Testen via internet is niet nieuw, maar de ontwikkelingen raken in een versnelling. Wouter Lucassen, lid van de COTAN, laat zien dat computerkracht en geavanceerde testtheorie toepassingen mogelijk maken die de psychodiagnostiek echt zullen vernieuwen. *De COTAN heeft onlangs een standpunt ingenomen over de situatie waarin een cliënt op afstand en zonder aanwezigheid van een psycholoog getest wordt. Ook vanuit de NIP-beroepscode zijn er over testen via internet een paar behartenswaardige opmerkingen te maken.*

EEN STAND VAN ZAKEN

INTERNET-PSYCHODIAGNOSTIEK

Testen via internet: de ontwikkelingen gaan hard. Daarbij moeten we niet alleen denken aan intelligentietests of vragenlijsten, al dan niet als aanpassing van een reeds bestaande papier-en-potloodtest. Dat is immers een ontwikkeling die zich al decennia voltrekt. Opvallend en uitdagend zijn juist die vormen van internet-diagnostiek die gebruik maken van de kracht van computers: opslagcapaciteit, complexe berekeningen en verwerkingssnelheid. In dit artikel zullen we daarvan diverse voorbeelden geven.

Ook van de multimedia-mogelijkheden die computers bieden, ondervindt de traditionele psychologische vragenlijst concurrentie. Zo is Dominic Interactive een screeningsinstrument om jonge kinderen op een aansprekende manier te laten rapporteren in hoeverre zij zich herkennen in de problemen van Dominic. Dominic kan een jongen of meisje zijn, spreekt net zo gemakkelijk Nederlands, Catalaans of Turks, en kan er met evenveel gemak Europees, Latijns-Amerikaans of Aziatisch uitzien (Kuijpers et al., 2014). Vernieuwend is ook dat onder Plato's motto 'Je leert iemand beter kennen door een uur spelen dan door een jaar praten' in toenemende mate *serious games* worden ontwikkeld die spannende 3D-omgevingen niet voor entertainment gebruiken maar voor selectie; zij hebben potentieel om constructen te meten waar traditionele assessments moeite mee hebben, zoals creativiteit en leiderschap (Kato & De Klerk, 2017).

Voor alle duidelijkheid: we[1] hebben het in dit artikel niet over screenings-tests van allerlei maatschappelijke instanties (bijvoorbeeld de autisme-test van PsyQ), nuttige oefentests (www.oefenassessment.nl), rechtstreeks aan de consument aangeboden gratis of betaalde tests (www.123test.nl) of pret-testjes ('Wat voor liefdestype ben jij?'). Hier gaat het over tests en toetsen die via internet in opdracht van een psycholoog of een gelijkwaardige professional worden aangeboden om te worden gebruikt voor een individuele beslissing of in het kader van een behandeling. Een uitputtend overzicht kan het niet zijn, het is veeleer een tussentijds beeld van diverse ontwikkelingen en hun consequenties voor de beroepspraktijk. We kijken hiernaar vanuit het testtechnische perspectief, de beroepspraktijk en de beroepsethiek.

EINDE VAN DE PAPIER-EN-POTLOODTEST?

Dat de papier-en-potloodtest zijn langste tijd heeft gehad, zou een voorbarige conclusie zijn. Er zijn tal van instrumenten die zich er niet voor lenen om via internet te worden af-

1 Met dank aan de COTAN-collega's Arne Evers, Karin Vermeulen en Wendy de Leng.

Veel kinderen vinden een papier-en-potloodtoets gewoon fijn

genomen. Bij de *Bayley Scales of Infant and Toddler Development* (Van Baar et al., 2014) zijn de items non-verbaal van karakter, wordt allerlei materiaal gebruikt en worden de responsen, de motorische reacties en het gedrag van het kind door de psycholoog geobserveerd en gescoord. De in de verkeerspsychologie gebruikte test voor *Peripheral Perception* (Schuhfried, 2012) vereist naast een computer apparatuur om in het perifere gezichtsveld bewegende lichtstimuli te kunnen presenteren alsmede een paneel en voetpedalen om daarop te kunnen reageren. Bij de bekende *Rey Visual Design Learning Test* (Wilhelm et al., 2010) moet getekend worden. De *Woord Fluency Test* (Mulder et al., 2006) ten slotte is een mondelinge test, die vraagt dat de psycholoog haperingen, herhalingen, fouten en perseveraties noteert. Enzovoort. Wellicht is het ook een respectabele reden dat je als psycholoog gewoon contact met je cliënt wilt hebben…

