Wetenschap

Situationele beoordelingstests in de schijnwerpers

Situationele beoordelingstests *(situational judgment tests, SJT's)* staan in de belangstelling. Studies in de VS melden hoge validiteitscoëfficiënten, een behoorlijke validiteitswinst, een goede acceptatie door kandidaten en een relatief geringe scoreachterstand van allochtonen. In onderstaande bijdrage worden de kenmerken van dit soort tests geschetst. De wijze van scoren wordt aangegeven en er wordt ingegaan op bestaand onderzoek naar psychometrische kenmerken. De voordelen van audiovisuele SJT's komen aan de orde. SJT's roepen echter ook kritische vragen op.

Paul van der Maesen de Sombreff, Marise Born, Karen van Oudenhoven-van der Zee en Denise Ruhe

Sinds 1990 is de aandacht voor situationele beoordelingstests (in het Engels: *situational judgment tests* of SJT's) sterk toegenomen. Bij dit type test worden kandidaten voor een functie geconfronteerd met een aantal situaties waar zij ofwel zelf op moeten reageren, of waarbij zij een aantal gegeven reacties moeten beoordelen op hun effectiviteit. Kenmerkend is dat de situaties representatief zijn voor de functie waar het om gaat. Situationele beoordelingstests kunnen de vorm hebben van gedragssimulaties, *paper & pencil*-opdrachten (bijvoorbeeld postbakoefeningen) of multimediale tests. Hoewel zij ontwikkeld kunnen worden voor allerlei verschillende competenties, draait het bij SJT's in de meeste gevallen om interpersoonlijke en praktisch-organiserende vaardigheden. SJT's zijn al lang in gebruik voor selectiedoeleinden, vooral voor managementfuncties en commerciële functies. Uit een overzicht van McDaniel, Morgeson, Finnegan, Campion en Braverman (2001) blijkt dat het eerste gebruik van SJT's dateert van 1920. In het Nederlands taalgebied is de Vragenlijst voor Commercieel Inzicht (Lievens, Steverdinck, Tjoa & Verhoeven, 1985) een bekend voorbeeld van een SJT. Voordelen van dit type instrument liggen in zijn hoge voorspellende waarde en in zijn acceptatie door kandidaten. Bovendien zijn SJT's aantrekkelijk omdat zij allochtone kandidaten minder lijken te benadelen dan bijvoorbeeld intelligentietests. Dat geldt zelfs bij schriftelijke varianten.

In dit artikel zullen we allereerst dieper ingaan op de meetpretentie van SJT's en hun plaats tussen andere simulatie-instrumenten. Daarnaast wordt aandacht besteed aan de manier waarop de beantwoording en de scoring verloopt, en aan hun psychometrische kwaliteiten en acceptatie door kandidaten.

We zullen ook ingaan op de multimediale variant van SJT's. In het bijzonder doen we verslag van de zogenaamde SQ-test, een multimedia SJT waarmee we de afgelopen jaren veel ervaring hebben opgedaan, en van een SJT-methode die we e-cartoon hebben genoemd en die vooral geschikt lijkt voor wervingsdoeleinden.

Ten slotte komt in de discussie een aantal vragen aan de orde die nadere opheldering vereisen. We eindigen met ons beeld van toepassing van SJT's in de toekomst.

Wat zijn SJT's?

Zoals aangegeven bestaan SJT's uit beschrijvingen van een aantal situaties. Meestal betreffen deze situaties een probleem waarmee iemand in het werk geconfronteerd wordt en dat van die persoon om een actie vraagt. De kandidaat wordt gevraagd om een reactie te geven of ontvangt een aantal alternatieve acties die de persoon in de gegeven situatie kan ondernemen. In plaats van acties kan het ook gaan om opinies of intenties die de persoon erop nahoudt. Meestal worden situaties en eventuele acties in de vorm van teksten aangeboden. Een voorbeeld van een schriftelijke SJT-situatie met acties is in Kader 1 opgenomen. Als acties reeds worden aangegeven, is het de bedoeling dat een kandidaat deze beoordeelt op effectiviteit (of kwaliteit) of dat een persoon zich uitspreekt over de waarschijnlijkheid dat de persoon zelf zo zou handelen.

SJT's zien er anders uit dan andere psychologische tests en vragenlijsten. Het is dan de vraag of ze ook iets anders meten. Wat opvalt is dat SJT's veelal een beroep doen op praktisch sociaal inzicht. Opdrachten van dit type komen ook wel voor in tests voor analytische intelligentie, bijvoorbeeld in de subtest 'Begrijpen' van de WAIS die een beroep doet op 'gezond verstand'. Items zijn van het type 'Er breekt brand uit in de bioscoop. (Wat doe je? a. Schreeuwen 'Brand!' et cetera.).

Kenmerkend voor deze opgaven die beroep doen op het gezond verstand (*common sense*) is dat er gevraagd wordt naar inzicht in wat het meest effectieve gedrag is in de gegeven situatie. 'Effectief' is gekoppeld aan doelen die in gedrag gerealiseerd moeten worden. Wat die doelen zijn en welke prioriteit zij hebben, wordt in het midden gelaten.

Volgens sommige auteurs is het construct achter SJT's fundamenteel anders dan het construct achter een instrument als de intelligentietest. De manier van kennis opdoen bij het soort

Achtergrondinformatie
De parkeerwachter die optreedt in de volgende interactie heeft de instructie dat, als een wielklem eenmaal is aangebracht, de autobezitter geen verhaal meer heeft.

Situatie
We zien een medewerker van de Parkeerwacht die de laatste hand legt aan het aanbrengen van een wielklem. De bezitter van de auto komt terug, ziet het tafereel en reageert met persoonlijke verwijten en verdachtmakingen, om te eindigen met: 'en nu die klem eraf'!

Acties
1. Ja, scheld me maar verrot, dat ben ik wel gewend, maar die klem blijft erop.
2. Doet u dat thuis ook zo? Lekker is dat. Die klem blijft erop, meneer! En nou ga ik ú wat vertellen als u hier weg wil tenminste.
3. Sorry hoor, ik kan er ook niks aan doen, dat zijn instructies, ik mag de klem er niet af halen.
4. Ik begrijp best dat u van de kook bent, maar ik kan de klem niet eraf halen. Als u nu even naar mij wilt luisteren?

Opdracht
Geef voor elk van de acties aan op een schaal van - - - 0 + + + hoe effectief u elk van de acties vindt.
Of: Geef voor elk van de acties aan op een schaal van - - - 0 + + + met welke waarschijnlijkheid u een soortgelijke reactie zou vertonen.

Kader 1. Voorbeeld van een SJT (naar Van der Maesen, Westen & De Veer, 2000)

problemen dat een SJT presenteert, zou heel anders zijn dan bij problemen die in een klassieke intelligentietest of in een onderwijsleersituatie gepresenteerd worden. Volgens andere auteurs is er geen apart construct nodig dat aan de basis zou liggen van prestaties op SJT's. Zij gaan ervan uit dat SJT's functiekennis meten. Beide standpunten komen hieronder aan de orde. Ook passeren resultaten van empirisch onderzoek de revue die iets zeggen over de constructvaliditeit van SJT's.

