Wanneer is kort te kort?

Over de geschiktheid van korte tests om selectiebeslissingen te nemen*

*Peter M. Kruyen***

Selectiepsychologen maken veelvuldig gebruik van psychologische tests om geschikte kandidaten te vinden. Uit de klassieke testtheorie volgt dat het verstandig is om gebruik te maken van lange tests om beslissingen te nemen over individuele kandidaten. Immers, hoe meer items men afneemt, hoe hoger de testbetrouwbaarheid, en hoe kleiner de invloed van toevallige meetfouten op het eindoordeel. In de praktijk bestaat er echter een grote behoefte aan korte tests. Zijn er toch situaties denkbaar waarin tests bestaande uit een klein aantal items geschikt zijn om selectiebeslissingen te nemen? Uitgaande van een hoge validiteit van korte testversies toonden Kruyen, Emons en Sijtsma (2012) aan dat het aantal benodigde items om selectiebeslissingen te nemen niet zozeer afhankelijk is van de betrouwbaarheid maar van de gewenste zekerheid op juiste beslissingen alsook van de base rate. In deze bijdrage leg ik uit hoe men de effecten van testverkorting op het risico op verkeerde selectiebeslissingen kan kwantificeren. Aan de hand van aanvullend simulatieonderzoek generaliseer ik daarnaast de conclusies van Kruyen et al. (2012) naar andere selectiescenario's, waaronder de inzet van adaptieve testen.

1 Inleiding

Willen organisaties optimaal kunnen presteren, dan is het van cruciaal belang dat zij hun werving- en selectieprocedures zorgvuldig vormgeven (Cook, 2009; Hunter & Schmidt, 1982). Het aannemen van ongeschikte kandidaten is kostbaar, evenals het afwijzen van geschikte kandidaten. Ook voor kandidaten zelf is het belangrijk dat zij voor functies worden geselecteerd die zij succesvol kunnen uitvoeren. Wanneer je geselecteerd wordt voor een functie waarvoor je ongeschikt bent, dan kan dat leiden tot veel stress en een negatief zelfbeeld (Smith & Smith, 2005, p. 1). Word je ten onrechte afgewezen dan is dat natuurlijk oneerlijk (International Test Commission, 2000).

* Ik bedank dr. Wilco Emons en prof. dr. Klaas Sijtsma voor hun bijdragen aan mijn proefschrift waarop deze bijdrage is gebaseerd. Dank ook aan prof. dr. Beate van der Heijden, dr. Wilco Emons, Esther de Weert MSc voor hun waardevolle commentaar op eerdere versies van deze tekst.

** Peter M. Kruyen is werkzaam bij het Institute for Management Research, Radboud University, Nijmegen. Correspondentieadres: Radboud University, Institute for Management Research, Postbus 9102, 6500 HC Nijmegen, tel. 024-3611934. E-mail: p.m.kruyen@fm.ru.nl.

De geschiktheid van sollicitanten werd vroeger vooral beoordeeld op basis van opleiding en relevante werkervaring. Tegenwoordig spelen psychologische eigenschappen een steeds belangrijkere rol (Drabbe, Drost, Klehe, Van Vianen & Boendermaker, 2008). Van de vele sollicitanten wil men die kandidaat selecteren die het beste past binnen het team, die gemotiveerd is, sociaal vaardig, communicatief en probleemoplossend, en die over de nodige cognitieve capaciteiten beschikt (Drabbe et al., 2008; Rothstein & Goffin, 2006; Tillema, 1998). Omdat psychologische eigenschappen niet direct waarneembaar zijn, maken selectiepsychologen veelvuldig gebruik van tests en vragenlijsten om kandidaten op deze eigenschappen te beoordelen (zie bijv. De Vries, De Vries, Born & Van den Berg, 2014).

Traditioneel bevatten veel psychologische tests een grote hoeveelheid items. Zo bestaat de test voor Niet Verbale Abstractie uit 40 items (Drenth, 1965), worden in de *Revised NEO Personality Inventory* de vijf persoonlijkheidskenmerken elk met 48 items gemeten (Costa & McCrae, 1992) en bestaat een recent ontwikkelde ecologische persoonlijkheidsvragenlijst (PPQ) uit 62 items (Butter, 2014). Ten eerste is de opname van veel items te wijten aan de theoretische breedte van de te meten constructen. Immers, hoe breder het veronderstelde inhoudsdomein, hoe meer items men nodig heeft om alle subdimensies adequaat te kunnen afdekken en dus begripsvaliditeit te waarborgen. Ten tweede wordt het grote aantal items gemotiveerd door inzichten uit de klassieke testtheorie (bijv. Brown, 1910; Spearman, 1910). Uit de klassieke literatuur kan men afleiden dat hoe meer items men afneemt om een specifiek afgebakend construct te meten, hoe hoger de testbetrouwbaarheid is en dus hoe kleiner de mogelijke invloed van toevallige meetfouten op het te vellen selectieoordeel.

In deze bijdrage bespreek ik de relatie tussen het aantal items, de testbetrouwbaarheid en het risico op verkeerde beslissingen over individuele kandidaten. Hoewel lange, betrouwbare tests te prefereren zijn vanuit het oogpunt van de klassieke testtheorie, bestaat er namelijk in de selectiepraktijk een grote behoefte aan korte tests (bijv. bestaande uit minder dan 20 items) om kandidaten te beoordelen (Burisch, 1997; Netemeyer, Pulling & Bearden, 2003). Korte tests zijn immers minder tijdrovend dan lange tests. Een ander bezwaar tegen het gebruik van een langere test om een specifiek construct te meten is dat items vaak veel op elkaar lijken. Dit levert het gevaar op dat kandidaten de selectieprocedure als langdradig en irrelevant ervaren, hetgeen negatief afstraalt op de selecterende organisatie.

De vraag rijst derhalve of er toch situaties denkbaar zijn waarin tests bestaande uit een klein aantal items geschikt zijn om selectiebeslissingen te nemen. Of moet men het gebruik van korte tests in de context van personeelsselectie categorisch afwijzen? Uitgaande van een hoge validiteit van de kortere testversies, concludeerden Kruyen et al. (2012) op basis van gesimuleerde data dat het aantal benodigde items niet zozeer afhankelijk is van de betrouwbaarheid, maar van de gewenste zekerheid op juiste beslissingen alsook van de *base rate* (d.w.z. de proportie geschikte kandidaten in de totale groep van kandidaten). In sommige omstandigheden geven vijf items met een voldoende betrouwbaarheid al genoeg zekerheid, maar in andere gevallen volstaat zelfs een 40-item test met hoge betrouwbaarheid niet.

In deze bijdrage licht ik eerst toe hoe men de zekerheid op juiste selectiebeslissingen kan kwantificeren. Vervolgens vat ik de aanpak en bevindingen van Kruyen et al. (2012) samen. Daarna bespreek ik de resultaten van enkele nieuwe simulatiestudies. Deze simulatiestudies geven inzicht in de mogelijke risico's van korte tests om beslissingen te nemen over kandidaten in een drietal nog niet eerder onderzochte selectiesituaties, te weten: (1) wanneer men twee kandidaten met elkaar wil vergelijken rekening houdend met de meetprecisie, (2) wanneer men kandidaten wil classificeren met korte, maar zeer betrouwbare tests, en (3) wanneer men gebruik wil maken van een adaptieve test. Aan het einde van dit artikel doe ik enkele aanbevelingen aan selectiepsychologen over het gebruik van (korte) tests in de context van personeelsselectie.