Ook als een internet-afname in principe mogelijk is, kunnen er goede redenen zijn dat toch niet te doen. Een actueel voorbeeld is de IEP Eindtoets voor groep 8 basisonderwijs (Bureau ICE, 2018). Gezien de grootschalige afname hadden de auteurs voor een internetapplicatie kunnen kiezen, maar men heeft een bewuste keuze voor een papier-en-potloodtoets gemaakt omdat veel kinderen dat gewoon fijn vinden. De antwoorden op de meerkeuze- en open vragen worden gewoon in het boekje geschreven (niet eens op een antwoordformulier) en de scoring vindt plaats met scan- en herkenningssoftware. Op basis van het advies van de Expertgroep Toetsen Primair Onderwijs en de COTAN-beoordeling is de IEP in 2018 door de minister van OCW opnieuw toegelaten als alternatief voor de Centrale Eindtoets van het College voor Toetsen en Examens.

Verder moet men zich realiseren dat de markt groot genoeg moet zijn om het ontwikkelen van een internet-test rendabel te maken. Een belangrijke beperking ten slotte is dat het automatisch digitaal beoordelen van productieve vaardigheden (spreken, schrijven) nog toekomstmuziek is.

VOORDELEN VAN INTERNET-TESTS

Aansluitend bij de reeds gegeven voorbeelden zijn er allerlei algemene voordelen te noemen die het testen via internet met zich meebrengt. Niet te onderschatten is het argument dat deze wijze van testen past bij de internet-generatie. De instrumenten zijn 24/7 beschikbaar. Ze zijn doorgaans gemakkelijk in gebruik en regelen zichzelf van instructie tot afronding. Ze zijn goedkoop in afname – al zijn de ontwikkelingskosten vaak hoog. De standaardisatie van de testafname is optimaal geregeld, mits duidelijk wordt voorgeschreven wat voor computer of tablet gebruikt mag worden (en of een smartphone ook mag…). En natuurlijk blijft ook de omgeving waarin de test wordt afgenomen een aandachtspunt. Er is aanzienlijk minder kans op fouten, zowel bij de invulling door de cliënt als bij de scoring. Gegevens voor onderzoek kunnen automatisch worden verzameld. Op de resultaten kan de software een analyse uitvoeren die dieper kan gaan dan wat een psycholoog gewoonlijk kan doen. Desgewenst is een rapportage onmiddellijk beschikbaar. Bovendien kan de rapportage direct worden gekoppeld aan een gz-, volg-, portfolio- of HRM-systeem.

Internet-tests kunnen ook van betekenis zijn als men door *Routine Outcome Monitoring* (ROM) structureel en herhaaldelijk de toestand van de cliënt wil meten, vanwege het gemak van een afname, die desgewenst ook bij de cliënt thuis kan plaatsvinden. De systematisch opgeslagen data kunnen niet alleen voor het primaire doel worden gebruikt (bezien of en hoe de behandeling voortgezet moet worden), maar ook op instellingsniveau geëvalueerd worden, ter beschikking komen voor wetenschappelijk onderzoek en ten behoeve van benchmarking worden ingediend.

SIMULATIES

Internet is niet alleen geschikt om tests en vragenlijsten die uit een reeks items bestaan af te nemen, maar bij uitstek ook om praktijksimulaties, zoals die in de HRM-adviespraktijk gebruikelijk zijn, te digitaliseren. Klassiek is de postbakoefening, die het vermogen van een kandidaat toetst om verbanden te leggen tussen verschillende informatiebronnen, tot een juist oordeel te komen, beslissingen te nemen, te plannen en effectief te delegeren. Ook *situational judgement tests* (SJT's) lenen zich uitstekend voor digitalisering en afname via internet. Oostrom et al. (2010) beschrijven een SJT waarbij de computer niet alleen gebruikt wordt om werkgerelateerde situaties in aansprekende videoclips aan kandidaten te presenteren, maar ook om hun reacties, die zij volkomen vrij mogen geven, via een webcam te registreren. Weliswaar kost het afzien van een meerkeuze-antwoordformat meer scoringstijd, maar daar staat winst tegenover wat betreft *face validity* en waardering door kandidaten.