SJT's meten tacit knowledge
Een psycholoog die reeds geruime tijd het oplossen van common-sense-problemen bestudeert is Robert Sternberg (Sternberg, 1997; Wagner & Sternberg, 1986; Sternberg, Wagner, Williams & Horvath,1995; Sternberg & Grigorenko, 2001). Sternberg et al. veronderstellen dat er een bijzondere manier van (leren) kennen geïmpliceerd is die ze aanduiden als *tacit knowledge*. Kenmerkend voor dit soort kennis is dat deze:
- niet (of moeilijk) te expliciteren (te beschrijven en over te dragen) is, omdat het om intuïtieve of voorbewuste kennis gaat. Bijvoorbeeld: de manier waarop je iets in een organisatie (maar ook in een cultuur die je niet kent) gedaan krijgt is een kwestie van *feeling*. Je krijgt deze kennis niet voorgeschoteld, maar leert deze door ervaring op te doen;
- gaat over procedures die met een bepaalde kans tot een resultaat leiden in plaats van over feiten die waar zijn of niet (procedurele in plaats van declaratieve kennis);
- gekoppeld is aan het bereiken van praktische doelen (waar declaratieve kennis een bepaalde stand van zaken betreft) in probleemsituaties die vaag (*ill-defined*) zijn;
- geleerd wordt door het zelf opdoen van ervaring in combinatie met observeren van rolmodellen (in plaats van declaratieve kennis die – volgens Sternberg - vaak boekenwijsheid is).

SJT's meten functiekennis
In tegenstelling tot wat Sternberg beweert is volgens Schmidt en Hunter (1993) het postuleren van een aparte kenvorm – zoals tacit knowledge – overbodig. Volgens hen is een SJT een test voor het meten van *job knowledge*. Het zou dus niet gaan om een aparte vorm van kennen, maar om kennis van een (al dan niet) functiespecifieke inhoud. De kwaliteit van die kennis is een functie van iemands intelligentie en de hoeveelheid opgedane ervaring in die functie (aantal jaren). Hoe intelligenter iemand is, hoe meer die persoon kan profiteren van zijn ervaring. Functiekennis correleert hoger met functiesucces dan intelligentie (Schmidt, Hunter, Outerbridge & Goff, 1988). Op de vraag of een SJT functiekennis meet, komen we nog terug in de laatste paragraaf.

Constructvaliditeit ter ondersteuning van de posities van tacit knowledge en van functiekennis
Sternberg en co-auteurs vinden in het algemeen geen hoge correlaties van hun Tacit Knowledge Tests en scores op intelligentietests (Sternberg & Grigorenko, 2001). De auteurs zien dat als empirische steun voor hun stelling dat er naast analytische ook praktische intelligentie bestaat. Wat wel opvalt is dat Sternberg et al. vaak Yale-studenten en managers als onderzoekspersonen gebruiken. Dit roept de vraag op in hoeverre *restriction of range* van intelligentietestscores wellicht verantwoordelijk is voor de lage correlaties.

Een recente meta-analyse van SJT-validiteiten laat een correlatie tussen SJT's en analytische intelligentietests (*g*) van .46. zien, hetgeen tamelijk hoog is (McDaniel et al., 2001). Zowel video-SJT's (Salgado & Lado, 2000; Weekley & Jones, 1997) als schriftelijke SJT's (Chan & Schmitt, 1997; Weekley & Jones, 1999) hebben een correlatie met intelligentiematen.

Sternberg et al. (1995) hebben aangetoond dat praktische intelligentie (gemeten met hun tests) toeneemt met de jaren, terwijl bij analytische intelligentie het omgekeerde het geval is. De kwestie of de kenvorm bepaalt of er sprake is van een apart construct achter SJT's is naar ons oordeel nog onbeslist.

SJT's zijn instrumenten die behoren tot de arbeidsproeven. Ze simuleren praktijksituaties en reacties in die situaties. Bij arbeidsproeven spelen altijd twee zaken een rol: de levensechtheid en de predictieve validiteit van de arbeidsproef. Levensechtheid is een aspect dat met de inhoudsvaliditeit en aan de indruksvaliditeit *(face validity)* van tests is verbonden.

Hoewel die vormen van validiteit van belang zijn voor de acceptatie van de test door kandidaat en opdrachtgever, en daardoor tevens voor het succes van de test als commercieel product, hoeft inhoudsvaliditeit niet per se altijd de predictieve validiteit te bevorderen. Het predictief succes van de klassieke intelligentietest bewijst dat.

Motowidlo et al. (1990, 1993) hebben zich met de vraag beziggehouden hoever men kan gaan in het opofferen van levensechtheid (*fidelity*) om toch nog een behoorlijk niveau van predictieve validiteit over te houden. Aan de ene kant staan instrumenten met een hoge levensechtheid (*high fidelity*) van situaties en acties. Deze simulaties benaderen het criterium qua levensechtheid, complexiteit en dynamiek het meest. Het realiseren van de wens om de werkelijkheid te benaderen gaat echter gepaard met hoge kosten, zoals het geval is met het assessment center (AC). Het dynamische element van acties (bijvoorbeeld het spel van woord en wederwoord met een rollenspeler in een AC) heeft nog meer keerzijden. Het kan ten koste gaan van standaardisering. Niet iedereen ontvangt dezelfde stimulus. Bovendien zijn er beoordelaars nodig, en liefst een flink aantal, om de noodzakelijke betrouwbaarheid op te leveren (Hofstee, 1999).

Het produceren van een actie in de open antwoordvorm kan gebeuren op drie manieren:

1. Beschrijven van een gedragsintentie (de vraag is: 'hoe zou je het doen?'). Zo gebeurt het bijvoorbeeld in het zogenaamde situationele interview.
2. Beschrijven of uitvoeren van gedrag ('wat is jouw eigen onmiddellijke reactie?').
3. Als 2, maar nu dynamisch: woord en wederwoord.

Aan de andere kant van het continuüm bevinden zich meer abstracte representaties, zoals verbale omschrijvingen van werksituaties (*low fidelity*). De werkelijkheidsgetrouwheid wordt nog geringer als de test aan de responskant van het gesloten type is. Bij een gesloten antwoordvorm zijn de acties al gegeven en van de persoon wordt een beoordeling op een schaal verlangd (zoals bij SJT's). Wat echter achterwege blijft is de productie van een actie.

SJT's zijn voorbeelden van *low fidelity* simulatie-instrumenten. Situaties worden als tekst op papier of op een monitor aangeboden en de reactie is een rangordening of effectiviteitsbeoordeling van eveneens tekstueel voorgelegde acties. Ook zogenaamde situatie-responsevragenlijsten zijn van dit type (Born, 1995). Motowidlo et al. (1990, 1993) noemen dit type instrumenten *situational inventories*.

