

Veel opleidingen beschouwen niet-cognitieve eigenschappen als relevant voor de studie en het toekomstige werkveld van studenten. Denk aan goede communicatieve vaardigheden, empathisch vermogen en een professionele houding. Die aandacht voor de niet-cognitieve component is ook steeds vaker terug te zien in de selectieprocedures van geneeskundeopleidingen. Wendy de Leng deed onderzoek naar de geschiktheid van een situationele beoordelingstest in dat selectieproces.

ONTWIKKELING EN TOEPASSING VAN EEN SITUATIONELE BEOORDELINGSTEST VOOR HET METEN VAN INTEGRITEIT

TOELATING VAN GENEESKUNDE- STUDENTEN

INTRODUCTIE

Sinds de invoering van de wet Kwaliteit in Verscheidenheid mochten opleidingen met een numerus fixus hun studenten niet meer toelaten op basis van loting, maar alleen door middel van selectie. Hierdoor staan de toelatingscommissies van deze opleidingen voor de taak om hun selectieprocedures zo in te richten dat de juiste student op de juiste plek terecht komt. Het afgelopen decennium zijn in Nederland diverse (promotie)onderzoeken uitgevoerd naar methoden voor het selecteren van studenten (De Visser, 2018; Lucieer, 2016; Niessen, 2018; Schreurs, 2020; Schripsema, 2017; Urlings-Strop, 2018; Wouters, 2017).

Eén aanpak is om de selectieprocedure te richten op een verscheidenheid aan eigenschappen die relevant zijn voor de betreffende opleiding. De selectie van studenten voor toelating tot het hoger onderwijs richt zich veelal op cognitieve eigenschappen, zoals intelligentie en studievoordigheden. Methoden voor het meten van deze eigenschappen zijn vaak gemakkelijk voorhanden (bijvoorbeeld het gemiddeld vwo-cijfer) en bewezen effectief in het voorspellen van toekomstige studieprestaties. Veel opleidingen beschouwen tevens niet-cognitieve eigenschappen als relevant voor de studie en het toekomstige werkveld van studenten (Duffy, 2011). Deze eigenschappen omvatten onder andere kenmerken zoals goede communicatieve vaardigheden, empathisch vermogen en een professionele houding. Deze relevantie van niet-cognitieve eigenschappen is ook

kenmerkend voor de opleiding geneeskunde, vanwege het directe contact dat een arts heeft met patiënten en hun naasten. Vandaar dat in veel geneeskunde curricula aandacht is voor aspecten zoals gespreksvoering en professionele ontwikkeling.

De aandacht voor de niet-cognitieve component van het artsenvak is ook steeds vaker terug te zien in de selectieprocedures van geneeskundeopleidingen. Zo vragen veel opleidingen aan hun kandidaten om een motivatiebrief te schrijven. Motivatiebrieven blijken echter weinig waardevol voor het maken van een zinvol onderscheid tussen kandidaten. Een andere veelgebruikte methode voor het in kaart brengen van niet-cognitieve eigenschappen is het interview, maar ook dat blijkt vaak ongeschikt voor selectie, omdat het moeilijk is om iedere kandidaat op dezelfde objectieve manier te beoordelen. Het structureren van een interview, bijvoorbeeld met behulp van meerdere korte rondes en meerdere interviewers, kan de waarde van het interview verhogen. Een nadeel hiervan is dat het voeren van meerdere gestructureerde interviews bij een groot aantal kandidaten, zoals voor de geneeskundeopleiding (circa 1000 kandidaten), tot een zeer arbeidsintensieve en kostbare procedure leidt (Siu & Reiter, 2009).

Een selectie-instrument dat al langere tijd wordt toegepast binnen personeelsselectie en voor de selectie van geneeskundestudenten buiten Nederland is de situationele beoordelingstest (*Situational Judgement Test*; *SJT*). Een *SJT*

Integriteitstests blijken na algemene intelligentietests de meest effectieve voorspeller te zijn van toekomstige werkprestaties

beschrijft aan de kandidaat een aantal lastige situaties in een relevante context, zoals de geneeskundeopleiding (zie voorbeelditems in kader 2). Deze scenario's worden gevolgd door mogelijke reacties op iedere situatie en de taak van de kandidaat is om de gepastheid van deze reacties te beoordelen gegeven de situatie (Weekley & Ployhart, 2006). De resultaten van eerdere studies wijzen op de voorspellende waarde van SJT's voor toekomstige werk- en studieprestaties (Lievens, 2013), vermoedelijk veroorzaakt door de coherentie tussen de testinhoud en context. Deze coherentie leidt er waarschijnlijk ook toe dat kandidaten overwegend positief reageren op de SJT als selectiemethode (Luschin-Ebengreuth et al., 2015). Verder laten SJT-scores kleinere etnische en socio-economische subgroepverschillen zien dan scores op traditionele cognitieve tests. Hierdoor kunnen SJT's mogelijk bijdragen aan een diverse populatie geneeskundestudenten (Lievens et al., 2016). Tot slot kan een SJT op gestandaardiseerde wijze afgenomen worden bij grote groepen kandidaten. Vanwege deze kwaliteiten heb ik in mijn proefschrift (De Leng, 2019) onderzocht of de SJT een geschikte methode is voor de selectie van geneeskundestudenten in Nederland.