2 Drie manieren om de zekerheid op juiste selectiebeslissingen te kwantificeren

2.1 Meetprecisie

In sommige selectiesituaties hecht men er grote waarde aan dat voor iedere kandidaat de juiste beslissing wordt genomen. Om in de taal van de testtheorie te spreken, men vindt het belangrijk dat het oordeel op basis van geobserveerde testscores X_+ overeenkomt met het oordeel dat men zou vellen als men toegang zou hebben tot de betrouwbare scores T. Selectiepsychologen maken bijvoorbeeld een fout wanneer zij stellen dat een specifieke kandidaat redelijk extrovert is afgaand op diens geobserveerde X_+ testscore terwijl zij op basis van diens betrouwbare score T hadden moeten concluderen dat de kandidaat juist introvert is.

In beginsel wordt de mate van zekerheid dat men de juiste beslissing neemt over individuele kandidaten bepaald door de mate waarin geobserveerde testscores gevoelig zijn voor toevallige omstandigheden (d.w.z. toevallige meetfouten). Deze gevoeligheid komt tot uitdrukking in de meetprecisie. Hoe kunnen we meetprecisie kwantificeren? In de praktijk gebruikt men het betrouwbaarheidsinterval *CI* om een schatting te maken van de precisie waarmee individuele scores worden gemeten (Sijtsma, 2009). Het *CI* is afhankelijk van de betrouwbaarheid $r_{XX'}$, de standaarddeviatie van de geobserveerde scoreverdeling en het gewenste significantieniveau (Harvill, 1991).

Wanneer men het belangrijk vindt dat voor iedere kandidaat de juiste beslissing wordt genomen, dan laat men de beslissing afhangen van het *CI*. Wil men bijvoorbeeld kandidaten indelen in twee categorieën (geschikt of ongeschikt) en daarbij rekening houden met de meetprecisie, dan deelt men alleen een kandidaat in één van de twee categorieën in wanneer de criteriumscore T_c niet in het *CI* ligt. Kandidaten wier *CI* onder T_c ligt, worden geclassificeerd als ongeschikt. Kandidaten wier *CI* boven T_c ligt, worden geclassificeerd als geschikt. Omvat het *CI* de T_c, dan dient men meer informatie te verzamelen over de betreffende kandidaat voordat men een selectiebeslissing kan nemen.

Opgemerkt moet worden dat veel selectiepsychologen hun tests niet op basis van de meetprecisie kiezen, maar op basis van de gerapporteerde betrouwbaarheid

$r_{XX'}$. Zij hanteren bijvoorbeeld als richtlijn dat tests geschikt zijn in selectieprocedures wanneer $r_{XX'} \geq .90$. Echter, de betrouwbaarheid zegt wel iets over de mate waarin de rangorde van kandidaten consistent blijft over herhaalde testafnames, maar zegt niet direct iets over de invloed van meetfouten op individuele scores (Mellenbergh, 1996; Sijtsma, 2009). Stel, men neemt tweemaal een test af bij de twee kandidaten Tom en Mirjam. Stel nu ook dat op beide meetmomenten Tom hoger scoort dan Mirjam, maar dat beide kandidaten 30 punten meer scoren bij de tweede testafname. In dit geval is de betrouwbaarheid nog steeds zeer hoog, maar is de meetprecisie twijfelachtig.

2.2 Classificatieconsistentie

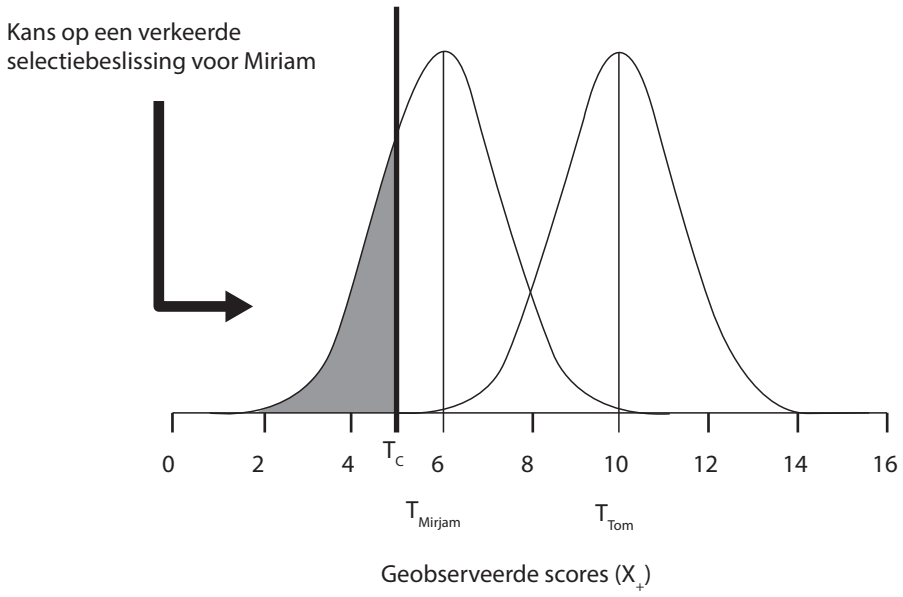
Sommige selecterende organisaties vinden het niet zo erg als voor individuele kandidaten de verkeerde beslissing wordt genomen zolang voor het leeuwendeel van de kandidaten de beslissing maar correct is. Men neemt in dit geval geen beslissing op basis van het *CI* maar wel op basis van de geobserveerde testscores X_+. Om de zekerheid op juiste beslissingen in deze situatie te kwantificeren maak ik gebruik van de zogenoemde classificatieconsistentie (CC, zie ook Emons, Sijtsma & Meijer, 2007).

Figuur 1 laat de hypothetische scoreverdeling zien van Tom en Mirjam wanneer het mogelijk zou zijn geweest om beide kandidaten oneindig vaak te testen. Afgaand op de spreiding van beide scoreverdelingen zien we dat Toms competentie (of psychologische construct waarin men geïnteresseerd is) even precies wordt gemeten als de competentie (of het psychologische construct waarin men geïnteresseerd is) van Mirjam. Stel dat men deze test zou gebruiken om kandidaten te classificeren als geschikt of ongeschikt zonder rekening te houden met de meetprecisie (d.w.z. kandidaten met een geobserveerde testscore X_+ gelijk of hoger dan T_c worden geselecteerd, kandidaten die lager scoren dan T_c worden afgewezen), dan zou voor Mirjam in ongeveer 30% van de testherhalingen een verkeerde beslissing worden genomen. Voor Tom is het risico op een verkeerde beslissing vrijwel nihil ondanks de gelijke meetprecisie.

Er wordt verder een onderscheid gemaakt tussen twee soorten van classificatieconsistentie. De classificatieconsistentie CC+ is de proportie geschikte kandidaten die in ten minste 90% van de herhaalde testafnames consistent als geschikt wordt geclassificeerd, zoals bij Tom maar niet bij Mirjam het geval is. De classificatieconsistentie CC- is de proportie ongeschikte kandidaten die in ten minste 90% van de herhaalde testafnames consistent als ongeschikt wordt geclassificeerd.

2.3 Predictieve waarden

Soms vindt men het niet belangrijk dat er een verkeerde beslissing over individuele kandidaten wordt genomen zolang het risico dat men ongeschikte kandidaten toelaat maar zo klein mogelijk is. In dit geval kan de zekerheid op juiste beslissingen worden uitgedrukt in de positieve predictieve waarde (PPW), dit is de proportie geschikte kandidaten in de totale groep van geselecteerde kandidaten. Daarnaast kan men het belangrijk vinden dat zoveel mogelijk ongeschikte kandidaten daadwerkelijk afgewezen worden, maar vindt men het minder belangrijk



Figuur 1 Classificatieconsistentie voor twee individuele sollicitanten

dat voor individuele kandidaten een kans bestaat om ten onrechte toch geselecteerd te worden. In dit geval kan de zekerheid op juiste beslissingen worden uitgedrukt in de negatieve predictieve waarde (NPW), dit is de proportie ongeschikte kandidaten in de totale groep van afgewezen kandidaten.