Prestatie versus stijl

Bij klassieke intelligentietest gaat het om testitems waarbij er sprake is van een objectief correct antwoord, en het is de taak van de kandidaat zo goed mogelijk te presteren, dat wil zeggen zoveel mogelijk opgaven goed te beantwoorden. We spreken hier van een *maximum performance test*. Met een persoonlijkheidsvragenlijst worden geen prestaties gemeten, maar wordt de typische manier van doen of stijl van de persoon beoordeeld. De opdracht van de kandidaat is hier om te beschrijven hoe hij of zij zich doorgaans gedraagt. Vragenlijsten van dit type worden daarom *typical-performance*-instrumenten genoemd. Het zijn nadrukkelijk geen tests, omdat ze geen prestaties meten, maar iemands stijl in het gedrag (Hofstee, 1999). Alhoewel het 'goede' antwoord intersubjectief bepaald wordt, behoren SJT's duidelijk tot het eerste type. De kunst voor kandidaten is om de situatie zo goed mogelijk in te schatten. Hun succes wordt daarbij uitgedrukt in de overeenstemming van het gegeven oordeel met expertoordelen of *common sense*.

Een groot voordeel van *maximum performance*-tests boven *typical performance*-tests is dat het probleem van sociale wenselijkheid geen rol speelt. Integendeel, het is juist de bedoeling dat de kandidaat datgene wat in de aangeboden situaties als sociale norm geldt zo accuraat mogelijk voorspelt. Dat wordt immers weerspiegeld in de norm die gebaseerd is op de consensus van een groep experts of leken.

Constructie en scoring van SJT's

Constructie

Situaties in SJT's kunnen van velerlei aard zijn, al naargelang de invalshoek die de ontwikkelaar heeft gekozen. De situaties kunnen betrekking hebben op tamelijk verschillende taken als het beslissen over investeringen, het plannen en organiseren, het oplossen van conflicten, het overhalen tot het doen van een aankoop, en het brengen van een moeilijke boodschap door een leidinggevende aan een medewerker. Het is vaak de bedoeling van de ontwikkelaar van SJT's om langs de weg van het kiezen van de inhoud van de SJT de meting te richten op een specifiek construct. Bijvoorbeeld: door situaties op te nemen die gaan over interacties tussen een adviseur en klanten hoopt de ontwikkelaar de test te kunnen richten op bepaalde competenties (klantgerichtheid, commerciële vaardigheid).

Bij de constructie van SJT's wordt vaak gebruik gemaakt van de kritieke-incidentenmethode (Flanagan, 1955). Deze houdt in dat ervaren mensen in de functie (Amerikaans: SME's, *subject matter experts*) functierelevante situaties beschrijven die moeilijk zijn, die tot schade kunnen leiden, en waarin het verschil tussen goede en minder goede medewerkers tot uiting komt. Deze situaties dienen vervolgens als basis voor de constructie van de testitems. Vaak worden de kritieke situaties verzameld door middel van mondelinge interviews met SME's.

Als een ontwikkelaar geen specifiek construct op het oog heeft, zal hij dergelijke interviews zonder veel structuur afnemen en de inhoud van de SJT baseren op wat hij van de expert aangedragen krijgt. In dat geval zal een sterk heterogeen samengestelde SJT ontstaan. Het is mogelijk dat in dat geval vele constructen ten grondslag liggen aan de overall-score op de SJT. Als de ontwikkelaar van de SJT wel een bepaald construct op het oog heeft, zal de ontwikkeling van een SJT een sterk gestructureerd verloop hebben en zullen de situaties betrekking hebben op het betreffende construct.

Scoring

In SJT's is het de bedoeling dat van acties de effectiviteit wordt beoordeeld. De wijze waarop de beoordeling wordt gedaan

varieert. Voor de gesloten antwoordvorm zijn we drie typen beantwoordingsinstructies tegengekomen, namelijk:

1. gedwongen keuze: geef aan wat de meest effectieve en wat de minst effectieve actie is
2. rangorden de acties naar effectiviteit
3. beoordeling op Likertschaal: beoordeel elke actie op effectiviteit.

Gebruik van de gedwongen keuzebeantwoording is niet aan te bevelen omdat die een minder betrouwbare score zal opleveren dan de Likertschaalbeantwoording (zie ook Legree, 1995). Bij die laatste antwoordwijze worden alle acties behorend bij een situatie beoordeeld, bijvoorbeeld op een vijfpuntsschaal. Dit betekent dat er evenveel items als acties zijn. Betrouwbaarheid hangt samen met het aantal items en dat is al snel groter bij Likert-antwoorden dan bij gedwongenkeuze-antwoorden. Hetzelfde argument geldt ook ten aanzien van de rangordeningsvorm. Betrouwbaarheid is een noodzakelijke voorwaarde voor validiteit. Bij meta-analyses naar validiteit van SJT's zou de manier van beantwoorden dan ook een interessante moderator van validiteitscoëfficiënten kunnen zijn.

Hoe worden de antwoorden vervolgens gescoord? We beperken ons tot de scoring van Likert-antwoorden. Antwoorden op een SJT zijn, in tegenstelling tot antwoorden op items van een (intelligentie)test, niet objectief goed of fout. De 'ware' effectiviteit van acties is een kwestie van subjectieve beoordeling. Daarbij wordt allereerst vaak gebruik gemaakt van een *expertbenadering*. Hierbij laat men ervaren werknemers uit organisaties als beoordelaars optreden. Zij worden geacht te beschikken over de vaardigheid die met de SJT gemeten wordt. Problematisch kan de vraag zijn welke superbeoordelaar of -beoordelaars deze experts op zijn of hun beurt moet(en) selecteren. Om een behoorlijke beoordelaarsbetrouwbaarheid te verkrijgen zou het aantal expert-beoordelaars behoorlijk groot moeten zijn.

Een andere benadering gaat uit van de consensus van een grote groep niet-experts. Legree (1995) en Hofstee en Schaapman (1990), bijvoorbeeld, maken duidelijk dat een consistente schatting van expertkennis kan worden verkregen door de beoordelingen van een grote groep niet-experts te middelen. Op die manier komt de gemeenschappelijke variantie bovendrijven en wordt de unieke variantie uitgefilterd, conform principes van de klassieke testtheorie (wet van de grote getallen, gevat in de Spearman-Brownfunctie; zie ook Hofstee, 1999). Het is bij uitstek de *common sense*-benadering. Zoals we hierboven reeds bespraken beschouwen Sternberg et al. praktische intelligentie als het voorspellen van *common sense*.

Naast de kwestie wie er als beoordelaars moeten optreden moet de constructeur van de SJT bepalen hoe de individuele oordelen en die van de beoordelaars met elkaar worden vergeleken en in een maat worden uitgedrukt.

Er zijn verschillende manieren om Likert-oordelen te scoren. We noemen er twee. De eerste is afstandsscoring: voor elk item dient de absolute afstand te worden bepaald tussen het oordeel van het individu en het gemiddelde expertoordeel. Vervolgens worden de afstanden over items gesommeerd. (Het woord 'expert' kan opgevat worden in een van de beide hierboven beschreven betekenissen.) De tweede is proportionaliteitsscoring: een oordeel krijgt een gewicht dat overeenkomt met de proportie experts dat ditzelfde oordeel heeft gekozen.

