and provide an overview of the current and desirable developments.

METHOD We explain what CAT is and why mental health care should warmly welcome this new development. We also outline the limitations of CAT, summarise the developments that have already been made nationally and internationally and consider some developments we think are desirable.

RESULTS PROMIS item banks for anxiety and depression for adults show that CAT is more efficient than instruments currently in use and is able to produce very precise outcomes.

CONCLUSION The first CATs for anxiety and depression will be available late 2017 or early 2018 for adults receiving mental health care in the Netherlands. Recent results are very impressive and CAT-technology will increase the efficiency of symptom measurements, bringing these measurements to a higher level.

TIJDSCHRIFT VOOR PSYCHIATRIE 59(2017)12, 767-744

KEY WORDS anxiety, computerised adaptive testing, depression, item response theory, routine outcome monitoring