De SJT in mijn proefschrift richt zich op het meten van integriteit, een eigenschap die getypeerd wordt door onder andere de aanwezigheid van karaktertrekken, zoals oprechtheid en eerlijkheid, en de afwezigheid van egocentrische attitudes, gedachten en overtuigingen. Integriteit maakt vaak deel uit van raamwerken en modellen die eigenschappen beschrijven die voor artsen zeer relevant zijn (bijv. Frank, Snell & Sherbino, 2015). Daarnaast beschouwen uiteenlopende medisch specialismen integriteit als één van de meest relevante eigenschappen (Patterson, Ferguson & Thomas, 2008). Integriteit wordt opgevat als een persoonlijkheidskenmerk (Lee et al., 2008) en daarmee als een relatief stabiele persoonlijkheidseigenschap. Integriteitstests blijken bovendien na algemene intelligentietests de meest effectieve voorspeller te zijn van toekomstige werkprestaties (Schmidt, Oh, & Shaffer,

2016). Deze kenmerken maken integriteit een interessante persoonlijkheidseigenschap voor selectie. Het gebruik van integriteit voor de toelating van adolescenten tot het hoger onderwijs vereist echter kritisch onderzoek, aangezien het hier niet gaat om personeelsselectie met economische utiliteit als doel, maar om de toelating van scholieren die het meest zullen profiteren van het aangeboden onderwijs. Bovendien laat de studie van Robins et al., (2001) zien dat, alhoewel persoonlijkheid in het algemeen stabiel is, er in de persoonlijkheid van jongvolwassenen kleine tot middelgrote veranderingen kunnen plaatsvinden. Onderzoek moet uitwijzen of het mogelijk is de selectie van geneeskundestudenten deels te baseren op niet-cognitieve eigenschappen zoals integriteit.

De introductie van een niet-traditionele test, zoals de SJT, in een selectieprocedure voor de geneeskundeopleiding vereist zorgvuldig onderzoek naar de kwaliteit van de test. Daarnaast is het waardevol te onderzoeken welke testkenmerken kunnen leiden tot een verbetering van de kwaliteit van de SJT, om zo de selectie van geneeskundestudenten te optimaliseren. Daarom zijn de volgende SJT-aspecten bestudeerd: ontwikkeling, scoring, verdraaiing, kandidaatreacties en voorspellende waarde. De resultaten van deze studies bieden meer inzicht in de toepassing van SJT's voor de selectie van geneeskundestudenten, zoals ik hieronder zal betogen.

ONTWIKKELING

Zoals reeds vermeld wordt de coherentie tussen de testinhoud en de relevante werk- of studiecontext gezien als een belangrijke bron van de goede voorspellende waarde van de SJT. De mogelijkheid om de test toe te spitsen op de eigen specifieke situatie is een van de belangrijkste voordelen van een SJT. Een nadeel van deze sterke coherentie met de context is echter dat het vaak onduidelijk is welke eigenschap met een SJT wordt gemeten, omdat realistische werk- en studietoepassingen vaak meerdere eigenschappen gelijktijdig aanspreken (Christian, Edwards & Bradley, 2010). Duidelijkheid over welke eigenschap met een test wordt gemeten, heeft een belangrijke meerwaarde voor selectie. Zo leidt die duidelijkheid tot meer inzicht in de positie van een test ten opzichte van de andere onderdelen van een selectieprocedure en biedt het kandidaten en toelatingscommissies meer transparantie wat betreft de eigenschappen waarop geselecteerd wordt (Chan & Schmitt, 2005).

Vanwege deze redenen is gestreefd naar het ontwikkelen van een constructgebaseerde SJT met een realistische inhoud voor het meten van integriteit bij kandidaten voor de

geneeskundeopleiding. Om dit doel te bereiken zijn twee ontwikkelmethoden gecombineerd: een empirische, inductieve methode en een theoretische, deductieve methode. De inductieve methode volgt de traditionele wijze van het ontwikkelen van een SJT op basis van zogeheten kritieke incidenten interviews (Flanagan, 1954). Bij deze interviews worden personen die nauw betrokken zijn bij de context waarvoor een SJT wordt ontwikkeld gevraagd naar concrete voorbeelden van uitzonderlijk goede en slechte gedragingen (zie kader 1). Deze kritieke incidenten vormen de input voor de scenario's en responsopties van de SJT en verhogen daarmee de realiteit van de test. Om de authenticiteit van de SJT te waarborgen hebben personen die betrokken zijn bij de beoordeling van het professioneel gedrag van geneeskundestudenten, medisch personeel en geneeskundestudenten bijgedragen aan de ontwikkeling van de test.

De deductieve methode is gebruikt om de testinhoud toe te spitsen op de eigenschap integriteit door twee bestaande integriteitgerelateerde instrumenten in te zetten om de ontwikkeling van de responsopties van de SJT te structureren. Als eerste is de Integriteit schaal van de HEXACO-persoonlijkheidsvragenlijst (zie bijv. De Vries & Born, 2013) gebruikt voor het creëren van geschikte responsopties (d.w.z. 'wat te doen'). Daarnaast is de *How I Think*-vragenlijst naar cognitieve verstoringen in het denken (Barriga & Gibbs, 1996), als uitgangspunt genomen voor het schrijven van ongeschikte responsopties (d.w.z. 'wat niet te doen').