Eigen onderzoek (Van der Maesen de Sombreff, 2002) van antwoorden van 180 personen op een multimedia-SJT (SQ-test, 72 items) liet zien dat de afstandsscoring (op basis van oordelen van zeven experts, te weten psychologen en personeelsadviseurs) een correlatie heeft van -.93 met proportionaliteitsscoring (waar de proporties zijn berekend op antwoorden van de hele groep van 180 personen). Het lijkt er dus op dat de twee methoden (de ene gebaseerd op de oordelen van weinig experts die behoorlijk met elkaar overeenstemmen, de andere op de oordelen van vele niet-experts die minder sterk met elkaar overeenstemmen) voor elkaar substitueerbaar zijn.

De correlatie tussen afstandsscoring op basis van gemiddelde expertoordelen en afstandsscoring op basis van gemiddelde oordelen van de hele groep respondenten was .91 (Van der Maesen, 2001). Dit resultaat bevestigt de voorspelling van Legree en Hofstee hierboven.

Het gebruik van consensus als norm voor scoring, impliceert dat conformering aan de status quo loont. Zeker in organisaties die in dynamische omgevingen opereren lijkt het van belang zowel bij de constructie van alternatieven als bij de beoordeling uit te gaan van het gedrag dat in de toekomst vereist zal zijn. In sommige situaties zal een open antwoordformaat te verkiezen zijn omdat dat formaat toelaat dat creatieve uitingen op hun merites worden beoordeeld.

Psychometrische merites van SJT's

In deze paragraaf behandelen we achtereenvolgens betrouwbaarheid en (toegevoegde) predictieve validiteit. Daarna gaan we in het kort in op enkele variabelen die van invloed kunnen zijn op de predictieve validiteit van SJT's.

Betrouwbaarheid

Er zijn verschillende vormen van betrouwbaarheid die relevant zijn voor SJT's. In de eerste plaats is de *overeenstemmingsbetrouwbaarheid* van beoordelingen door experts relevant. In SJT's wordt de scoring van antwoorden gebaseerd op het kwaliteitsoordeel dat beoordelaars over die antwoorden hebben gegeven. SJT's lijken wat dit aspect betreft op AC's. Het verschil zit erin dat de assessoren in een SJT zich maar één keer hoeven uit te spreken over de kwaliteit van antwoorden. Daarna worden die beoordelingen ingeblikt in de rekenregels en consistent toegepast. Een manier om overeenstemming te berekenen is coëfficiënt alpha, waar beoordelaars behandeld worden als items. Alpha is hoger naarmate het aantal beoordelaars groter is en naarmate hun beoordelingen meer met elkaar overeenkomen. Verwijdering van beoordelaars met afwijkende profielen zal de alpha-coëfficiënt verhogen. Nadeel van alpha is dat die maat niet signaleert dat er systematische verschillen tussen beoordelaars in mildheid of strengheid kunnen zijn.

Een tweede vorm van betrouwbaarheid is *hertestbetrouwbaarheid*. Het valt te verwachten dat de standaardisatie van opdrachten en rekenregels in SJT's gunstig uitwerkt op de betrouwbaarheid. Wat dit aspect betreft wint een SJT het waarschijnlijk van AC-oefeningen waarin meestal niet dezelfde rolspelers en assessoren optreden.

Een derde vorm van betrouwbaarheid heeft te maken met hoe goed de items die deel uitmaken van de test één en dezelfde dimensie meten. Het gaat om de *interne consistentie* van de test. Als de betrouwbaarheid van de test laag is, dan meet de test met de verschillende items steeds iets anders; de kans dat een persoon op een test die uit iets andere items bestond een heel andere uitslag zou hebben, is heel groot. Dit is onwenselijk. De betrouwbaarheid zegt iets over de uitwisselbaarheid van de test. Als je een SJT met andere items zou ontwikkelen op basis van dezelfde constructieprincipes dan zou die parallelle test naar verwachting tot vergelijkbare uitslagen van de personen leiden als de oorspronkelijke test. Er is ons geen onderzoek naar de interne consistentie van SJT's bekend.

Predictieve validiteit

Er is veel onderzoek verricht naar de predictieve validiteit van SJT's, met gunstige resultaten. Motowidlo et al. (1990, 1993) vonden ongecorrigeerde coëfficiënten in de buurt van de .30.

McDaniel et al. (2001) verrichtten een meta-analyse op 102 correlationele studies naar de voorspelling van personeelsbeoordelingen door SJT's en vonden een validiteit van .26. Na correctie voor attenuatie in verband met verschillen in betrouwbaarheid van het criterium was de coëfficiënt .34. De auteurs concluderen: 'This level of validity is comparable to such commonly used selection measures as assessment centers (Gaugler, Rosenthal, Thornton & Benson, 1987), employment interviews (McDaniel, Whetzel, Schmidt & Mauer, 1994), and biographical data measures (Rothstein, Schmidt, Erwin, Owens & Sparks, 1990).' (p. 736).

Salgado en Lado (2000) rapporteren op basis van een meta-analyse naar de validiteit van videotests gebaseerd op 12 studies een validiteit van .25 (na correctie .56). Videotests zijn te beschouwen als een vorm van SJT's. Daarnaast is het van belang de toegevoegde predictieve waarde aan te tonen boven traditionele intelligentietests. Immers, de beste predictor van werksucces, die bovendien tegen lage kosten te meten is, is de cognitieve capaciteitentest (Schmidt & Hunter, 1998). Bovendien lijken, zoals eerder gezegd, scores op SJT's redelijk samen te hangen met intelligentiescores.

Salgado en Lado (2000) vonden dat de validiteit van een cognitieve capaciteitentest en een videotest samen .10 hoger is dan de validiteit van een cognitieve capaciteitentest alléén. Ook Weekley en Jones (1999) lieten zien dat een SJT toegevoegde voorspellende waarde heeft ten opzichte van een cognitieve capaciteitentest. McDaniel et al. (2001) schatten dat de validiteit van de gecombineerde score van SJT en cognitieve capaciteitentest .40 is tegenover .32 à .34 voor alleen een cognitieve capaciteitentest.

Factoren van invloed op de predictieve validiteit

McDaniel et al. (2002) vonden dat SJT's die ontwikkeld waren op basis van een functieanalyse meer valide waren dan SJT's die dat niet waren. De coëfficiënten zijn .38 versus .29. Functieanalyse kan op verschillende manieren een rol spelen, in de eerste plaats natuurlijk bij het selecteren van situaties en bij het maken van de situatiescripts, en in de tweede plaats bij het selecteren en maken van scripts voor acties. In de derde plaats speelt functieanalyse ook een rol bij het maken van de rekenregels voor de scoring van antwoorden. Bij een functiegerichte aanpak zou men ervaren en succesvolle medewerkers kunnen betrekken als beoordelaars. Er is veel voor te zeggen om managers als experts voor de scoring in te zetten (zoals bij AC's) omdat managers (zouden moeten) weten wat (vanuit de bedrijfsstrategie) effectief is en wat niet en omdat ze bij uitstek de 'cultuurdragers' (zouden moeten) zijn.