ADAPTIEVE TESTS

De papier-en-potloodinstrumenten waarmee psychologen over het algemeen werken, hebben een vaste lengte en worden door de cliënt van het eerste tot en met het laatste item ingevuld, met uitzondering van bepaalde mondelinge tests die instap- en afbreekregels kennen. Een gefixeerde lineaire test vormt echter vaak een onnodige belasting voor de cliënt: het zijn veel items en bij capaciteitentests zijn de eerste items kinderachtig makkelijk en de laatste frustrerend moeilijk. En bij een persoonlijkheidsvragenlijst kan een cliënt die op zes vragen van een extraversie-schaal 'nee!' heeft geantwoord geïrriteerd raken als er daarna nog een aantal vragen over hetzelfde onderwerp volgt –vragen waarop de psycholoog het antwoord wel zou kunnen raden. Een adaptieve test is hiervoor de oplossing, maar dat kan werkelijk alleen per computer.

Bij een gecomputeriseerde adaptieve test (CAT) wordt na ieder antwoord de score opnieuw geschat en wordt uit de vragenbank een vervolgvraag uitgezocht die de meeste informatie oplevert. De CAT stopt als de vereiste mate van nauwkeurigheid bereikt is. Een CAT is niet alleen afgestemd op de individuele cliënt, maar vooral ook flink korter dan een traditionele test. Zo simpel echter als het principe is, zo complex is de uitvoering. Hoewel het idee al minstens vijftig jaar oud is (Lord, 1968), begint de toepassing in Nederland nu pas op gang te komen en dan nog mondjesmaat. Twee factoren zijn voor het volwassen worden van de CAT een noodzakelijke impuls geweest: de ontwikkeling van de Item Respons Theorie (IRT) en de beschikbaarheid van krachtige computers.

Generaties psychologen zijn groot geworden met de klassieke testtheorie (Gulliksen, 1950), een theoretisch bouwwerk dat op drie eenvoudige uitgangspunten is gebaseerd: een testscore bestaat uit een betrouwbare score plus een meetfout, de gemiddelde meetfout in een populatie van personen is 0 en de betrouwbare score en meetfouten zijn niet gecorreleerd. De Item Respons Theorie – zie voor een introductie Furr & Bacharach (2014); zie ook kader – slaat een geheel andere weg in: IRT beschrijft de reactie van een persoon op een item. Anders gezegd: IRT beschrijft de kans dat de persoon met een bepaalde mate van de te meten eigenschap een vaardigheidsitem correct zal beantwoorden of zal instemmen met een attitude-item, afhankelijk van de kenmerken van dat item. In wezen is IRT dus een theorie over *gedrag*, wat impliceert dat van IRT-modellen – er zijn er vele – kan worden getoetst of ze een adequate beschrijving van de testdata zijn.

IRT-psychometrie biedt enorme kansen aan de psychodiagnostiek. Dat wil echter niet zeggen dat er binnen afzienbare tijd afscheid genomen gaat worden van de klassieke testtheorie en daarop gebaseerde tests. IRT-werk vraagt een aanzienlijke ontwikkelingsinspanning (grote kalibratiesteekproeven, grote itembanken en complexe software), dus we zullen nog vele jaren met klassiek geconstrueerde tests moeten werken. Ondanks bepaalde beperkingen is daar gelukkig niet zoveel mis mee.

TOEPASSINGEN VAN CAT

Adaptieve tests maken zoals gezegd langzamerhand hun entrée. Zo is de *ROUTE 8* van A-VISION een erkende eindtoets basisonderwijs en wordt de *Wiscat-pabo* van Cito ingezet om te toetsen of de rekenvaardigheid van pabo-studenten aan de landelijk vastgestelde norm voldoet. Ook HRM-adviesbureaus achten de investering in adaptieve instrumenten

ITEM RESPONS THEORIE

Het meest bekende IRT-model is het drie-parameter logistische model:

$$P(Y_{ij}=1 \mid \theta_i) = c_j + (1 - c_j)\ \frac{e^{(a_j(\theta_i - b_j))}}{1 + e^{(a_j(\theta_i - b_j))}}$$