3 Aanpak en belangrijkste bevindingen van Kruyen et al. (2012)

Kruyen et al. (2012) onderzochten de effecten van testverkorting op de CC en predictieve waarden aan de hand van gesimuleerde data in een drietal selectiescenario's: (1) selectie van kandidaten die gelijk of hoger scoren dan de criteriumscore op één of meerdere tests (d.w.z. *cut score selectie*), (2) selectie van kandidaten die het hoogst scoren op één of meerdere tests (d.w.z. *top down selectie*) en (3) selectie van kandidaten wiens scoreprofiel het meeste lijkt op het scoreprofiel van de ideale kandidaat. In deze paragraaf vat ik hun aanpak en bevindingen kort samen.

Concreet simuleerden Kruyen et al. (2012) testdata voor 100.000 hypothetische kandidaten die zij duizend maal dezelfde 40-item test voorlegden wanneer een beslissing werd genomen op basis van één test of duizend maal dezelfde vijf tests wanneer een beslissing werd genomen op basis van meerdere tests. Voor de selectie van kandidaten op basis scoreprofielen bestudeerden zij de situatie waarin de ideale kandidaat gemiddeld scoorde op elk van de vijf 40-item tests. Kandidaten wiens scoreprofiel het meest overeenkwam met het profiel van de ideale kandidaat, werden geselecteerd. In elk van deze drie onderzochte scenario's werd gebruikgemaakt van 40-item tests met een hoge betrouwbaarheid ($r_{XX'}$ = .90).

Vervolgens onderzochten zij de effecten op de CC en predictieve waarden wanneer de 40-item tests werden ingekort tot respectievelijk 20, 15, 10 en 5 items, waarbij de betrouwbaarheid van de 5-item test nog steeds als voldoende beschouwd kon worden ($r_{XX'}$ = .70). Kruyen et al. (2012) namen aan dat de korte testversies even valide waren als de langere versies.

Op basis van hun simulatiestudie stelden Kruyen et al. (2012) dat een voldoende hoge testbetrouwbaarheid niet garandeert dat het risico op selectiefouten wordt geminimaliseerd. Het aantal benodigde items blijkt namelijk afhankelijk van de gewenste zekerheid op juiste beslissingen alsook van de *base rate* en verschilt daarbij voor de groep geschikte kandidaten en ongeschikte kandidaten. Zo heeft men aan vijf items met voldoende betrouwbaarheid reeds genoeg wanneer de *base rate* laag is, men vooral belang hecht aan een hoge NPW en men het niet zo erg vindt dat voor een groot deel van de geschikte kandidaten een aanzienlijk risico bestaat om ten onrechte te worden aangemerkt als ongeschikt. Tegelijkertijd concludeerden zij dat zelfs 40 betrouwbare items niet voldoende zijn bij een lage *base rate* wanneer men het liefst wil dat alle geschikte kandidaten ook daadwerkelijk geselecteerd worden. De resultaten in hun onderzoek waren vergelijkbaar voor tests bestaande uit *dichotome* items en tests bestaande uit *polytome* items en verschilden niet veel tussen het drietal onderzochte selectiescenario's (d.w.z. cut score selectie, top down selectie of met het scoreprofiel van de ideale kandidaat).

Dat de impact van testverkorting groter is voor CC+ en CC- dan voor NPW en PPW kan als volgt worden verklaard. Stel dat er drie personen solliciteren op een bepaalde functie en dat deze personen alle drie even geschikt zijn. Laten wij ook aannemen dat deze drie personen een risico hebben van 30% om – door toevallige meetfouten – lager dan T_c scoren. CC+ is dus gelijk aan 0 in dit voorbeeld. Wanneer nu twee van de drie kandidaten het criterium inderdaad niet halen, dan is de PPW nog steeds 1. Immers, het effect van testverkorting voor individuele kandidaten is sterker dan de impact van testverkorting op het groepsniveau vanwege het sterke effect van het verwijderen van items op de meetprecisie van individuele scores.

Ook is het begrijpelijk dat de effecten van testverkorting afhankelijk zijn van de *base rate* en daarmee verschillen voor geschikte en ongeschikte kandidaten. Hoe lager de *base rate* (oftewel hoe hoger criteriumscore T_c) hoe kleiner het aantal kandidaten dat feitelijk geschikt is. Echter, hoe lager de *base rate*, hoe groter het effect van testverkorting op het risico op een verkeerde beslissing voor de groep geschikte kandidaten zal zijn. Immers, bij een lage *base rate* heeft een relatief groot aantal geschikte kandidaten een betrouwbare score vlakbij de criteriumscore, zoals Mirjam in Figuur 1, en dus een grote kans om onterecht te worden afgewezen. Daarentegen zullen bij een lage *base rate* verreweg de meeste ongeschikte kandidaten een betrouwbare score hebben die ver van de criteriumscore ligt. De lagere meetprecisie als gevolg van testverkorting heeft in dit laatstgenoemde geval voor de meeste ongeschikte kandidaten geen nadelige gevolgen. Bij een hogere *base rate* (of een lagere criteriumscore T_c) treedt juist het tegenovergestelde effect op. In dat geval is het effect van testverkorting groter voor de groep ongeschikte kandidaten dan voor de groep geschikte kandidaten. Er zijn dan relatief veel ongeschikte kandidaten met een betrouwbare score in de buurt van de

criteriumscore, terwijl de meeste geschikte kandidaten een betrouwbare score hebben die ver van de criteriumscore vandaan ligt.

4 De geschiktheid van korte tests om selectiebeslissingen te nemen nader bestudeerd

Met behulp van dezelfde simulatieprocedures als beschreven in Kruyen et al. (2012) onderzocht ik voor deze bijdrage de geschiktheid van korte tests in drie belangrijke, nog niet eerder onderzochte, selectiesituaties, te weten: (1) wanneer men twee kandidaten met elkaar wil vergelijken rekening houdend met de meetprecisie; (2) wanneer men kandidaten wil classificeren als geschikt of ongeschikt met behulp van een zeer betrouwbare test (al dan niet rekening houdend met de meetprecisie) en (3) wanneer met kandidaten wil classificeren als geschikt of ongeschikt met behulp van een adaptieve test.
Kruyen et al. (2012) bestudeerden de effecten van testverkorting onder de conditie dat men een 5-item test gebruikte met een betrouwbaarheid van $r_{XX'}$ = .70. Men kan zich afvragen in hoeverre hun conclusies standhouden wanneer de betrouwbaarheid van de kortste test wordt opgeschroefd tot $r_{XX'}$ = .90. Anders ook dan Kruyen et al. (2012) presenteer ik de resultaten voor *polytome* items in plaats van de resultaten voor *dichotome* items omdat *polytome* items veelvuldig gebruikt worden in de context van personeelsselectie. Gekozen is voor items met een 5-puntsschaal, waarbij kandidaten items kunnen scoren op een schaal die loopt van 0 tot 4.

4.1 Het vergelijken van twee kandidaten rekening houdend met meetprecisie

- *Selectiesituatie.*

Stel dat een team selectiepsychologen een vergelijking wil maken tussen twee specifieke kandidaten en men daarbij rekening wil houden met de meetprecisie. In deze situatie wordt dus alleen een beslissing genomen wanneer de geobserveerde scores significant van elkaar verschillen, oftewel het *CI* van de verschilscores groter of kleiner dan 0 is. De complicatie in deze situatie is dat men rekening moet houden met meetfouten in beide geobserveerde scores en dus gebruik moet maken van de standaarddeviatie van de geobserveerde verschilscores in plaats van de standaarddeviatie van de geobserveerde scores om het *CI* te bepalen (Harvill, 1991).