Acceptatie van SJT's

Acceptatie van een test hangt samen met aspecten als de persoonlijk beoordeelde relevantie en rechtvaardigheid van de test en heeft betrekking op de maatschappelijke kwestie van scoreverschillen tussen allochtone en autochtone kandidaten. We behandelen deze twee punten.

Beoordeelde relevantie en rechtvaardigheid

Eén van de kenmerken waaraan SJT's hun populariteit ontlenen (tenminste in de VS) is de hoge indruksvaliditeit (*face validity*) van de methode. Wat de klaarblijkelijke relevantie en eerlijkheid van de methode betreft wint deze het gemakkelijk van de cognitieve capaciteitentest, maar ook van de persoonlijkheidsvragenlijst en zelfs van het AC. Indruksvaliditeit is vooral belangrijk in tijden van een krappe arbeidsmarkt waarin bedrijven grote moeite moeten doen om de aandacht te trekken van een slinkend aantal intreders. Chan en Schmitt (1997) lieten verder zien dat de acceptatie van de audiovisuele SJT nog weer significant groter is dan van de tekstuele variant.

Een aspect dat in relatie tot rechtvaardigheid aandacht behoeft is het feit dat SJT's niet waardevrij zijn. Ze kunnen bekritiseerd worden vanwege het feit dat ze de status quo bevestigen, doordat de *common sense* of de opvatting van de gevestigde orde als norm geldt voor de scoring. Personen met een deviant reactiepatroon kunnen zich benadeeld voelen vanuit de gedachte dat hun creatieve afwijkende oplossingen eveneens tot effectieve uitkomsten hadden kunnen leiden. In de praktijk zal dit doorgaans geen bron van grote weerstand bij kandidaten vormen, maar het is wel belangrijk om als beoordelaar dit feit bij duidelijk deviante scoorders in ogenschouw te nemen. Aan de andere kant is het onvermijdelijk dat de gevestigde orde een rol speelt bij het beoordelen van prestaties van mensen. Dat gebeurt bijvoorbeeld bij personeelsbeoordeling.

Scoreverschillen tussen groepen

De aandacht voor SJT's is ook gewekt omdat blijkt dat niet-blanke personen – in vergelijking met blanke personen – het

op SJT's veel minder slecht doen dan op een cognitieve capaciteitentest. De kwestie van scoreverschillen tussen bevolkingsgroepen die leiden tot gereduceerde kansen van minderheidsgroepen om aangesteld te worden (*adverse impact*) is in de VS een juridisch struikelblok voor selecterende organisaties. Waar het scoreverschil tussen autochtonen en allochtonen gemiddeld één standaarddeviatie is, kan dat bij SJT's gereduceerd worden tot onder de .50 (Motowidlo et al., 1990, 1993).

Conclusies over psychometrische merites van SJT's

Het lijkt goed mogelijk om betrouwbare expertoordelen voor SJT's te ontwikkelen. Over de interne consistentie van SJT's is nog weinig bekend.

Relevant is de vraag of een verbetering van de validiteit met .06 tot .10 die volgt uit het toevoegen van een SJT aan een cognitieve capaciteitentest, de moeite waard is. Bij 'waarde' gaat de aandacht vooral uit naar economisch rendement. Voor de beantwoording van deze vraag is meer informatie vereist: additionele kosten van een SJT, selectieratio, standaarddeviatie van prestaties uitgedrukt in euro's. Met behulp van het utiliteitsmodel van Brogden, Cronbach en Gleser is uit te rekenen dat deze schijnbaar geringe winst in validiteit een behoorlijk rendement kan opleveren. Een ander waardeaspect heeft te maken met de acceptatie van SJT's door kandidaten. Dit aspect kan ook invloed hebben op het rendement doordat de organisatie met de beter geaccepteerde selectieprocedure (mét SJT's) beter gekwalificeerde kandidaten aantrekt. De acceptatie wordt ook beïnvloed door de geringere scoreachterstand die allochtone kandidaten blijken te hebben op SJT's in vergelijking met die op andere tests, in het bijzonder op de intelligentietest (Hunter & Hunter, 1984).

Multimediale SJT's

Hieronder zullen we achtereenvolgens twee voorbeelden van SJT's presenteren: de SQ-test en de e-cartoon. Beide vallen in de categorie simulatie-instrumenten. Zowel situaties als reacties zijn audiovisueel. Moderne informatietechnologie maakt het mogelijk om SJT's in multimediavorm aan te bieden. Een voordeel van het multimediaal presenteren van SJT's is dat het mogelijk wordt naast de inhoudelijke boodschap ook aspecten als non-verbale informatie (gelaatsuitdrukkingen en andere lichaamstaal), paralinguale informatie (toonhoogte, intensiteit, snelheid van spreken) en contextinformatie (bijvoorbeeld de werkomgeving) over te dragen. De gepresenteerde werkelijkheid is rijker dan met woorden in tekstvorm kan worden overgedragen. De betere *fidelity* zal vermoedelijk ook leiden tot een hogere acceptatie dan de tekstuele varianten.

Onderzoek van Weekley en Jones (1997) en Salgado en Lado (2000) laat zien dat video-SJT's minder sterk dan *paper & pencil* SJT's correleren met scores op intelligentietests. Chan en Schmitt (1997) hebben bovendien aangetoond dat tekstuele SJT's een veel grotere scoreachterstand van allochtone kandidaten te zien geven dan audiovisuele SJT's, terwijl met beide methoden toch hetzelfde gemeten wordt.

SQ-test

Een test die maximaal gebruik maakt van ICT-mogelijkheden is de SQ-test (Van der Maesen de Sombreff, 2002; Born, Van der Maesen de Sombreff, Ruhe & Van der Zee, 2001). SQ staat voor *sociaal quotiënt*. Deze term is gekozen analoog aan de term die doorgaans gebruikt wordt voor het IQ (cognitieve capaciteiten).

De SQ-test is bedoeld om in de eerste fase van een selectieprocedure inzicht te verkrijgen in de sociale competenties van kandidaten. Om als eerste-fase-instrument geschikt te zijn moet de test voldoen aan eisen van gebruiksgemak en lage kosten. Aan deze eisen is bij de SQ-test voldaan door het instrueren, afnemen, scoren en rapporteren te automatiseren.

De SQ-test maakt een internetverbinding met online databases voor het verifiëren van logins, voor het opslaan van antwoorden en voor het scoren en rapporteren met online rekenregels en rapportagemodules. De beveiliging van de testdata is goed te waarborgen. De verwerking van de resultaten is geautomatiseerd waardoor de behaalde score van de kandidaat na afname van de test direct bekend is. Dit vormt een efficiencyvoordeel (Guion, 1998).

De SQ-test bestaat uit een serie problematische interacties (minstens vijftien), van het type waarvan aan het begin van dit artikel een voorbeeld is opgenomen. Elke situatie wordt voorafgegaan door een korte, mondeling gepresenteerde achtergrondinstructie. De hoofdpersoon uit de interactie laat vier oplossingen in echt gedrag horen en zien. De kandidaat beoordeelt elk van deze acties op effectiviteit. Er zijn vijf mogelijke beoordelingen op een Likert-schaal, op dezelfde wijze als in het voorbeeld aan het begin van dit artikel. Met het doornemen van instructie en oefenscènes neemt de test ongeveer één uur in beslag. Scènes en acties kunnen herhaaldelijk worden afgespeeld en beoordelingen kunnen worden aangepast. Overslaan van items is niet mogelijk.