In termen van een capaciteitentest geformuleerd beschrijft deze formule (waarin e het grondtal van de natuurlijke logaritme is) de kans dat persoon i met vaardigheid θ_i bij test Y een correct antwoord zal geven op item j, dat gekenmerkt wordt door discriminerend vermogen a_j, moeilijkheid b_j en gokkans c_j. De formule maakt het wezenlijke verschil met de uitgangspunten van de klassieke testtheorie zichtbaar: $\theta_i - b_j$ staat voor de confrontatie tussen de vaardigheid van de persoon en de moeilijkheid van het item.

steeds vaker de moeite waard, zoals Ixly op het gebied van intelligentie en persoonlijkheid.

De potentie van CAT's om efficiënte én nauwkeurige inschattingen te maken, blijft ook op gz-terrein niet onopgemerkt. Als huisarts en praktijkondersteuner vlot een inzicht willen krijgen in klachten (angst, depressie, stress, positieve en negatieve symptomen van psychose) en krachten (vriendschap, emotionele steun) van een patiënt, is een CAT het aangewezen hulpmiddel bij de triage. Van Bebber (2018) laat zien dat zo'n instrument kan bijdragen aan *personalized medicine*: tijdig psychosesymptomen herkennen, antipsychotica-dosering verfijnen, rekening houden met negatieve psychose-symptomen, terugval beperken en herstel van functioneren monitoren. Overigens wordt op basis van de CAT de ernst van de klachten iets lichter ingeschat dan wanneer de praktijkondersteuner zelf een inschatting maakt en wordt na raadpleging van het scoreprofiel over het algemeen een minder zwaar zorgniveau geadviseerd.

Paap et al. (2018) laten een andere overtuigende meerwaarde zien. Veelal zijn schalen van een vragenlijst gecorreleerd. Toch telt in een traditionele vragenlijst een bepaald item doorgaans slechts mee voor de ene schaal waar het toe gerekend wordt. Gezien de correlaties tussen de dimensies kan IRT echter gebruik maken van het feit dat items van de ene dimensie ook informatie over de score op de andere dimensies bevatten. De auteurs voerden een simulatie uit gebaseerd op typische gz-data, namelijk vier sterk gecorreleerde ($r > .75$) dimensies en een meerpunts-antwoordschaal. De dimensies waren: vermoeidheid, COPD-klachten[2], algemeen fysiek functioneren en deelname aan sociale activiteiten. De conclusie: één multidimensionele CAT is korter en nauwkeuriger dan vier aparte unidimensionele CAT's. Om maar te zwijgen van de efficiëntie van een traditionele vragenlijst met vier schalen.

Een belangrijk verschil tussen de klassieke testtheorie en IRT is dat de klassieke testtheorie slechts een globale uitspraak over de betrouwbaarheid van een test toelaat, namelijk de doorgaans gerapporteerde coëfficiënt alfa van Cronbach. IRT daarentegen biedt de mogelijkheid de meetnauwkeurigheid op verschillende scoreniveaus te bepalen. Dit is een belangrijk praktisch voordeel, bijvoorbeeld als een test de pretentie heeft dat een bepaalde score de optimale grenswaarde is om te concluderen dat de cliënt meer bij een klinische groep dan bij een groep 'normale' cliënten past. Men mag dan verlangen dat de test juist in de omgeving van die grenswaarde een optimale nauwkeurigheid heeft. Een CAT kan specifiek daarop worden gericht.

CAT-BELEVING

Het is maar de vraag of een CAT, die immers gericht is op optimale snelheid, efficiëntie en nauwkeurigheid, ook door cliënten als optimaal wordt ervaren. Als een kind een toets moet maken, dan is het wenselijk dat die toets start met een aantal niet te moeilijke opgaven om 'erin te komen' en dat het bij de volgende vragen niet een kans van 50% op een juist antwoord heeft, wat meet-technisch optimaal zou zijn, maar bijvoorbeeld van 70%: aangenaam uitdagend.