De combinatie van deze twee ontwikkelmethoden heeft geresulteerd in de totstandkoming van de Integriteit SJT, bestaande uit 57 scenario's, elk gevolgd door vier responsopties, die beoordeeld dienen te worden op een zespuntsschaal lopend van 'zeer ongepast' tot 'zeer gepast' (zie kader 2 voor twee voorbeelditems). De SJT is voor onderzoeksdoeleinden afgenomen bij een groep aspirant-geneeskundestudenten samen met vier andere integriteit-gerelateerde vragenlijsten. De significante samenhang tussen de Integriteit SJT en de integriteitgerelateerde vragenlijsten duiden op een goede convergente validiteit van de SJT.

Een opvallende bevinding was dat de subscore op basis van de ongeschikte responsopties (gebaseerd op de *How I Think*-vragenlijst) een sterkere convergente validiteit had dan de subscore op basis van de geschikte responsopties (gebaseerd op de Integriteit schaal van de HEXACO). Deze bevinding is vermoedelijk te verklaren doordat het in ambigue situaties, zoals die omschreven in de SJT-scenario's, vaak duidelijker is wat men niet moet doen dan wat men wel moet doen. Dikwijls zijn er immers verschillende manieren waarop een situatie op een gepaste manier opgelost kan worden. De uiteindelijk gekozen weg is afhankelijk van de karakteristieken van de context en de persoon. Het op een gepaste manier oplossen van een conflict met een leidinggevende is bijvoorbeeld afhankelijk van de hiërarchie in een organisatie. Ongepaste reacties leiden daarentegen bijna altijd tot een negatieve uitkomst, waardoor de beoordeling

KADER 1. THEMA'S EN VOORBEELDEN UIT KRITIEKE INCIDENTEN INTERVIEWS

THEMA'S	VOORBEELDEN
Meeliftgedrag	Niets/weinig bijdragen aan een groepsopdracht.
Communicatie	Opgeven voor iets en dan niet komen opdagen.
Mailgedrag	Mailadres met een scheldwoord erin.
Tentamengedrag	Telefoon meenemen naar het toilet tijdens een tentamen.
Feedback	Niets doen met de feedback van een docent.
Plagiaat	Overnemen van het verslag van een andere student.
Fraude	Handtekening plaatsen op aanwezigheidslijst voor een ander.
Gedrag kleinschalig onderwijs	Medestudenten niet laten uitpraten.
Agressief of dreigend gedrag	Medestudenten of medewerkers uitschelden.
Liegen persoonlijke omstandigheden	Liegen over het overlijden van een naaste.
Geheimhouding verbreken	In de trein uitgebreid praten over patiënten.
Ongeordend	Onvoorbereid en zonder spullen naar de les gaan.

van dergelijke reacties minder afhankelijk is van de specifieke context- en persoonskenmerken (Stemler, Aggarwal & Nithyanand, 2016). Het toespitsen van de SJT op het vermogen om ongepaste responsies te herkennen kan daarom de potentie van de test om geschikte en ongeschikte kandidaten van elkaar te onderscheiden verbeteren.

SCORING

Een opvallend kenmerk waarin de SJT verschilt van traditionele, cognitieve selectie-instrumenten zoals kennistests, is de afwezigheid van eenduidig correcte antwoorden op de items. Een SJT toetst geen feitelijke kennis (Bergman et al., 2006). Om een SJT te kunnen scoren wordt gewoonlijk een scoringsleutel ontwikkeld op basis van de antwoorden van

KADER 2. VOORBEELDITEMS UIT DE INTEGRITEIT SJT

SCENARIO

Michelle, Tim en Sarah moeten samen een groepsopdracht maken. De groep maakt een eerste afspraak om te bespreken hoe ze de opdracht gaan aanpakken. Tijdens deze afspraak blijkt Sarah de opdracht al gemaakt te hebben. Sarah vertelt dat ze tijd over had en daarom de opdracht al heeft gemaakt, maar dat ze het niet erg vindt om haar werk te delen met Michelle en Tim. Tim is wel tevreden met deze afspraak.

Beoordeel voor elk van de onderstaande reacties hoe gepast ze zouden zijn voor Michelle (omcirkel s.v.p.)