Figuur 1 laat de testinterface zien: een SQ-scène met vier acties en beoordelingsschaal. Scène en acties zijn door klikken met de muis te activeren. Op dit moment zijn er twee standaard SQ-tests op de markt, de zogenaamde entry-level versie

Figuur 1. Testinterface SQ-test

gericht op hoger opgeleiden die de arbeidsmarkt betreden en een SQ-test voor leiding geven. Daarnaast zijn er maatwerkversies voor tandartsen en artsen ontwikkeld.

De SQ-test is gericht op het meten van sociale intelligentie, waarbij sociale intelligentie opgevat wordt als 'inzicht in de emoties, motieven, waarden en belangen van andere mensen en het in praktijk brengen van dat inzicht om doelen te bereiken'. Uitgangspunt is dat het inzicht in sociale situaties een noodzakelijke voorwaarde is voor het vertonen van effectief gedrag. Iemand die een sociale situatie niet doorziet zal zeker niet goed handelen. Het doel van de test is te voorspellen hoe succesvol een kandidaat zal zijn in het vertonen van sociaal effectief gedrag in het werk. De SQ-test veronderstelt geen specifieke voorkennis over de functie.

Als theoretische basis bij de ontwikkeling van actiealternatieven is gebruik gemaakt van een bekend model van interpersoonlijk gedrag, namelijk de 'Roos van Leary' (1957). Leary deelt sociale interacties in aan de hand van twee dimensies: *Control* en *Affiliation*. Gedrag op de *control*-dimensie wordt beschreven in termen van dominantie versus ondergeschiktheid; gedrag op de *affiliation*-dimensie in termen van vriendelijkheid versus vijandigheid. Kruising van beide assen resulteert in vier typen gedragingen: *proactief* (vriendelijk-dominant), *receptief* (vriendelijk-ondergeschikt), *defensief* (vijandig-ondergeschikt) en *offensief* (vijandig-dominant). In de SQ-test reflecteert elk van de vier reacties een van de vier kwadranten uit het model van Leary. Dit maakt het mogelijk om naast een terugkoppeling in termen van sociale intelligentie (of *tacit knowledge*) een terugkoppeling te geven over de door kandidaten meest geprefereerde stijl. Omdat de instructie de persoon ertoe aanzet een *maximum performance*-benadering te kiezen, is de stijlscore (*typical performance*) slechts een toegift. Die extra informatie kan voor een persoon wel nuttig zijn. Bijvoorbeeld: de *maximum-performance*-score is laag. Een hoge offensieve stijlscore verklaart vervolgens waarom de SQ-score achterblijft. Blijkbaar vindt de persoon offensief gedrag heel effectief, terwijl experts dat in het algemeen niet effectief vinden.

Voor de berekening van de score is gekozen voor afstandsscoring. Een rekenprogramma in de test vergelijkt voor elk van de beoordeelde acties het effectiviteitsoordeel van de kandidaat met het gemiddeld effectiviteitsoordeel van experts. De absolute afstandsmaten worden opgeteld tot een overallscore. Deze score wordt genormeerd. Hoe kleiner de som van de afstanden, hoe hoger de normscore van sociale intelligentie. Deze SQ-score wordt uitgedrukt op een decielschaal. Naast een algemene score rapporteert de SQ-test ook deelscores. Beoordelingen van scènes die qua inhoud bij elkaar horen worden tot één score gecombineerd. In de SQ-test voor entry-level-niveau worden bijvoorbeeld deelscores berekend voor omgang met collega's, omgang met klanten en omgang met leidinggevenden. In Figuur 2 wordt een deel van een rapportage van de SQ-test leiding geven gepresenteerd. De rapportage is via internet op te vragen.

Uit onderzoek naar de betrouwbaarheid van de SQ-test entry-level bleek de overeenstemming tussen beoordelaars hoog ($\alpha = .95$). Expertscores waren daarbij gebaseerd op de onafhankelijke beoordelingen van zeven ervaren psychologen en personeelsmanagers. De interne consistentie gebaseerd op een proefgroep van 181 personen was eveneens hoog ($\alpha = .89$). Omdat de deelscores van de SQ-test op minder antwoorden zijn gebaseerd dan bij de SQ-score het geval is, zijn de betrouwbaarheden van de deelscores lager dan van de SQ-score. De predictieve validiteit van de SQ-test entry-level wordt momenteel onderzocht.

E-cartoon

Een SJT kan ook in de wervingsfase een belangrijke rol vervullen. Het gebruik van internet om de student via een banensite kennis met de organisatie te laten maken wordt steeds populairder. Studenten worden uitgenodigd om zich op de banensite van het bedrijf in te schrijven en hun cv achter te laten. Een nuttige banensite heeft naast een wervende functie ook een selecterende functie. Om de werkzoekende te interesseren dient de website aantrekkelijk, uitdagend en onderscheidend te zijn van de rest. Tegelijkertijd is het voor de organisatie belangrijk om informatie met toegevoegde waarde van de werkzoekende te verkrijgen. Een aantrekkelijke banensite is naar onze mening interactief, prestatiegericht en geeft een snelle feedback aan de bezoeker. Een SJT die bestaat uit *real-life*-werksituaties vormt daarom een interessante mogelijkheid. Die SJT's doen beroep op algemene vaardigheden zoals conflicthantering, vaardigheden in overtuigen of onderhandelen. Mensen die op zoek zijn naar werk worden door de SJT uitgedaagd het oordeel van experts uit de betreffende organisatie over de verschillende reacties te voorspellen.

Het vormgeven van een SJT in een e-cartoon (cartoonanimaties met bewegende beelden en geluid) geeft die SJT's extra aantrekkingskracht omdat ze levensecht zijn en vooral de jeugd aanspreken.

Tegelijkertijd krijgen werkzoekenden een *realistic job preview* (Wanous, 1992). SJT's bieden immers de gelegenheid om te zien welke situaties ze in het werk kunnen tegenkomen en hoe reacties op die situaties binnen het bedrijf gewaardeerd worden. Werkzoekenden kunnen zelf bepalen of de situaties die ze er kunnen tegenkomen, hen wel aanstaan. Ze zullen zichzelf selecteren, op basis van motivationele match of mismatch.