Eggen & Verschoor (2006) demonstreren een alternatieve manier om items te selecteren die hiermee rekening houdt zonder afbreuk te doen aan de nauwkeurigheid van de meting. Als bij een examen na tien vragen de vooraf ingestelde nauwkeurigheid van de vaardigheidsschatting zou zijn bereikt, kan men het als heel onbevredigend ervaren gezakt te zijn op een toetsje van tien vragen en weinig kans te hebben gehad om te laten zien wat men kan. Evenmin geeft het voldoening om te slagen op een toetsje van tien vragen, want 'Wat stelt dat nou helemaal voor?' In het laatste geval is er dus duidelijk spanning tussen de optimale efficiëntie en het door de kandidaat en diens omgeving ervaren civiel effect.

Een ander nadeel kan zijn dat sterke kandidaten een ander soort vragen kunnen krijgen dan minder sterke doordat de CAT immers vragen selecteert die het best passen bij het vaardigheidsniveau van de kandidaat. Dit probleem vraagt om bewaking door een toets-matrijs die ervoor zorgt dat bij alle kandidaten over alle deelonderwerpen in een juiste verhouding vragen worden aangeboden. Als het domein lezen bestaat uit de sub-domeinen 'techniek & woordenschat', 'begrijpen', 'interpreteren', 'evalueren', 'samenvatten' en 'opzoeken', dan moet het immers niet zo zijn dat aan zwakkere kandidaten geen enkele vraag die bijvoorbeeld 'evalueren' meet, wordt aangeboden.

IRT- CAT-techniek biedt dus belangrijke voordelen, maar het is van groot belang dat er bij de implementatie van een CAT ook gedacht wordt vanuit de beleving van een cliënt.

SOCIALE WENSELIJKHEID

Als vragenlijsten ergens last van hebben, is het wel van verschijnselen als antwoordtendenties, sociaal wenselijk antwoorden, *faking* of *malingering*. IRT biedt hier aanzien-

2 *Chronic Obstructive Pulmonary Disease* (chronische obstructieve longziekte).

lijk betere kansen dan alle andere pogingen om op dergelijke verschijnselen grip te krijgen. Allereerst kan er bij de constructie van worden geprofiteerd. Waar bijvoorbeeld de c-parameter bij een capaciteitentest of een vaardigheidstoets staat voor de kans dat een persoon met een minimale kennis het item toch goed zal beantwoorden, zou dit itemkenmerk bij een vragenlijst geduid kunnen worden als gevoeligheid voor sociaal wenselijke beantwoording.

Maar er zijn méér innovaties mogelijk. Doordat, zoals gezegd, IRT een theorie over testgedrag is, kan niet alleen worden onderzocht of de items zich volgens het betreffende model gedragen, maar ook of de *personen* dat doen. Daardoor kunnen bij een door IRT gestuurde testafname *person-fit* statistieken worden berekend die kunnen worden gebruikt om respondenten met afwijkende antwoordpatronen te identificeren.

METEN VIA TEKSTEN

Computers zijn bij uitstek in staat om grote hoeveelheden gegevens over personen op te slaan en te verwerken, niet alleen van gestructureerde data, zoals antwoorden op een vragenlijst, maar ook van ongestructureerde. Zo worden teksten steeds vaker gebruikt als bron van informatie bij het stellen van een diagnose. He (2013) demonstreerde dat men met *text mining* van korte verhalen waarin patiënten vertelden hoe het met ze ging, onderscheid kan maken tussen mensen met een hoog of een laag risico op het ontwikkelen van een posttraumatische stress-stoornis. De overeenkomst met de diagnose van een psychiater was aanzienlijk (82%). Extra effectief en efficiënt bleek een Bayesiaanse aanpak: de schatting uit de *text mining* wordt gebruikt om een prior (een eerste schatting van de kansverdeling) op te stellen van de latente trek, waarna een IRT-vragenlijst op deze startinformatie voortborduurt. Door de combinatie kan de vragenlijst korter en dus minder belastend voor de cliënt zijn, terwijl de nauwkeurigheid van de detectie zelfs toeneemt.