- *Simulatieprocedure.*

Om het effect van testlengte in dit scenario te bepalen liet ik een miljoen paren van hypothetische kandidaten duizend maal dezelfde 40, 20, 15, 10 en 5-item testversie maken waarbij de betrouwbaarheid van de langste test gelijk was aan $r_{XX'}$ = .99 en de kortste test een betrouwbaarheid had van $r_{XX'}$ = .90. De langste testversie bestond uit items met een grote variëteit in populariteit om de kans te maximaliseren dat verschillen tussen de hypothetische kandidaten konden worden gedetecteerd. Voor de kortere testversies selecteerde ik de items zo dat de

grote spreiding in item-populariteit behouden bleef. In de gesimuleerde dataset rekende ik vervolgens voor iedere testversie voor ieder paar van kandidaten het 90% *CI* van de verschilscore uit.

- *Simulatieresultaten.*

In dit scenario zijn korte tests zeer problematisch. Stel dat het team van selectiepsychologen wil dat voor ieder paar van kandidaten in iedere (hypothetische) testafname een significant verschil kan worden vastgesteld, dan neemt de proportie van paren van kandidaten voor wie dat mogelijk is af van .64, tot .51, .43, .33, en zelfs .21 bij een reductie van testlengte van 40 tot 20, 15, 10 en 5 items. Wanneer men twee kandidaten met elkaar wil vergelijken en daarbij rekening wil houden met meetprecisie, is het dus aan te bevelen om gebruik te maken van lange tests met een zeer hoge betrouwbaarheid *of* te accepteren dat er een aanzienlijk risico bestaat dat men geen significant verschil tussen de twee kandidaten kan vaststellen en dus geen beslissing kan nemen.

4.2 Classificatie met behulp van korte tests met een zeer hoge betrouwbaarheid

- *Selectiesituatie.*

In dit scenario worden kandidaten die ten minste gelijk aan de criteriumscore T_c scoren geclassificeerd als geschikt, terwijl kandidaten die lager dan de criteriumscore scoren geclassificeerd worden als ongeschikt. In de praktijk kiest men er soms voor om geen uitspraak te doen over kandidaten voor wie de criteriumscore in het 90% *CI* ligt. Immers, men wil een hoge mate van zekerheid dat men de juiste beslissing neemt voor individuele kandidaten voordat men tot een oordeel overgaat. Deze beslisregel werd niet door Kruyen et al. (2012) bestudeerd in de door hen onderzochte scenario's, maar ik neem hem hier mee.

- *Simulatieprocedure.*

In dit scenario liet ik een miljoen hypothetische kandidaten duizend maal dezelfde 40, 20, 15, 10 en 5-item testversie maken. Wederom was de betrouwbaarheid van de langste test gelijk aan $r_{XX'} = .99$ en voor de kortste test $r_{XX'} = .90$. Anders dan in het scenario waarin twee kandidaten met elkaar worden vergeleken (zie 4.1), selecteerde ik nu voor de kortere testversies items met een populariteit in de buurt van de criteriumscore om zodoende de kans te vergroten dat voor hypothetische kandidaten in de buurt van de criteriumscore de juiste beslissing werd genomen.

De criteriumscore T_c is in dit scenario afhankelijk van de *base rate*. In simulatieonderzoek is T_c eenvoudig te bepalen omdat men toegang heeft tot de verdeling van betrouwbare scores T. Wanneer men bijvoorbeeld kiest voor een *base rate* van 12.50%, dan neemt men de betrouwbare score van die kandidaat die nog net bij de 12.50% kandidaten met de hoogste betrouwbare scores hoort als criteriumscore. In deze bijdrage werd zo bijvoorbeeld voor de 5-item test de criteriumscore gesteld op 18 bij een *base rate* van 12.50%. Ik onderzocht de gevolgen van testverkorting voor een *base rate* van 12.50%, 37.50%, 50%, 62.50% en 87.50%.

Tabel 1 *Resultaten voor cut score selectie met behulp van korte betrouwbare tests*

J	$r_{XX'}$	PPW	NPW	CC+	CC-	$\|CI\|>T_c$
			Base rate = 12.50%			
40	.99	.93	.99	.78	.96	.82
20	.97	.91	.99	.68	.94	.61
15	.96	.90	.99	.62	.92	.56
10	.94	.88	.99	.55	.90	.41
5	.90	.85	.99	.42	.85	.23
			Base rate = 37.50%			
40	.99	.96	.98	.84	.90	.72
20	.97	.95	.97	.78	.85	.48
15	.96	.94	.97	.75	.83	.43
10	.94	.93	.97	.68	.79	.27
5	.90	.91	.96	.53	.69	.13
			Base rate = 50.00%			
40	.99	.97	.97	.87	.88	.70
20	.97	.96	.96	.81	.82	.47
15	.96	.96	.96	.79	.78	.42
10	.94	.95	.95	.73	.73	.26
5	.90	.93	.95	.57	.61	.13
			Base rate = 62.50%			
40	.99	.98	.96	.90	.85	.70
20	.97	.97	.95	.86	.78	.47
15	.96	.97	.94	.83	.75	.42
10	.94	.96	.94	.78	.68	.26
5	.90	.96	.91	.69	.52	.13
			Base rate = 87.50%			
40	.99	.99	.93	.96	.77	.80
20	.97	.99	.90	.94	.66	.58
15	.96	.99	.89	.92	.59	.52
10	.94	.99	.87	.91	.51	.42
5	.90	.99	.85	.86	.43	.25

NB *J* = Testlengte; $r_{XX'}$ = Testbetrouwbaarheid; PPW = Positieve predictieve waarde; NPW = Negatieve predictieve waarde; CC+ en CC- = Proportie van kandidaten voor wie in ten minste 90% van de testreplicaties de juiste beslissing kon worden genomen indien men geschikt is (CC+) of ongeschikt is (CC-); $|CI|>T_c$ = Proportie van kandidaten voor wie de criteriumscore niet in het 90% betrouwbaarheidsinterval lag.

- *Simulatieresultaten.*

Tabel 1 toont de effecten van testverkorting voor dit scenario. Voor een *base rate* van 12.50% en een testlengte 5 bedraagt de PPW .85, de NPW . 99, de CC+ .42 en

de CC- .85 (zie Kolom 3 tot 6 in Tabel 1). Zo'n hoge PPW, NPW en CC- zijn voor de praktijk zeer acceptabel, maar dat geldt waarschijnlijk niet voor een CC+ van .42. Een CC+ van .42 betekent dat 68% van de geschikte kandidaten in ten minste 10% van de (hypothetische) testreplicaties als ongeschikt wordt bevonden. Voor een *base rate* van 87.50% zijn de resultaten gespiegeld, zoals te verwachten viel op basis van Kruyen et al. (2012): de CC+ bedraagt nu .86 en de CC- bedraagt .43.

Vergelijkt men de resultaten in deze bijdrage met de gevonden effecten in Kruyen et al. (2012), dan zijn de gevolgen van testverkorting voor de predictieve waarden en CC minder extreem. In het bijzonder zijn de verschillen kleiner voor tests bestaande uit 20, 15 en 10 items wanneer de betrouwbaarheid $r_{XX'}$ hoger is dan .90 dan wanneer de betrouwbaarheid van deze tests nabij .80 ligt zoals in het simulatieonderzoek van Kruyen et al. (2012). Dus korte tests, met een zeer hoge betrouwbaarheid, zijn meer geschikt voor dit scenario dan tests met een hoge betrouwbaarheid, behalve bij een zeer hoge (of lage) *base rate* en men veel waarde hecht aan een hoge CC+ (of CC-).