		Zeer ongepast				Zeer gepast	
1.	Het aanbod van Sarah accepteren omdat dat makkelijkst is.	1	2	3	4	5	6
2.	Het aanbod van Sarah accepteren omdat het zonde is om dezelfde opdracht twee keer te maken.	1	2	3	4	5	6
3.	Sarah bedanken voor het aanbod en voorstellen het werk meer te verdelen.	1	2	3	4	5	6
4.	Samen met Tim het werk van Sarah aanvullen om toch een bijdrage aan de opdracht te leveren.	1	2	3	4	5	6

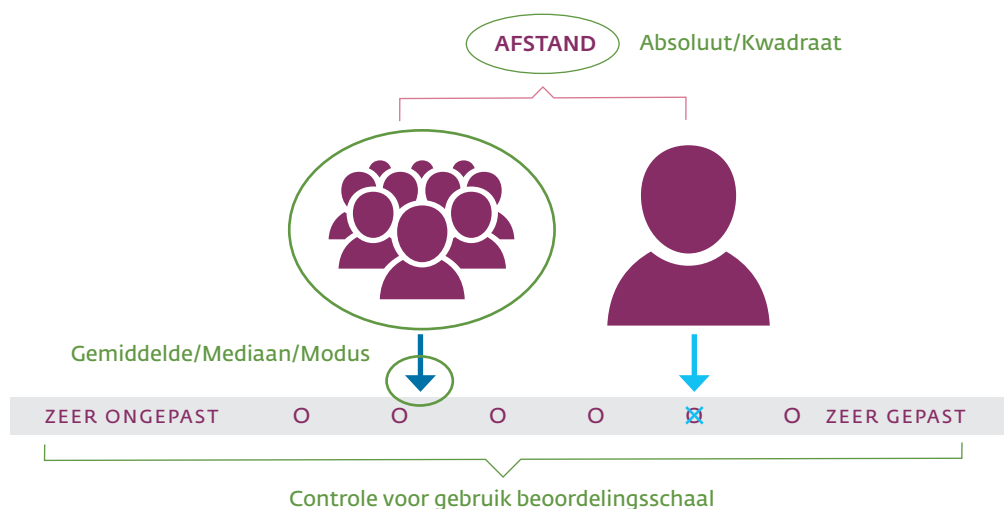
SCENARIO

Stacey hoort twee medestudenten praten over de cursus praktische vaardigheden van gisteren. Tijdens deze cursus moeten studenten oefenen met het uitvoeren van lichamelijk onderzoek op elkaar. De twee studenten roddelen over Maarten, een medestudent die een wijnvlek heeft over een groot deel van zijn lichaam.

Beoordeel voor elk van de onderstaande reacties hoe gepast ze zouden zijn voor Stacey (omcirkel s.v.p.)

		Zeer ongepast				Zeer gepast	
1.	Doen alsof ze niets hoort omdat het haar niet aangaat.	1	2	3	4	5	6
2.	Haar medestudenten vertellen dat ze hen hoort roddelen over Maarten.	1	2	3	4	5	6
3.	De studenten vertellen dat ze moeten stoppen met hun gesprek omdat ze de vertrouwelijkheidsregels schenden.	1	2	3	4	5	6
4.	Haar medestudenten niet aanspreken omdat zij haar anders vinden zeuren.	1	2	3	4	5	6

FIGUUR 1. SCORINGSMETHODEN



De horizontale balk representeert de zes-puntschaal waarop responsies beoordeeld moeten worden. Het antwoord van de kandidaat (lichtblauwe pijl) wordt vergeleken met het samengevatte antwoord van de referentiegroep (donkerblauwe pijl). Binnen deze methode zijn vier factoren (zie tekst) gevarieerd om 28 verschillende scoringsmethoden te verkrijgen.

een groep experts. De antwoorden van de experts vormen het referentiepunt waarmee de antwoorden van de kandidaten worden vergeleken. De literatuur toont aan dat deze vergelijking op veel verschillende manieren uitgevoerd kan worden en dat de wijze van scoring invloed kan hebben op de kwaliteit van de SJT (McDaniel et al., 2011). Desondanks is de keuze voor de scoringsmethode vaak een arbitraire. Om meer inzicht te krijgen in hoe de manier van scoren de testkwaliteit beïnvloedt, heb ik in mijn proefschrift 28 verschillende scoringsmethoden onderzocht.

Deze scoringsmethoden zijn gevormd door vier factoren te variëren. De *eerste factor* betreft de manier van controleren voor systematische fouten die ontstaan door individuele verschillen in het gebruik van beoordelingsschalen. Sommige mensen zijn meer geneigd om vooral het midden van een beoordelingsschaal te gebruiken (bijv. 'enigszins (on) gepast'), terwijl anderen juist meer geneigd zijn vooral de uitersten van een schaal te gebruiken (bijv. 'zeer (on)gepast'). Deze individuele voorkeuren zeggen vaak niets over de eigenschap die een test beoogt te meten, maar beïnvloeden wel de score en veroorzaken daarmee systematische onnauwkeurigheden in de testscore. Er zijn methoden beschikbaar om te controleren voor deze individuele voorkeuren in het gebruik van beoordelingsschalen, waarvan

er in deze studie twee zijn toegepast. De *tweede factor* is de referentiegroep waarmee de antwoorden van de kandidaten worden vergeleken, deze bestond uit experts (d.w.z. 13 personen betrokken bij het onderwijs en de beoordeling van professioneel gedrag van geneeskundestudenten) of niet-experts (d.w.z. de groep kandidaten zelf). De *derde factor* omvat de wijze waarop het referentiepunt wordt vastgesteld, waarbij het gemiddelde, de mediaan of de modus van de antwoorden van de referentiegroep wordt gebruikt. Tot slot betreft de *vierde factor* de afstand tussen het antwoord van de kandidaat en het referentiepunt, te weten absoluut of gekwadrateerd. Door te kwadrateren hebben antwoorden die sterker afwijken van het referentiepunt een grotere negatieve invloed op de score (zie figuur 1).