Analyse op basis van woordtype (*Linguistic Inquiry and Word Count*) is een andere benadering. Bij de analyse van e-mails waarmee cliënten zich voorstelden aan het begin van het doorlopen van een zelfhulpboek, bleek uitval positief samen te hangen met het aantal negatieve emotiewoorden en lichaamsgerichte woorden, en negatief met het totaal aantal woorden, het aantal zintuiglijke woorden en het aantal prestatiegerichte woorden. Uitval bij de cursus kon met een specificiteit van 91% voorspeld worden en de sensitiviteit voor het afmaken van de interventie bedroeg 84% (Westerhof, 2018).

TESTAFNAME ZONDER TOEZIEND OOG

Als een persoon een test thuis maakt zonder dat een testleider daarop toeziet (*unproctored*), dan weet men in principe niet of het echt de cliënt is die achter het scherm zit, of er hulpmiddelen worden gebruikt en of er andere personen aanwezig zijn die zich met de beantwoording bemoeien. Het is weliswaar een groot voordeel dat een testafname overal kan plaatsvinden, maar zeker bij capaciteitentests moet men toch oplossingen bedenken (zoals wachtwoord, legitimatie, webcam, gezichtsherkenning, *screen capture*) om de betrouwbaarheid van de testafname te controleren. In een getrapte selectiesituatie kan men eerlijk gedrag bevorderen door aan te kondigen dat bij kandidaten die doorgaan naar de volgende ronde een parallelversie van de test onder toezicht zal worden afgenomen.

De COTAN heeft in de Algemene Standaard Testgebruik (Nederlands Instituut van Psychologen, 2017) het uitgangspunt geformuleerd dat bij niet-cognitieve tests de invloed van een *unproctored* of *proctored* afname op factorstructuur, criteriumvaliditeit en gemiddelden als verwaarloosbaar kan worden beschouwd, maar dat bij capaciteitentests grote voorzichtigheid geboden is. De basisregel is dat de normeringssituatie en de beoogde afnameconditie overeen moeten komen en dat equivalentie aannemelijk moet worden gemaakt (COTAN, 2018).

Kwaliteitscriteria voor psychologische tests zijn onverkort ook op internet-tests van toepassing

COTAN-CRITERIA

Het COTAN-beoordelingssysteem spreekt zich uit over de uitgangspunten van de testconstructie, de kwaliteit van het testmateriaal, de kwaliteit van de handleiding, de normen, de betrouwbaarheid, de begripsvaliditeit en de criteriumvaliditeit. Extra aandacht wordt besteed aan de vraag of het instrument eerlijk is jegens relevante maatschappelijke subgroepen – bijvoorbeeld wat betreft sekse, leeftijd of culturele achtergrond (fairness). Al deze kwaliteitscriteria voor psychologische tests zijn onverkort ook op internet-tests van toepassing. In de huidige editie van het beoordelingssysteem

De psycholoog moet aan de cliënt kunnen uitleggen hoe in een geautomatiseerde rapportage het systeem tot conclusies komt

(Evers et al., 2009) wordt aan computertests al speciale aandacht besteed; het is zeker dat dit in de herziening, die naar verwachting medio 2020 zijn beslag zal krijgen, nog nadrukkelijker het geval zal zijn.

Hoe gemakkelijk maakt internet het niet om normgegevens te verzamelen door respondenten door middel van linkjes uit te nodigen! Dat doet echter niets af aan de eis dat normgroepen representatief voor een welomschreven doelgroep moeten zijn, en dat dit aangetoond moet kunnen worden. Ook moet de verzamelsituatie overeenkomen met het doel waarvoor de test gebruikt gaat worden. Als men bijvoorbeeld personen heeft gevraagd een capaciteitentest op vrijwillige basis en zonder consequenties thuis in te vullen, terwijl het instrument bedoeld is om in een selectiesituatie te worden afgenomen, dan komen deze situaties duidelijk niet overeen.

Net zoals bij een papier-en-potloodtest verlangt het COTAN-beoordelingssysteem dat een internet-test goed beveiligd is. Dit betreft de toegang tot de test (gebruikersnaam, wachtwoord, legitimatie), een https-verbinding tijdens de afname, een bescherming tegen het bekend raken van items – bij adaptieve tests zijn methodes à la Sympson & Hetter (1985) effectief tegen het al te vaak aanbieden van bepaalde items –, de toegang tot een server waarop data zijn opgeslagen, en de ontkoppeling tussen testdata en persoonsgegevens.

WAT ZEGT DE BEROEPSCODE?