Het verhaal wordt echter anders wanneer men alleen beslissingen wil nemen over personen voor wie de criteriumscore niet in het 90% *CI* ligt (zie Kolom 7 in Tabel 1). In dit geval neemt de proportie kandidaten over wie men een beslissing kan nemen schrikbarend af als gevolg van testverkorting. Zo daalt deze proportie van .82 voor de 40-item test tot .23 voor de 5-item test voor een *base rate* van 12.50%. Opnieuw blijkt dat korte tests – ook al hebben zij een zeer hoge betrouwbaarheid – niet geschikt zijn om beslissingen te nemen over individuele kandidaten wanneer men rekening wil houden met de meetprecisie.

4.3 Classificatie met behulp van adaptieve tests

- *Selectiesituatie.*

Een belangrijke nieuwe ontwikkeling binnen de werving- en selectiepraktijk is het gebruik van adaptieve selectietests (Schakel, 2012). Bij deze computergestuurde vorm van toetsen wordt de test *optimaal* afgestemd op de geteste kandidaat. Een belangrijk voordeel hiervan is dat je kandidaten niet langer lastigvalt met onnodige vragen, dat wil zeggen vragen die te eenvoudig of te lastig zijn voor de betreffende kandidaten. Adaptieve selectietests worden bijvoorbeeld ingezet om PABO-studenten te selecteren op basis van hun rekenvaardigheden (Straetmans & Eggen, 2007). Ook aspirant-politieagenten krijgen een adaptieve capaciteitentest voorgelegd in het selectieproces (Staatscourant, 2012).

Adaptieve selectietests werken als volgt. Een sollicitant start met enkele items van gemiddelde moeilijkheid (of populariteit in geval van persoonlijkheidstests). Afhankelijk van zijn/haar antwoorden op deze items wordt een volgend item gekozen. Scoort een kandidaat hoog op een item, dan wordt vervolgens een moeilijker (minder populair) item aan deze kandidaat voorgelegd of anders een gemakkelijker (populairder) item. De adaptieve test eindigt op het moment dat de test een voldoende precieze meting heeft opgeleverd of wanneer alle items gebruikt zijn. Wat als voldoende precies wordt beschouwd, is een keuze van de selectiepsycholoog en hangt af van de specifieke toepassing.

Wat zijn de consequenties voor testlengte wanneer men een adaptieve test inzet? Leidt het gebruik van een adaptieve test ertoe dat men altijd veel minder items nodig heeft om met voldoende zekerheid beslissingen te nemen? Om deze vraag te beantwoorden legde ik in een nieuwe simulatiestudie aan 1000 hypothetische kandidaten een adaptieve test voor, bestaande uit 40 items. Ik besloot iedere kandidaat zoveel items voor te leggen totdat de criteriumscore T_c niet meer in zijn/haar betrouwbaarheidsinterval *CI* viel. Op dat moment had ik voldoende zekerheid dat een kandidaat terecht werd geselecteerd of afgewezen.

- *Simulatieprocedure.*

In deze simulatiestudie bestudeerde ik een aantal verschillende condities. Ten eerste onderzocht ik of de vereiste testlengte verschilde voor een adaptieve test bestaande uit *dichotome* items of *polytome* items met een 5-puntsschaal. Daarnaast legde ik de hypothetische kandidaten items voor afkomstig uit een betrouwbare 40-item test ($r_{XX'}$ = .90) *of* items afkomstig uit een zeer betrouwbare 40-item test ($r_{XX'}$ = .99). Ook bestudeerde ik de consequenties voor de keuze van een 90% of 95% *CI*. Ten slotte onderzocht ik het aantal benodigde items voor iedere kandidaat wederom voor een *base rates* van 12.50%, 37.50%, 50.00%, 62.50% en 87.50%. Voor iedere hypothetische kandidaat herhaalde ik de procedure honderdmaal om de CC te kunnen schatten. Om dit simulatieonderzoek uit te voeren maakte ik gebruik van het softwarepakket R en verschillende programmacodes, ontwikkeld door Nydick (2015).

- *Simulatieresultaten.*

De uitkomsten van deze kleine simulatiestudie waren vergelijkbaar voor de tests bestaande uit *dichotome* items en de tests bestaande uit *polytome* items. Ik beperk mij in deze bijdrage tot de bespreking van de resultaten voor de tests bestaande uit *polytome* items. Verder leverde de keuze voor een *base rate* van 12.50% en 87.50% en voor de *base rate* van 37.50% en 62.50% dezelfde, maar gespiegelde resultaten op voor de geschikte en ongeschikte kandidaten. Ik bespreek alleen de resultaten voor de *base rate* van 50.00%, 62.50% en 87.50%.

De eerste kolom van Tabel 2 geeft de resultaten voor een adaptieve test bestaande uit 40 betrouwbare items wanneer selectiebeslissingen werden genomen op basis van het 90% *CI* en een *base rate* van 50.00%. In deze conditie bleken er gemiddeld genomen 8.13 items nodig voor geschikte kandidaten (Rij 1) en 9.64 items voor ongeschikte kandidaten (Rij 5) voordat de procedure werd beëindigd omdat *of* de criteriumscore T_c niet langer in het 90% *CI* lag *of* omdat er na alle 40 items nog steeds geen beslissing kon worden genomen. 70% procent van de geschikte kandidaten werd in deze conditie in ten minste 90% van de testreplicaties geselecteerd (CC+, Rij 3) terwijl 66% van de ongeschikte kandidaten in ten minste 90% van de testreplicaties werd afgewezen (CC-, Rij 8).

Tabel 2 *Resultaten voor cut score selectie met behulp van een adaptieve test bestaande uit polytome items*

		betrouwbare 40-item test ($r_{XX'}$ = .90)						zeer betrouwbare 40-item test ($r_{XX'}$ = .99)					
		90% *CI*			95% *CI*			90% *CI*			95% *CI*		
	Base rate	50.00%	62.50%	87.50%	50.00%	62.50%	87.50%	50.00%	62.50%	87.50%	50.00%	62.50%	87.50%
						Geschikte kandidaten							
J		8.13	7.65	3.91	11.96	9.24	5.01	7.68	6.66	3.92	10.23	7.95	4.59
Sd(*J*)		8.69	8.79	5.45	11.33	10.32	7.29	8.36	7.48	5.21	10.10	9.35	6.52
CC+		.70	.74	.89	.65	.70	.87	.78	.83	.92	.75	.79	.91
Onbeslist		.11	.11	.03	.20	.15	.05	.09	.06	.02	.13	.10	.03
						Ongeschikte kandidaten							
J		9.64	11.96	16.58	12.83	14.47	21.06	8.49	10.20	14.57	10.27	12.27	17.48
Sd(J)		8.77	9.67	9.50	11.13	11.49	11.36	8.10	8.74	9.79	10.12	10.54	11.52
CC-		.66	.58	.37	.63	.53	.27	.77	.69	.55	.74	.66	.50
Onbeslist		.16	.21	.37	.20	.28	.47	.10	.14	.30	.12	.18	.34

NB *J* = Gemiddelde aantal benodigde items per kandidaat voordat de adaptieve test werd beëindigd omdat de criteriumscore niet meer in het betrouwbaarheidsinterval *CI* viel; *Sd*(*J*) = Standaarddeviatie van het aantal items per kandidaat voordat de adaptieve test werd beëindigd; CC+ en CC- = Proportie van kandidaten voor wie in ten minste 90% van de testreplicaties de juiste beslissing werd genomen indien men geschikt is (CC+) of ongeschikt is (CC-); Onbeslist = Proportie van kandidaten voor wie in ten minste 50% van de replicaties geen beslissing kon worden genomen.