Deze scoringsmethoden zijn toegepast op dezelfde SJT en de resulterende scores zijn onderling vergeleken op betrouwbaarheid, constructvaliditeit en etnische subgroep-verschillen. De interne consistentie, berekend met coëfficiënt alfa, blijkt sterk afhankelijk van de scoringsmethode – wat vermoedelijk het gevolg is van het effect van de scoringsmethode op de scorevariantie. In de literatuur is veel kritiek op het gebruik van de traditionele coëfficiënt alfa (Cronbach's alfa) voor het schatten van de betrouwbaarheid van een SJT, vanwege de multidimensionaliteit van de test (Catano,

Het ultieme doel van het uitbreiden van een selectieprocedure met niet-traditionele selectiemethoden is het voorspellen van relevante toekomstige prestaties van kandidaten

Brochu & Lamerson, 2012). Ons resultaat onderschrijft de beperkte bruikbaarheid van coëfficiënt alfa, omdat deze sterk beïnvloed wordt door de vaak arbitraire keuze voor een scoringsmethode. De scoringsmethode had een minder sterke invloed op de constructvaliditeit (onderzocht aan de hand van correlaties met de persoonlijkheidsdimensies neuroticisme, altruïsme en consciëntieusheid) en op subgroepverschillen tussen kandidaten met een Nederlandse achtergrond en kandidaten met een niet-westerse migratieachtergrond.

De factor met de grootste invloed op de testkwaliteit was de manier van controleren voor individuele verschillen in het gebruik van een beoordelingsschaal. Een laatste opvallende bevinding was dat de referentiegroep geen verschillen in testkwaliteit veroorzaakte – een mogelijk gevolg van het gegeven dat de groep niet-experts groter en representatiever was dan de groep experts. De algemene conclusie van deze studie is dat de optimale scoringsmethode niet bestaat en dat het gebruik van een SJT voor de *high-stakes* selectie van geneeskundestudenten moet samengaan met een zorgvuldige, goed onderbouwde keuze voor de methode waarmee de test gescoord zal worden.

VERDRAAIING

Een veelgehoord kritiekpunt over het gebruik van niet-cognitieve instrumenten voor selectie is de vatbaarheid van deze instrumenten voor de opzettelijke verdraaiing van antwoorden. Daarmee proberen kandidaten bewust een betere indruk te maken tijdens de selectie en de kans op toelating te vergroten. Over de vraag of dit fenomeen, in het Engels *faking*, wel of niet problematisch is voor selectie bestaat in de literatuur veel controverse. Niettemin is er een aanzienlijke consensus dat mensen kunnen faken op niet-cognitieve instrumenten en dat de meeste mensen zullen faken in situaties waarin veel op het spel staat (Cook, 2016).

De capaciteit en motivatie om te faken is tevens aangetoond ten aanzien van SJT's, hoewel hun gevoeligheid voor *faking* kleiner lijkt dan die van persoonlijkheidsvragenlijsten. De mate van *faking* wordt sterk beïnvloed door de responsinstructies van een SJT. Als de kandidaat geïnstrueerd wordt de responsies te beoordelen op basis van wat

gedaan zou moeten worden – in plaats van wat hij/zij werkelijk zou doen – dan blijkt een SJT minder vatbaar voor *faking*. Deze kennisgebaseerde instructies – in tegenstelling tot gedragsmatige instructies – geven een SJT meer het karakter van een kennistest – en kennis is moeilijker om te faken (Nguyen, Biderman & McDaniel, 2005). Hoewel de kennisgebaseerde instructies inderdaad resulteren in een verminderde gevoeligheid voor *faking* in *low-stakes* situaties (bijv. loopbaanbegeleiding), blijkt het in *high-stakes* situaties (bijv. selectie) niet afdoende om *faking* op een SJT volledig uit te sluiten (Lievens, Sackett & Buyse, 2009).

Faking heb ik onderzocht door dezelfde SJT met kennisgebaseerde instructies tweemaal af te nemen bij dezelfde groep kandidaten, onder verschillende omstandigheden. De eerste afname (T1) was in een *low-stakes* situatie, namelijk tijdens een vrijwillige coachingsdag waarop kandidaten informatie ontvingen over de selectieprocedure, maar waar geen selectie plaatsvond. De tweede afname (T2) was tijdens een *high-stakes* situatie, namelijk tijdens de selectiedag waarop kandidaten studievoordigheidstoetsen moesten maken. Ondanks dat de SJT tijdens beide afnames alleen voor onderzoeksdoeleinden werd afgenomen, was de verwachting dat kandidaten zouden proberen een betere indruk te maken op T2, omdat er dan meer op het spel stond dan op T1.