Een groot voordeel voor het bedrijf is dat reeds in de wervingsfase interessante kandidaten geselecteerd worden op competentie en motivatie. Wanneer de testresultaten van de werkzoekende een bepaalde kritieke waarde overschrijden dan gaat – indien de kandidaat daarmee instemt – een e-mail met enkele persoonlijke gegevens naar de recruitment-adviseur en kan deze contact opnemen met de betreffende persoon. De persoon zelf krijgt na afloop van de test meteen bericht hoe goed hij of zij gepresteerd heeft en of er contact met hem of haar wordt opgenomen door de organisatie (snelle signalering en follow-up). Het nadeel van SJT's die via wervingskanalen worden aangeboden, is dat iemand anders dan de persoon die zijn naam eraan verbindt de antwoorden kan

	1	2	3	4	5	6	7	8	9	10
SQ-Score	■	■	■	■	■	■	■	■		

Deze persoon heeft een hoge score gehaald. Zijn beoordelingen komen goed overeen met de beoordelingen die ervaren leidinggevenden hebben gegeven van de situaties in deze test. Dat betekent dat hij in de positie van leidinggevende goed aanvoelt wat de emoties en motieven van medewerkers zijn dat hij een goed inzicht heeft wat in die situaties het te bereiken resultaat is. Deze persoon ziet goed in welk gedrag effectief is en welk niet om gewenste doelen te bereiken en hij ziet in dat de ene situatie een andere stijl van leidinggeven vereist dan een andere situatie.

	1	2	3	4	5
Aanspreken op resultaten	■	■	■	■	■

Deze persoon doorziet zeer goed situaties waarin een leidinggevende medewerkers moet aanspreken op hun voortgang en hun resultaten en weet zeer goed welk gedrag in deze situaties gegeven de omstandigheden het meest effectief is.

	1	2	3	4	5
Aanspreken op sociaal gedrag	■	■	■	■	■

Deze persoon doorziet zeer goed situaties waarin een leidinggevende medewerkers moet aanspreken op hun sociaal gedrag en hun attitudes en weet zeer goed welk gedrag in deze situaties gegeven de omstandigheden het meest effectief is.

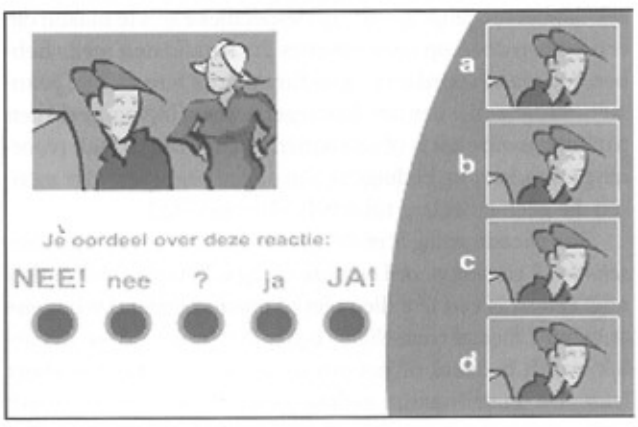
	1	2	3	4	5
Motiveren	■	■			

Deze persoon doorziet situaties waarin een leidinggevende medewerkers moet motiveren en inspireren duidelijk minder goed dan andere kandidaten en heeft minder inzicht dan andere kandidaten in welk gedrag in deze situaties gegeven de omstandigheden het meest effectief is.

	1	2	3	4	5
Coachen	■	■			

Deze persoon doorziet situaties waarin een leidinggevende medewerkers moet coachen duidelijk minder goed dan andere kandidaten en heeft minder inzicht dan andere kandidaten in welk gedrag in deze situaties gegeven de omstandigheden het meest effectief is.

Figuur 2. Rapportage (SQ-test leidinggeven)



Figuren 3 en 4. Interfaces van e-cartoon

hebben gegeven. Daarom is het ook nodig dat de organisatie aangeeft dat de persoon in de selectiefase nog een gecontroleerde beproeving te wachten staat.

Tevens krijgt de persoon informatie over de door hem geprefereerde interactiestijl. Dit maakt dat de test voor de persoon naast uitdagend ook leerzaam is. E-cartoons zijn gemakkelijk via internet te versturen. De scoring en rapportage vindt ook plaats via een webserver. Figuren 3 en 4 geven schermafdrukken van een e-cartoon. Als Likertschaal is een door Hofstee (1999) aanbevolen vorm gekozen.

Conclusies

Eén van de manieren waarop de psychologie haar nut kan bewijzen is het leveren van methoden om toekomstig arbeidsgedrag beter te voorspellen. Het lijkt erop alsof, meer dan vroeger, medewerkers naast cognitieve capaciteiten (IQ) sociale intelligentie (SQ) in hun 'pakket' moeten hebben om succesvol te zijn in het uitoefenen van hun beroep (Ferris, Witt & Hochwarter, 2001). Vanuit dat oogpunt beschouwd lijkt het interessant om aandacht te besteden aan de methode van SJT's. SJT's lijken geschikt om sociale competenties te meten. Er is ten opzichte van cognitieve capaciteitentests een duidelijk toegevoegde voorspellende waarde van SJT's en ook op acceptatie scoren SJT's goed. Bovendien lijken SJT's een goed alternatief te zijn voor klassieke tests waar het gaat om selectie van allochtone medewerkers. SJT's zijn op efficiënte wijze af te nemen, te scoren en te rapporteren.

Ondanks deze positieve geluiden, is er ook een aantal vraagtekens te plaatsen bij de methode. In de eerste plaats is het de vraag of SJT's *functiespecifieke voorkennis* vereisen. We hebben steeds beargumenteerd van niet. Maar is dit uitgangspunt terecht? Een SQ-test die selecteert voor de geneeskundeopleiding laat bijvoorbeeld een arts zien in overleg met een patiënt. Wanneer SJT's dit type functiespecifieke situaties afbeelden, is het de vraag of een kandidaat zich de competenties waarop de SJT beroep doet, al heeft kunnen eigen maken. Ze hebben immers geen on-the-job-ervaring opgedaan. Als Schmidt en Hunter (1993) gelijk hebben met hun stelling dat SJT's functiekennis meten, dan zouden SJT's niet gebruikt mogen worden als selectie-instrument. Wij denken zelf dat het mogelijk is om functiespecifieke SJT's te maken die een beroep doen op competenties die kandidaten reeds hebben opgedaan voordat ze in de functie zelf terecht zijn gekomen. Situaties die een arts kan tegenkomen, bijvoorbeeld een patiënt die onzeker is of een patiënt die uit zijn rol valt (voorrang eisen, korting bedingen), zijn ook te begrijpen voor mensen die geen opleiding tot arts hebben gevolgd.

Een tweede vraag is of SJT's gevoelig zijn voor *sociale wenselijkheid*. Het antwoord op deze vraag is bevestigend, sterker nog: een SJT is een test die meet of iemand inziet dat het ene antwoord sociaal wenselijker is dan het andere. Met 'wenselijk' wordt bedoeld of het antwoord tot een gewenst effect leidt. De mate waarin gedrag gewenst/effectief is, wordt geoperationaliseerd door meervoudige beoordelingen van experts. Of een effect wenselijk is, hangt op zijn beurt samen met de doelen die moeten worden bereikt. De prioriteit van doelen is situationeel en cultureel bepaald. Om een voorbeeld te geven: in de meeste gevallen is een boven-samen-communicatiestijl van leidinggevende jegens een medewerker effectief. Als de medewerker er echter de kantjes van afloopt kan een boven-tegen-benadering het meest effectief zijn. Bovendien zal binnen een sterk hiërarchische organisatie bovengedrag jegens een bovengeschikte minder sociaal wenselijk zijn dan in een plattere en informelere organisatie.