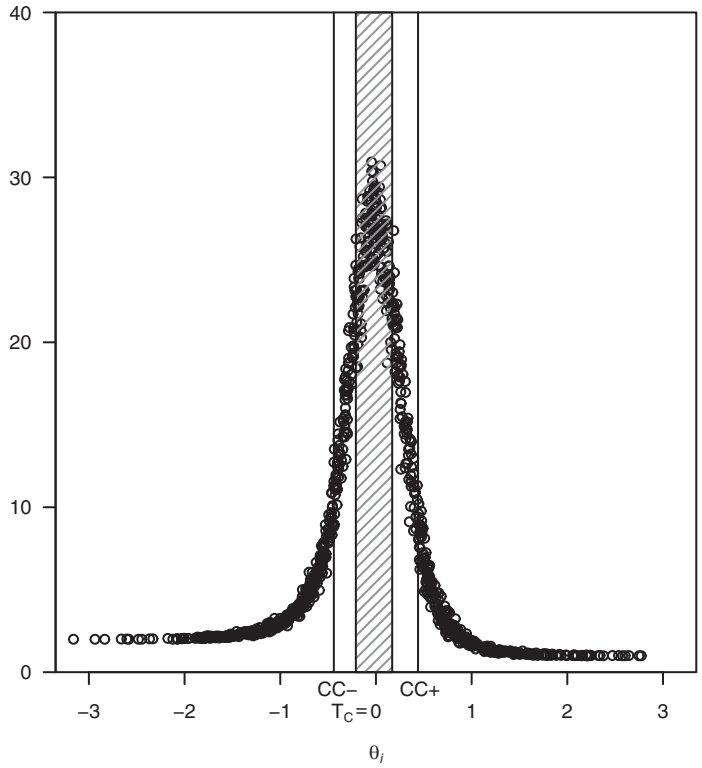
De spreiding van het aantal voorgelegde items in deze conditie was echter groot (zie Tabel 2, Kolom 1, Rij 2 en Rij 6), hetgeen ook blijkt uit Figuur 2. Hoe dichter een betrouwbare score bij de criteriumscore lag, hoe meer items nodig waren voordat de procedure werd beëindigd. Daarentegen gold dat voor kandidaten met een betrouwbare score die veraf lag van T_c al 1 tot 3 items voldoende waren om tot een beslissing te komen. Verder blijkt ook dat 11% van de geschikte kandidaten en 16% van de ongeschikte kandidaten in ten minste 50% van de testreplicaties niet konden worden geclassificeerd nadat alle 40 items afgenomen waren (zie Tabel 2, Kolom 1, Rij 4 en Rij 8). Deze kandidaten bevinden zich in het gearceerde gedeelte van Figuur 2. Op basis van Figuur 2 kan men ook concluderen dat het niet veel zin had om, bijvoorbeeld, meer dan 20 items aan deze kandidaten voor te leggen. De kans was miniem dat het laten beantwoorden van nog meer items wel tot een classificatiebeslissing zou leiden.

De keuze voor een hogere *base rate* laat dezelfde trend zien als bij niet-adaptieve tests (zie Tabel 2, Kolom 1-3). Hoe hoger de *base rate*, hoe minder items er gemiddeld nodig waren voor geschikte kandidaten. Immers, hoe hoger de *base rate* hoe groter de proportie geschikte kandidaten met een betrouwbare score die veraf ligt van T_c. Voor ongeschikte kandidaten geldt het tegenovergestelde. Een hogere *base rate* betekent dat er verhoudingsgewijs veel ongeschikte kandidaten zijn met een betrouwbare score in de buurt van T_c. Evenzo is CC+ hoger en CC- lager voor hogere *base rates*.

Wordt er gebruikgemaakt van een 95% *CI* (zie Tabel 2, Kolom 4-6) dan blijkt dat gemiddeld twee tot vier items meer nodig waren voordat de testprocedure werd beëindigd dan wanneer beslissingen werden genomen op basis van het 90% *CI*. Het voorleggen van deze extra items resulteerde echter niet in een substantieel hogere CC+ en CC-. Tegelijkertijd viel de proportie kandidaten voor wie in ten minste 50% van de replicaties geen beslissing kon worden genomen, hoger uit, in het bijzonder voor de groep ongeschikte kandidaten bij een *base rate* van 87.50%.

Vergelijkt men de resultaten voor de adaptieve tests bestaande uit items met een zeer hoge betrouwbaarheid (zie Tabel 2, Kolom 7-12) met de adaptieve tests bestaande uit items met een hoge betrouwbaarheid (zie Tabel 2, Kolom 1-6), dan valt op dat voor de groep geschikte kandidaten de extra betrouwbaarheid weinig meerwaarde had: CC+ en de proportie kandidaten voor wie geen beslissing genomen kon worden, waren ongeveer gelijk. Daarentegen had de extra betrouwbaarheid wel een positief effect voor de groep ongeschikte kandidaten: CC- viel substantieel hoger uit terwijl de groep ongeschikte kandidaten voor wie men geen beslissing kon nemen, in ten minste 50% van de testreplicaties flink lager was.

Vergelijkt men de resultaten van de adaptieve tests met de resultaten van de niet-adaptieve, standaardtests in Kruyen et al. (2012, in het bijzonder Tabel 2) en Scenario 2 in deze bijdrage (zie Tabel 1), dan blijkt dat het gebruik van een adaptieve test ertoe leidt dat men gemiddeld veel minder items nodig heeft dan bij niet-adaptieve tests om vergelijkbare prestaties te bereiken. Concreet komen de resultaten van de adaptieve test met betrouwbare items overeen met de resultaten van een niet-adaptieve test met 20 betrouwbare items bij een 90% *CI* en 15 items bij een 95% *CI*. De resultaten van de adaptieve test met zeer betrouwbare items zijn vergelijkbaar met de resultaten van een niet-adaptieve test met 15 zeer betrouw-



NB θ_i = De naar θ getransformeerd betrouwbare score T voor gesimuleerde sollicitant i (adaptieve tests plaatsen respondenten op θ-schalen in plaats van geobserveerde scoreschalen X_+) waarbij de criteriumscore T_c = 0 correspondeert met een *base rate* van 50%. J = Het aantal door gesimuleerde sollicitant i beantwoorde items (gemiddelde over 100 replicaties). Voor de gesimuleerde sollicitanten links van CC- en rechts van CC+ werd in ten minste 90% van de replicaties de juiste beslissing genomen (d.w.z. correct afwezen / correct geselecteerd). Voor de gesimuleerde sollicitanten in het gearceerde deel was het in ten minste 50% van de replicaties niet mogelijk om een beslissing te nemen.

Figuur 2 ***Gemiddelde testlengte in een adaptieve test met een betrouwbare 40-item test ($r_{XX'}$ = .90) voor 1000 gesimuleerde sollicitanten bij cut score selectie en een base rate van 50%***

bare items bij een 90% *CI* en 10 items bij een 95% *CI*. Echter, hoewel er gemiddeld genomen minder items nodig zijn in een adaptieve test, betekent dat niet dat een adaptieve test het risico op verkeerde beslissingen voor iedere kandidaat minimaliseert. Zo was CC+ (CC-) nog steeds laag bij lage (hoge) *base rates* en kon men een substantieel deel van de kandidaten nog steeds niet met voldoende zekerheid classificeren nadat alle 40 items waren voorgelegd.