Faking is als eerste onderzocht door te kijken naar een verandering in het gebruik van de beoordelingsschaal, omdat uit eerder onderzoek blijkt dat *faken* gerelateerd is aan een toename in het gebruik van de uitersten van een beoordelingsschaal (Van Hooft & Born, 2012). Daarnaast is gekeken naar het scoreverschil tussen T1 en T2, berekend met behulp van drie scoringsmethoden: één die niet controleert en twee die wel controleren voor systematische fouten veroorzaakt door individuele verschillen in het gebruik van beoordelingsschalen. Een hogere SJT-score op T2 dan T1 wordt beschouwd als een indicatie voor *faking*.

Kandidaten gebruikten meer extreme beoordelingsschaalpunten op T2 dan op T1, wat erop duidt dat kandidaten anders antwoorden op de SJT op T2. Of die toename in de gebruikmaking van extreme beoordelingsschaalpunten inderdaad leidt tot een toename in de SJT-score (indicatie van

faking), blijkt sterk afhankelijk van de gehanteerde scoringsmethode. Als de scoringsmethode controleert voor individuele verschillen in het gebruik van de beoordelingsschaal, dan wordt op T2 een hogere score waargenomen dan op T1, duidend op de aanwezigheid van faking. Een opvallende bevinding is dat, als de scoringsmethode *niet* controleert voor individuele verschillen in beoordelingsschaalgebruik, de resulterende SJT-score lager is op T2 dan op T1, duidend op de afwezigheid van faking. De resultaten laten dus zien dat dezelfde kandidaten in een high-stakes situatie anders reageren dan in een low-stakes situatie op dezelfde SJT met kennisgebaseerde instructies. Of dit verschil leidt tot een toename of afname in de SJT-score van T1 naar T2 hangt af van de wijze waarop de SJT wordt gescoord.

De voornaamste conclusie van deze studie is dus dat de aanwezigheid van faking op een SJT onontdekt kan blijven als bij het scoren van de test geen rekening wordt gehouden met individuele verschillen in het gebruik van een beoordelingsschaal.

KANDIDAATREACTIES

Zoals vermeld in de introductie draagt de verbinding tussen de testinhoud en de context eraan bij dat kandidaten positief reageren op de SJT. De reacties van kandidaten voor de geneeskundeopleiding op de Integriteit SJT waren echter niet eenduidig positief. Er was vooral kritiek op de vermeende vatbaarheid van de SJT voor faking.

Onderzoek naar de kandidaatreacties op de SJT is relevant omdat negatieve reacties onder andere kunnen leiden tot verminderde testmotivatie, de terugtrekking van mogelijk geschikte kandidaten (Chan et al., 1997) en een grotere kans dat kandidaten een klacht indienen of in beroep gaan tegen de selectiebeslissing. Daarnaast kunnen er demografische subgroep-verschillen in kandidaatreacties bestaan. Zo blijkt uit Amerikaans onderzoek dat etnische minderheidsgroepen soms negatiever reageren dan de etnische meerderheid op selectietests (Schmitt & Chan, 1999). Wanneer subgroepen verschillend reageren op een SJT, kan dit negatieve consequenties hebben voor een eerlijke selectie. Vandaar dat in deze studie reacties op de SJT zijn vergeleken voor subgroepen op basis van geslacht en etnische en socio-economische achtergrond.

De reacties op de SJT zijn uitgevraagd in een enquête onder kandidaten voor de geneeskundeopleiding. De SJT werd kort omschreven en samen met twee voorbeelditems gepresenteerd aan de kandidaten. In de voorbeelditems werden twee testkenmerken gemanipuleerd; enerzijds de

responsinstructies (kennisgebaseerd versus gedragsmatig) en anderzijds het responsformat (meerkeuze versus beoordelingsschaal). Van deze twee testkenmerken wordt aangenomen dat zij de gepercipieerde vatbaarheid van de SJT voor faking beïnvloeden. Gedragsmatige instructies worden beschouwd als ontvankelijk voor faking (zie de paragraaf Verdraaiing), evenals beoordelingsschalen die veel worden gebruikt in faking-gevoelige zelfrapportage-instrumenten, zoals persoonlijkheidsvragenlijsten. De verwachting is dat een SJT met deze kenmerken minder positieve kandidaatreacties ontvangt omdat eerder onderzoek heeft aangetoond dat kandidaten positiever reageren op selectiemethoden die moeilijk te faken zijn (Uggerslev, Fassina & Kraichy, 2012). Door de twee testkenmerken te manipuleren zijn vier verschillende condities gecreëerd, waaraan respondenten willekeurig toegewezen zijn.

Uit de antwoorden op de enquête bleek tegen de verwachting in dat kandidaten positiever zijn over een SJT met een beoordelingsschaal dan een SJT met een meerkeuzeformat. Een verklaring hiervoor is dat een beoordelingsschaal het mogelijk maakt genuanceerdere antwoorden te geven – wat vermoedelijk beter aansluit bij de dilemma-achtige aard van de SJT-items. Verder beschouwen kandidaten een SJT met kennisgebaseerde instructies als moeilijker te faken dan een SJT met gedragsmatige instructies; dit is in overeenstemming met de empirische bevinding dat SJT's met kennisgebaseerde instructies minder vatbaar zijn voor faking.