In de Beroepsethiek van het NIP (Nederlands Instituut van Psychologen, 2015) komt het woord internet-tests niet voor, maar deze gedragsregels laten geen misverstand bestaan over de zorgvuldigheid en deskundigheid die vereist is bij het verrichten van psychodiagnostisch onderzoek. Bovendien spreekt artikel 17 zich expliciet uit over het gebruik van relatief nieuwe methoden: 'Bij het toepassen van nieuwe methoden of het betreden van nieuwe toepassingsgebieden gaan psychologen zorgvuldig en voorzichtig te werk.'

Ook bij internet-tests dient de psycholoog volgens artikel 63 ervoor te zorgen dat de cliënt vrijelijk, specifiek, geïnformeerd en ondubbelzinnig toestemming geeft en dat er afspraken worden gemaakt over doel en werkwijze, het soort gegevens dat wordt verzameld, en over de wijze waarop en hoe lang deze gegevens worden bewaard. Dit geldt uiteraard ook onverkort voor het recht op inzage en afschrift van het dossier en op correctie en blokkering van de rapportage. Ook waar het de test zelf betreft heeft de psycholoog ten volle de regie over én de professionele verantwoordelijkheid voor de afname, scoring, interpretatie en rapportage.

De psycholoog moet aan de cliënt kunnen uitleggen hoe in een geautomatiseerde rapportage het systeem tot conclusies komt. Stel dat in een selectiepraktijk de kandidaat over een rapport, waarin persoonlijkheidsscores door middel van rekenregels vertaald worden in competentiescores, aan de psycholoog vraagt waar een bepaalde conclusie op gebaseerd is. Het antwoord 'Ik kan dat niet precies uitleggen, dat heeft het systeem zo berekend' is onacceptabel. De psycholoog dient immers zelf te begrijpen en daardoor te kunnen uitleggen hoe het systeem werkt. Evenmin is het acceptabel als de psycholoog geen toegang tot de resultaten van een internet-test kan geven. Weliswaar hoeft aan een cliënt blijkens een uitspraak van de Autoriteit Persoonsgegevens geen inzage op itemniveau te worden verleend, maar een begeleide inzage in ruwe en genormeerde scores moet wel degelijk kunnen worden geboden en het systeem behoort dat mogelijk te maken.

Over de neurale netwerken waarop *text mining* is gebaseerd, is het geven van uitleg overigens nagenoeg onmogelijk. Dergelijke zichzelf trainende algoritmen hebben een sterk black box-karakter. Maken zulke expertsystemen de psycholoog dan overbodig? Dat is te sterk gezegd – de psycholoog is méér dan zijn instrumentarium. Maar dat er een spanningsveld is tussen commerciële belangen van een uitgever en de transparantie voor de cliënt die psychologen graag willen betrachten, is duidelijk. 'Computer says no' is niet een tekst die we als psycholoog gemakkelijk over onze lippen krijgen. Maar als *big data* gedrag beter kunnen voorspellen dan psychologische tests, waarom zouden we dergelijke innovatieve methoden dan niet omarmen en in onze diagnostiek integreren?

Tot slot het recht om vergeten te worden: de Wet op de geneeskundige behandelingsovereenkomst (WGBO) en de Beroepscode specificeren bewaartermijnen van het psychologische dossier, maar na het verstrijken van zo'n termijn of als

de cliënt eerder een verzoek tot verwijdering doet, dient de psycholoog daarvoor te zorgen. Er kunnen echter omstandigheden zijn dat de psycholoog niet aan zo'n verzoek kan voldoen, bijvoorbeeld omdat er een klachtprocedure loopt waarbij de psycholoog het dossier nodig heeft om verweer te kunnen voeren. Ook zou een zorgverzekeraar na het verstrijken van een termijn nog een controle op rechtmatigheid kunnen willen uitvoeren. In dat geval zou een inhoudelijk maximaal gestripte versie van het dossier een oplossing kunnen zijn, mits zorgvuldig beargumenteerd en gedocumenteerd.

Kortom: de psycholoog is ook bij internet-testen en dossiervoering in de *cloud* onverkort verantwoordelijk voor alles wat er in de professionele relatie met de cliënt gebeurt.