5 Slotbeschouwing

Gezien de praktische voordelen is het logisch dat selectiepsychologen inzetten op korte tests, d.w.z. tests bestaande uit minder dan 20 items. Een aantal auteurs beargumenteren daarbij dat het gebruik van korte tests geoorloofd is zolang een hoge betrouwbaarheid gegarandeerd is (bijv. Fu, Liu & Yip, 2007; Marteau & Bekker, 1992; Olatunji et al., 2010). Echter, om de geschiktheid van korte tests in de context van personeelsselectie te bepalen, dient men in eerste instantie niet naar de gerapporteerde testbetrouwbaarheid te kijken maar naar de gewenste zekerheid op een juiste beslissing en de *base rate*.

Wat zijn nu belangrijke punten waar selectiepsychologen rekening mee dienen te houden wanneer zij geconfronteerd worden met vraagstukken rondom de geschiktheid van korte tests om selectiebeslissingen te nemen? In de eerste plaats beveel ik aan om beslissingen over de inzet van psychologische tests en vragenlijsten niet te laten leiden door testlengte maar te baseren op informatie over de begripsvaliditeit (d.w.z. hoe zeker men is dat de test meet wat hij zou moeten meten volgens de testaanbieder). Een test kan nog zo lang of kort zijn, als individuele testscores niet geïnterpreteerd kunnen worden, dan heeft men in het minst erge geval een 'kat in een zak' gekocht. In deze bijdrage nam ik impliciet aan dat de korte testversies even valide waren als de langere testversies.

Ten tweede, zouden selectiepsychologen moeten nagaan in hoeverre zij een bepaald risico op verkeerde selectiebeslissingen ethisch verantwoord vinden, alsmede hoe men de kosten inschat van het aannemen (afwijzen) van ongeschikte (geschikte) kandidaten. Principieel en belangrijker dan de vraag hoeveel items volstaan, is welke risico's op verkeerde beslissingen de selecterende organisatie wil dragen en of men het wel of niet belangrijk vindt dat voor individuele kandidaten de juiste beslissing wordt genomen.

Pas wanneer men zeker weet dat men een valide test heeft *en* men bepaald heeft welke risico's op onjuiste selectiebeslissingen men nog verantwoord vindt, komt testlengte in beeld. Deze bijdrage liet zien dat de benodigde testlengte in de eerste plaats afhankelijk is van de gewenste zekerheid op juiste beslissingen. Wanneer men rekening wil houden met de meetprecisie – bijvoorbeeld wanneer men individuele kandidaten met elkaar wil vergelijken of individuele kandidaten wil classificeren in verschillende categorieën – dan toont deze bijdrage dat korte tests ongeschikt zijn, zelfs als hun betrouwbaarheid zeer hoog is.

Als men voor het leeuwendeel van de kandidaten voldoende zekerheid wil dat de juiste beslissing wordt genomen maar men niet geïnteresseerd is in meetprecisie, dan is de geschiktheid van korte tests afhankelijk van de *base rate* en verschilt de impact van testverkorting voor geschikte en ongeschikte kandidaten. Wanneer de *base rate* laag is, bieden zelfs tests bestaande uit 40 zeer betrouwbare items geen garantie dat voor het merendeel van de geschikte kandidaten de juiste beslissing wordt genomen. Wil men bij een lage *base rate* dat juist ongeschikte kandidaten buiten de deur worden gehouden, dan bieden 5 betrouwbare items reeds voldoende zekerheid. Bij een hoge *base rate* geldt juist het tegenovergestelde, dan biedt een korte test juist voldoende zekerheid dat voor het grootste gedeelte van de geschikte kandidaten de juiste beslissing wordt genomen maar is dit *niet* het

geval voor de ongeschikte kandidaten (zelfs niet als de betrouwbaarheid zeer hoog is).

Sommige organisaties zijn in het geheel niet geïnteresseerd in individuele kandidaten en willen alleen dat de proportie geschikte kandidaten in de groep geselecteerde kandidaten en/of de proportie ongeschikte kandidaten in de groep afgewezen kandidaten zo groot mogelijk is. In dat geval hangt de geschiktheid van korte tests in de eerste plaats af van de betrouwbaarheid. Is de betrouwbaarheid voldoende dan geldt dat wederom de geschiktheid afhankelijk is van de *base rate* en verschilt de impact van testverkorting voor geschikte en ongeschikte kandidaten (Kruyen et al., 2012). Is de betrouwbaarheid zeer hoog, zoals in deze bijdrage onderzocht, dan geldt voor iedere *base rate* dat korte tests zonder problemen kunnen worden gebruikt.

In deze bijdrage verkende ik ook de mogelijkheden van computergestuurde adaptieve tests om selectiebeslissingen te nemen. Een belangrijk voordeel van adaptieve tests is dat sollicitanten geen onnodige vragen worden voorgelegd. Het computeralgoritme slaat items die te makkelijk of te lastig zijn voor een specifieke kandidaat over. Hierdoor kan er veel tijdwinst behaald worden en blijven kandidaten gemotiveerd. Inderdaad, uit mijn kleine simulatiestudie bleek dat – wanneer men een adaptieve test gebruikt om sollicitanten te classificeren in twee categorieën – men voor een grote groep kandidaten minder dan 15 items nodig heeft voordat de procedure beëindigd wordt. Echter, dat de adaptieve test wordt afgebroken, betekent niet dat men de zekerheid heeft dat voor iedere kandidaat de juiste beslissing wordt genomen. Sterker nog, zelfs nadat alle 40 items voorgelegd waren, was het voor een substantieel deel van de (hypothetische) kandidaten nog steeds niet mogelijk om een beslissing te nemen. In toekomstig onderzoek kan men bestuderen of andere beslisregels bij adaptief testen leiden tot betere resultaten.

Adaptieve testen hebben bovendien ook enkele belangrijke nadelen. Zo maken adaptieve tests gebruik van item respons modellen. Dit impliceert dat selectiepsychologen moeten onderzoeken of het gekozen model realistisch is voordat men het kan inzetten in selectieprocedures. Hierbij komt dat een groot aantal keuzes moet worden gemaakt in de inrichting van de test die lang niet altijd even gunstig uitpakken (zoals bleek uit mijn eigen kleine simulatiestudie). Daarnaast vereist al dit werk specialistische psychometrische kennis. Voor vervolgonderzoek roept dit de vraag op of de problemen die adaptief testen met zich meebrengen niet zwaarder wegen dan de voordelen van minder items.

In deze bijdrage ging ik voorbij aan het vraagstuk van de ervaren of subjectieve testlengte. Op basis van een uitgebreide literatuurstudie naar testverkortingspraktijken (Kruyen, Emons & Sijtsma, 2013) kan men concluderen dat tests bestaand uit minder dan 20 items over het algemeen als kort beschouwd worden, terwijl tests bestaande uit 40 of meer items zeker als lang worden gezien. Echter, een 40-item test kan als kort worden beschouwd wanneer het weinig tijd kost om alle items te beantwoorden en, wellicht nog belangrijker, wanneer de test als leuk en nuttig wordt gezien. Evenzo kan een 5-item test kandidaten veel tijd kosten, zeker wanneer de test als weinig uitdagend of zinloos wordt ervaren. Voor selec-

tiepsychologen is de ervaren testlengte, naast de 'objectieve' testlengte, een belangrijk punt om in de praktijk rekening mee te houden.
Hoewel adaptieve tests waarschijnlijk steeds belangrijker worden om geschikte kandidaten te selecteren, is de rol van gestandaardiseerde tests nog lang niet uitgespeeld. Immers, deze tests zijn eenvoudig af te nemen, hun scores zijn makkelijk te interpreteren, en zij hebben een bewezen staat van dienst. Wanneer selectiepsychologen gevraagd worden naar hun oordeel over de inzet van *korte* gestandaardiseerde tests, raad ik hen aan om de toename in risico's op verkeerde beslissingen als gevolg van testverkorting bewust af te wegen tegen de verwachte praktische voordelen. Wil men het zekere voor het onzekere nemen, dan is aan te bevelen om gebruik te blijven maken van tests bestaande uit een groot aantal items om een specifiek construct te meten in plaats van over te stappen naar kortere testversies. Hierbij doet men er goed aan om kandidaten uit te leggen waarom het belangrijk is om geconcentreerd en gemotiveerd de vele vragen te beantwoorden. Korte tests mogen dan wel efficiënt zijn, ze zijn lang niet altijd effectief.