Een derde vraag die vaak gesteld wordt is of een SJT die *inzicht* meet iets kan zeggen over *handelen*. Er kunnen redenen zijn om inzicht niet te gebruiken. Bijvoorbeeld: iemand kan verlegen zijn. Dat betekent dat hij in Leary-termen een communicatiestijl hanteert waarin het aspect O (onder) overweegt. In persoonlijkheidstermen is deze persoon minder assertief (of scoort hij lager op een sociale invloed, wat een deelaspect is van extraversie, Big Five factor I). Een andere mogelijkheid is dat een persoon simuleert dat hij veel S (samen) in zijn communicatiestijl heeft, maar dat hij, als het er echt op aankomt, zijn eigen belang najaagt. Zo iemand simuleert (weer in termen van Leary) S maar is T (tegen, of, in persoonlijkheidstermen, hij scoort laag op vriendelijkheid/respect, Big Five factor II).

De voorbeelden hierboven maken duidelijk dat het hebben van inzicht niet per se resulteert in het gebruiken van dat inzicht. Toch heeft validiteitsonderzoek tot nu toe duidelijk gemaakt dat in het algemeen inzicht lijkt samen te gaan met handelen. De predictieve validiteit van SJT's is hoog. De criteria waarmee de SJT's zijn gecorreleerd, zijn meestal beoordelingen van succes in de functie. Men mag aannemen dat succes is af te lezen aan het effect van handelen en niet alleen aan inzicht.

Samenvattend lijken SJT's de aandacht te verdienen van selecterende instanties die de predictieve kracht van hun procedures willen vergroten, belang hechten aan het oordeel van hun kandidaten over de aantrekkelijkheid en rechtvaardigheid van de procedure en die werk willen maken van het aantrekken van allochtone kandidaten.

Dr. P. van der Maesen de Sombreff is werkzaam bij Van der Maesen Advies in Den Haag. Mw dr. M. Born is werkzaam bij de Vakgroep Psychologie aan de Erasmus Universiteit Rotterdam. Mw prof.dr. K. van Oudenhoven-van der Zee is werkzaam aan de Rijksuniversiteit Groningen. Mw drs. D. Ruhe is werkzaam bij de Belastingdienst in Utrecht. E-mail: <advies@vandermaesen.com>.

Literatuur

Born, M.Ph. (1995). *Het meten van prestatiegerichtheid. Een Situatie-Response vragenlijst.* Academisch proefschrift, Vrije Universiteit Amsterdam.

Born, M.Ph., Van der Maesen de Sombreff, P.E.A.M., Ruhe, D. & Van der Zee, K.I. (2001). *Validation of a multimedia Situational Judgment Test for social intelligence.* Paper presented at the HRM conference on organizational renewal: Challenging HRM. Nijmegen, The Netherlands, November 15.

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159.

Ferris, G.R., Witt, L.A. & Hochwarter, W.A. (2001). Interaction of social skill and general mental ability on job performance and salary. *Journal of Applied Psychology, 86,* 1075-1082.

Flanagan, J.C. (1955). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C. & Benson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493-511.

Guion, R.M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Erlbaum.

Hofstee, W.K.B. (1999). *Principes van beoordeling. Methodiek en ethiek van selectie, examinering en evaluatie,* Lisse: Swets & Zeitlinger.

Hofstee, W.K.B. & Schaapman, H. (1990). Bets beat polls: averaged predictions of election outcomes. *Acta Politica, 25,* 257-270.

Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance, *Psychological Bulletin, 96,* 72-98.

Leary, T. (1957). *Interpersonal diagnosis of personality.* New York: Ronald.

Legree, P.J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence, 28,* 544-566.

Lievens, S., Steverdinck, H., Tjoa, A. & Verhoeven, C. (1985). *Handleiding vragenlijst voor Commmercieel Inzicht,* Lisse: Swets & Zeitlinger.

Maesen de Sombreff, P.E.A.M. van der (2002). *Handleiding sq-test.* (Ongepubliceerde handleiding). Den Haag: Van der Maesen Advies.

Maesen de Sombreff, P.E.A.M. van der, Westen, G. & de Veer, J. (2000). Sociale intelligentie en multimedia. *Gids voor Personeelsmanagement, 79* (7/8), 38-43.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A. & Braverman, E.P. (2001). Predicting job performance using situational judgment tests: a clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

Motowidlo, S.J., Dunnette, M.D. & Carter, G.W. (1990). An alternative selection procedure: the low fidelity simulation. *Journal of Applied Psychology, 75,* 640-647.

Motowidlo, S.J. & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66,* 337-344.

Rothstein, H.R., Schmidt, F.L., Erwin, F.W., Owens, W.A. & Sparks, C.P. (1990). Biographical data in employment selection: can validities be made generalizable? *Journal of Applied Psychology, 75,* 175-184.

Salgado, J.F. & Lado, M. (2000). *Validity generalization of video tests for predicting job performance ratings.* Paper presented at 15th Annual Conference of SIOP, New Orleans, April 14-16.

Schmidt, F.L., Hunter, J.E., Outerbridge, A.N. & Goff, S. (1988). Joint relation of job experience and ability with job performance: test of three hypotheses, *Journal of Applied Psychology, 73, 1,* 46-57.

Schmidt, F.L. & Hunter, J.E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 8-9.*

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of reseach findings. *Psychological Bulletin, 124,* 262-274.

Sternberg, R. J. (1997). Tacit knowledge and job success. In N. Anderson & P. Herriot (Eds.), *International handbook of selection and assessment* (p. 200-213). London: Wiley.

Sternberg, R.J., Wagner, R.K., Williams, W.M. & Horvath, J.A. (1995). Testing common sense. *American Psychologist, 50,* 912-927.

Sternberg, R.J. & Grigorenko, E.L. (2001). Practical intelligence and the Principal. Unpublished Report www.temple.edu/LSS/pubseries2001-2.pdf, Yale University.

Wagner, R.K. & Sternberg, R.J. (1986). *Tacit Knowledge Inventory for Managers, Research Instrument, Form A.* The Psychological Corporation.

Weekley, J.A. & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50,* 25-49.

Weekley, J.A. & Jones, C. (1999). Further studies of situational tests, *Personnel Psychology, 52,* 679-700.

Wanous, J.P. (1992). *Organizational entry, recruitment, selection, orientation and socialization of newcomers.* Mass.: Addison-Wesley.

Summary

Spotlight on situational judgment tests

P. van der Maesen de Sombreff, M. Born, K. van Oudenhoven-van der Zee, D. Ruhe

Situational judgment tests (SJT) recently have received much attention. Research into the validity of SJTs has reported a high predictive validity, a considerable incremental validity, a high acceptability score of SJTs and less adverse impact for applicants from minority groups. This article describes the characteristic features of SJTs and how answers are scored. It describes the psychometrical merits of SJTs. Special attention is devoted to multimedia SJTs. Some critical questions about SJTs are considered.