OVER DE AUTEUR

Wouter Lucassen is lid van de Commissie Testaangelegenheden Nederland (COTAN) en oud-lid van het College van Beroep. Dit artikel is een uitwerking van zijn inleiding op de Landelijke Supervisorendag Basisaantekening Psychodiagnostiek van 7 december 2018. Email: wluc@xs4all.nl.

Summary

AN UPDATE ON INTERNET TESTING
W. LUCASSEN

Taking full advantage of computer power undoubtedly adds to the already known advantages of internet testing, as multimedia inventories in child psychodiagnostics or 'serious games' in personnel selection show. The benefits of computerized adaptive tests for both client and psychologist are discussed and illustrated. A plea is made for a more intensive use of item response theory to tackle problems of response tendencies, social desirability, faking and malingering. Some recent developments, such as text mining for the diagnosis of posttraumatic stress disorder, are mentioned. Addressing the unproctored use of psychological tests, the Dutch Committee on Tests and Testing has issued an addendum to its test quality review system. Special attention is paid to ethical questions with regard to internet testing.

Literatuur

Bureau ICE (2018). *IEP Eindtoets. Verantwoording afname 2018 en normeringsonderzoek 2019*. Culemborg: Bureau ICE.

COTAN (2018). *Herzien addendum unproctored gegevensverzameling*. Te downloaden op https://tinyurl.com/yxlcnpm2.

Eggen, T.J.H.M., & Verschoor, A.J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*, 379-393.

Evers, A., Lucassen, W., Meijer, R.R. & Sijtsma, K. (2009). *COTAN Beoordelingssysteem voor de kwaliteit van tests, geheel herziene versie*. Amsterdam: NIP/COTAN.

Furr, R.M. & Bacharach, V.R. (2014). *Psychometrics: An introduction*. Second Edition. Thousand Oaks, CA: Sage Publications.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

He, Q. (2013). *Text Mining and IRT for Psychiatric and Psychological Assessment*. Ph.D. thesis, University of Twente, Enschede.

Kato, P.M. & De Klerk, S. (2017). Serious Games for Assessment: Welcome to the Jungle. *Journal of Applied Testing Technology, 18*(S1), 1-6.

Kuijpers, R.C.W.M., Otten, R.,Vermulst, A.A., Pez, O., Bitfoi, A. et al. (2014). Cross-country construct validity of the 'Dominic Interactive'. *The European Journal of Public Health, 24*(S2).

Lord, F.M. (1968). *Some Test Theory for Tailored Testing*. Princeton, NJ: Educational Testing Service.

Mulder, J.L., Dekker, P.H. & Dekker, R. (2006). *Woord-Fluency Test / Figuur-Fluency Test, Handleiding*. Leiden: PITS.

Nederlands Instituut van Psychologen (2015). *Beroepscode voor psychologen 2015*. Utrecht: NIP.

Nederlands Instituut van Psychologen (2017). *Algemene Standaard Testgebruik*. Utrecht: NIP.

Oostrom, J.K., Born, M.P., Serlie, A.W. & Van der Molen, H.T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology, 19*, 532-550.

Paap, M.C.S., Kroeze, K.A., Glas, C.A.W., Terwee, C.B., Van der Palen, J. & Veldkamp, B.P. (2018). Measuring Patient-Reported Outcomes Adaptively: Multidimensionality Matters! *Applied Psychological Measurement, 42*(5), 327-342.

Schuhfried, G. (2012). *Vienna Test System Traffic*. Mödling: Schuhfried.

Sympson, J.B. & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association*, 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Van Baar, A.L., Steenis, L.J.P. & Verhoeven, M. (2014). *Bayley-III-NL Afnamehandleiding*. Amsterdam: Pearson Assessment and Information.

Van Bebber, J. (2018). *Computerized adaptive testing in primary care: CATja*. Proefschrift Rijksuniversiteit Groningen.

Westerhof, G. (2018). *Verhaal - Digitaal. Narratieve technologie voor gezondheid en zorg*. Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar Narratieve Psychologie en Technologie, Universiteit Twente.

Wilhelm, P., Swaab, H., Serlier-van den Bergh, A.H.M.L. & Van der Heijden, K.B. (2010). *Rey Visual Design Learning Test*. Houten: Bohn Stafleu van Loghum.