Praktijkbox

Wat betekenen de resultaten voor de praktijk?

- Bij de keuze van psychologische tests en vragenlijsten voor selectiedoeleinden mag men zich niet laten leiden door testlengte, maar dient men zich eerst en vooral te baseren op validiteitsinformatie *en* de zekerheid waarmee men juiste beslissingen wil nemen.
- Als men toch korte tests gebruikt, kan dit leiden tot verkeerde beslissingen. Of het gebruik van korte tests leidt tot verkeerde beslissingen is echter afhankelijk van de gewenste zekerheid op juiste beslissingen alsook van de *base rate* en verschilt daarbij voor de groep geschikte en ongeschikte kandidaten.
- In sommige omstandigheden bieden 5 items al voldoende zekerheid op een juiste beslissing, maar in andere gevallen volstaan zelfs 40 items met een hoge betrouwbaarheid niet, zelfs niet wanneer men gebruikmaakt van een adaptieve test.
- Wanneer men toch een kortere (al dan niet adaptieve) test overweegt, kunnen de risico's op verkeerde beslissingen aanzienlijk groter zijn voor individuele kandidaten dan voor de selecterende organisatie, zelfs wanneer testbetrouwbaarheid hoog is.
- Wil men het zekere voor het onzekere nemen, dan is het aan te bevelen om gebruik te maken van tests bestaande uit veel items om selectiebeslissingen te nemen.

Literatuur

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322. doi: 10.1111/j.2044-8295.1910.tb00207.x

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality, 11*, 303-315. doi: 10.1002/(SICI)1099-0984(199711)11:4<303::AID-PER292>3.0.CO;2-#

Butter, R. (2014). Ecologische schalen als personeelspsychologisch antwoord op situationele gedragsverschillen: Ontwikkeling van een consciëntieusheidschaal voor promovendi. *Gedrag & Organisatie, 27*, 290-308.

Cook, M. (2009). *Personnel selection: Adding value through people* (Ffifth edition). Oxford, UK: Wiley-Blackwell. doi: 10.1002/9780470742723

Costa, P.T., & McCrae, R.R. (1992). *The NEO PI/FFI manual supplement*. Odessa, FL: Psychological Assessment Resources.

De Vries, A., De Vries, R.E., Born, M.Ph., & Van den Berg, R. (2014). Persoonlijkheid als voorspeller van werkprestaties en contraproductief werkgedrag. *Gedrag & Organisatie*, *27*, 407-427.

Drabbe, J.P., Drost, D., Klehe, U.-C., Van Vianen, A.E.M., & Boendermaker, W. (2008). Personeelsselectie in Nederland: Aanbevelingen voor selectie van personeel in tijden van krapte. Gedownload van www.dare.uva.nl/document/126587

Drenth, P.J.D. (1965). *Test voor Niet Verbale Abstractie: Handleiding*. Amsterdam: Swets & Zeitlinger.

Emons, W.H.M., Sijtsma, K., & Meijer, R.R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12*, 105-120. doi: 10.1037/1082-989x.12.1.105

Fu, K., Liu, K.Y., & Yip, P.S.F. (2007). Predictive validity of the Chinese version of the Adult Suicidal Ideation Questionnaire: Psychometric properties and its short version. *Psychological Assessment, 19*, 422-429. doi: 10.1037/1040-3590.19.4.422

Harvill, L.M. (1991). An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice*, *10*, 33-41.

Hunter, J.E., & Schmidt, F.L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M.D. Dunnette & E.A. Fleishman (Eds.), *Human performance and productivity: Human capability assessment* (Vol. 1, pp. 233-284). Hillsdale, NJ: Lawrence Erlbaum Associates.

International Test Commission. (2000). *International guidelines for test use*. Gedownload op 4 juni 2012 van www.intestcom.org.

Kruyen, P.M., Emons, W.H.M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing, 12*, 321-344. doi: 10.1080/15305058.2011.643517

Kruyen, P.M., Emons, W.H.M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. International Journal of Testing, 13, 223-248. doi: 10.1080/15305058.2012.703734

Marteau, T.M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology, 31*, 301-306. doi: 10.1111/j.2044-8260.1992.tb00997.x

Mellenbergh, G.J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*, 293-299. doi: 10.1037//1082-989X.1.3.293

Netemeyer, R.G., Pulling, C., & Bearden, W.O. (2003). Observations on some key psychometric properties of paper-and-pencil measures. In A.G. Woodside & E.M. Moore (Eds.), *Advances in business marketing and purchasing: Essays by distinguished scholars* (pp. 115-138). New York: Elsevier Science.

Nydick, S.W. (2015). *catIrt: An R package for simulating IRT-based computerized adaptive tests*. R package version 0.5-0.

Olatunji, B.O., Sawchuk, C.N., Moretz, M.W., David, B., Armstrong, T., & Ciesielski, B.G. (2010). Factor structure and psychometric properties of the Injection Phobia Scale-Anxiety. *Psychological Assessment, 22*, 167-179. doi: 10.1037/a0018125

Rothstein, M.G., & Goffin, R.D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, *16*, 155-180. doi: 10.1016/j.hrmr.2006.03.004

Schakel, L. (2012). *Online computer-based testing in human resource management: Contributions from item response theory*. Gedownload van www.irs.ub.rug.nl/ppn/352188936.

Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing, 9*, 167-194. doi: 10.1080/15305050903106883

Smith, M., & Smith, P.C. (2005). *Testing people at work: Competencies in psychometric testing*. Malden, MA: BPS Blackwell.

Spearman, C.C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Staatscourant. (2012). *Regeling van de Minister van Veiligheid en Justitie van 1 november 2012 nr. 1 tot wijziging van de Regeling aanstellingseisen politie 2002 in verband met nieuwe normen voor het geschiktheidonderzoek en enkele andere wijzigingen*. Gedownload op 15 oktober 2014 van www.zoek.officielebekendmakingen.nl/stcrt-2012-25901.html.

Straetmans, G.J.J.M., & Eggen, T.J.H.M. (2007). Wiscat-pabo: Computergestuurd adaptief toetspakket rekenen, *Onderwijsinnovatie*, *3*, 17-27.

Tillema, H.H. (1998). Assessment of potential, from assessment centers to development centers. *International Journal of Selection and Assesment, 6*, 185-191. doi: 10.1111/1468-2389.00088

When short is too short: The suitability of short psychological tests for personnel selection

Peter Kruyen, Gedrag & Organisatie, Volume 28, December 2015, nr. 4, pp. 337-355.

Psychological tests and questionnaires are commonly used in personnel recruitment and selection procedures. Because test length has a considerable impact on assessment time and costs, short tests consisting of fewer than 20 items, for instance, are preferable to long tests. Simulated data show that reducing test length can have a substantial impact on the risk of making incorrect selection decisions. However, the impact of shortening tests varies across situations. In this article, I illustrate testing settings in which five items are sufficient to make reliable decisions about individual candidates. I also present scenarios in which 40 items are still insufficient to draw reliable conclusions, even when using an adaptive test. The conclusion is that long tests are preferable to short tests for making personnel selection decisions.

Key words: personnel selection, test length, measurement precision, classification consistency, decision errors