De subgroepen op basis van geslacht en etnische en socio-economische achtergrond verschillen niet in hun kandidaatreacties op de SJT. Interactie-effecten tussen de demografische subgroep-variabelen en testkenmerken impliceerden echter wel dat subgroepen mogelijk verschillen in hun voorkeur voor bepaalde SJT-kenmerken. De conclusie uit deze studie is dat de reacties van subgroepen van kandidaten mogelijk positief beïnvloed kunnen worden door het wijzigen van de kenmerken van een SJT.

VOORSPELLEND WAARDE

Het ultieme doel van het uitbreiden van een selectieprocedure met niet-traditionele selectiemethoden, zoals de SJT, is het voorspellen van relevante toekomstige prestaties van kandidaten. De relevant geachte uitkomstmaat ten aanzien van de Integriteit SJT is de beoordeling van het professioneel gedrag van eerstejaars geneeskundestudenten. Om de voorspellende waarde van de SJT voor deze uitkomst te onderzoeken is gebruikgemaakt van *machine learning*-technieken.

Machine learning vindt langzamerhand ingang binnen personeelsselectie (Verhagen, 2020). Het is vermoedelijk slechts een kwestie van tijd voordat het eveneens zal worden toegepast voor de toelating tot het hoger onderwijs. In machine learning wordt een set gegevens aangeboden aan een computer, waaruit deze automatisch de onderliggende patronen 'leert'. Op basis van deze patronen, weergegeven door middel van algoritmen, worden vervolgens weer nieuwe gegevens geclassificeerd (Witten, Frank & Hall, 2011).

De voornaamste reden om in mijn promotieonderzoek gebruik te maken van machine learning is de lage prevalentie van (gesignaleerd) onprofessioneel gedrag onder geneeskundestudenten. Minder dan één op de tien studenten ontving een ontoereikende beoordeling van zijn/haar professioneel gedrag. Hoewel slechts een klein aantal studenten problemen met professionaliteit heeft, kan deze kleine groep wel grote moeilijkheden voor zichzelf, de opleiding of anderen veroorzaken. Daarom is het belangrijk deze groep vroeg te identificeren.

Een SJT maakt het mogelijk om niet-cognitieve eigenschappen op een gestandaardiseerde wijze te meten bij grote groepen kandidaten

Het probleem met het voorspellen van een zeldzame uitkomst is dat traditionele analysetechnieken (bijv. logistische regressie) alle studenten als 'professioneel' zullen classificeren, omdat de classificatie van de meeste studenten dan correct is (de meerderheid is immers professioneel). Deze hoge classificatienauwkeurigheid geldt in dat geval echter alleen voor de groep professionele studenten. Alle onprofessionele studenten worden daarentegen incorrect geclassificeerd als 'professioneel', terwijl voor deze groep een hoge classificatienauwkeurigheid juist wenselijk is. Deze methodologische moeilijkheid in het voorspellen van zeldzame uitkomsten heet het *base rate*-probleem (Mercaldo, Lau & Zhou, 2007). Machine learning wordt vaak toegepast wanneer sprake is van het *base rate*-probleem, zoals bij de classificatie van zeldzame ziektes of frauduleuze geldtransacties. Binnen het domein van machine learning bestaan verschillende methoden om het *base rate*-probleem aan te pakken.

Machine learning heb ik toegepast op de classificatie van studenten die ten aanzien van hun professioneel gedrag het oordeel 'verdient aandacht' hebben ontvangen (in tegenstelling tot het oordeel 'gaat (zeer) goed'). De classificatie is gebaseerd op verschillende cognitieve en niet-cognitieve variabelen, waaronder de Integriteit SJT. De classificatienauwkeurigheid, sensitiviteit (d.w.z. proportie correct geclassificeerde onprofessionele studenten) en specificiteit (d.w.z. proportie correct geclassificeerde professionele studenten) zijn vergeleken voor zes verschillende machine learning-algoritmen. Daarnaast zijn drie verschillende methoden toegepast om het *base rate*-probleem aan te pakken. Ten eerste is *undersampling* gebruikt, waarbij de veelvoorkomende klasse wordt verkleind door personen uit die klasse te verwijderen. Ten tweede is *oversampling* toegepast, waarbij de zeldzame klasse wordt vergroot door gegevens toe te voegen op basis van de kenmerken van bestaande personen uit die zeldzame klasse. Ten derde is gebruikgemaakt van *cost-sensitive learning*, waarbij de juiste classificatie van iemand uit de zeldzame klasse meer waarde krijgt toegekend dan de juiste classificatie van iemand uit de veelvoorkomende klasse.

Het *base rate*-probleem werd in de resultaten weerspiegeld door een lage classificatienauwkeurigheid voor de groep onprofessionele studenten en een hoge classificatienauwkeurigheid voor de groep professionele studenten. De meest effectieve methode om de classificatie van niet-professionele studenten te vergroten, en dus het *base rate*-probleem te verminderen, bleek de *oversampling*-techniek, vermoedelijk omdat met deze techniek geen waardevolle informatie wordt verwijderd. De algoritmen die resulteerden in de hoogste classificatienauwkeurigheid voor onprofessionele studenten waren de zogeheten *k-nearest neighbour* en de *neural network* algoritmen. Deze algoritmen laten een hoge mate van flexibiliteit toe, hetgeen erop duidt dat de onderliggende patronen in de onderzochte dataset niet gemakkelijk te vangen zijn met een simpel, lineair verband.

Om het aandeel van de Integriteit SJT in de classificatie van professioneel gedrag te onderzoeken, is de bijdrage van iedere variabele geëvalueerd. Hieruit bleek dat de Integriteit SJT weinig bijdraagt aan de classificatie van professioneel gedrag. Het ontbreken van een verband tussen de Integriteit SJT en de onderzochte uitkomstmaat is waarschijnlijk het gevolg van een combinatie van factoren, waaronder een niet-eenduidige conceptualisatie van wat onder een professionele arts/geneeskundestudent wordt verstaan en een onnauwkeurige meting van het laag prevalentie criterium onprofessioneel gedrag.

Ondanks de afwezigheid van een substantiële bijdrage van de Integriteit SJT geeft deze studie meer inzicht in hoe machine learning toegepast kan worden voor de selectie van geneeskundestudenten.

CONCLUSIE EN AANBEVELINGEN

Een SJT maakt het mogelijk om niet-cognitieve eigenschappen op gestandaardiseerde wijze te meten bij grote groepen kandidaten. Een SJT is een veelzijdig instrument dat kan voorkomen in verschillende vormen en maten en het wijzigen van de testkenmerken blijkt de testkwaliteit te kunnen beïnvloeden. De Integriteit SJT had geen voorspellende waarde voor de beoordeling van het professioneel gedrag van eerstejaars geneeskundestudenten; vandaar dat de SJT momenteel niet wordt ingezet voor de daadwerkelijke selectie van geneeskundestudenten.

Daarentegen wordt de toepassing van de Integriteit SJT voor andere gebruiksdoelen verkend, bijvoorbeeld als hulpmiddel voor scholieren om te beslissen of de geneeskundestudie bij hen past of als onderdeel van het onderwijs gericht op de professionele ontwikkeling van studenten. Verder zal toekomstig onderzoek moeten uitwijzen of de Integriteit SJT voorspellende waarde heeft voor uitkomsten op de langere termijn (bijv. prestaties tijdens de coschappen). Ondanks het ontbreken van de voorspellende waarde van de Integriteit SJT kan op basis van de resultaten van dit promotieonderzoek een aantal aanbevelingen worden gedaan voor gebruikers en ontwikkelaars van integriteit-gebaseerde SJT's.

De eerste aanbeveling is dat de toepassing van bestaande theoretische modellen kan helpen bij de ontwikkeling van een SJT en om de test meer te richten op een specifieke eigenschap. Daarnaast kan het zinvol zijn om onderscheid te maken tussen responsopties die beschrijven 'wat te doen' en responsopties die beschrijven 'wat niet te doen', omdat de testkwaliteit kan verschillen tussen deze twee typen responsopties. Het derde advies is om een pilotstudie uit te voeren om te kijken welke scoringsmethode het meest geschikt is, aangezien de manier waarop een SJT gescoord wordt van grote invloed kan zijn op de testkwaliteit. Verder wordt voor SJT's die gebruikmaken van beoordelingsschalen aanbevolen om rekening te houden met hoe individuen deze schalen gebruiken. De vierde aanbeveling is dat men met het wijzigen van de kenmerken van een SJT een positieve invloed kan uitoefenen op de kandidaatreacties van specifieke subgroepen.

Tot slot vormt dit onderzoek een aanzet voor de wijze waarop machine learning kan worden toegepast voor de

voorspelling van uitkomsten met een lage prevalentie voor de selectie van geneeskundestudenten. Zo kan machine learning helpen bij het lineair en/of non-lineair combineren van verschillende soorten informatie over de kandidaten en daarmee ondersteuning bieden bij de selectiebeslissing. Verder kan een methode zoals cost-sensitive learning bijdragen aan het expliciet maken van de doelen van selectie; welke beslissingsfout verdient bijvoorbeeld meer gewicht: het onterecht afwijzen van een geschikte kandidaat of het onterecht toelaten van een ongeschikte kandidaat?

De implementatie van een SJT in een high-stakes selectiesituatie vereist nauwkeurig onderzoek naar hoe SJT-kenmerken de kwaliteit van de test beïnvloeden. Bovengenoemde aanbevelingen kunnen onderzoekers en toelatingscommissies mogelijk helpen bij de ontwikkeling en het gebruik van SJT's en daarmee een potentiële bijdrage leveren aan het optimaliseren van de selectie van geneeskundestudenten. Toekomstig onderzoek moet eveneens uitwijzen of de resultaten ook van toepassing zijn op andere contexten dan de geneeskundeopleiding.

OVER DE AUTEUR

Wendy de Leng promoveerde in 2019 aan het Erasmus MC onder begeleiding van prof. dr. ir. Axel Themmen, prof. dr. Marise Born en dr. Karen Steger-Jager. Aan haar proefschrift is de Abbas dissertatieprijs 2019 toegekend. Sinds 2018 werkt De Leng als eindredacteur testangelegenheden bij het NIP. E-mail: wendy.deleng@psynip.nl.

Summary

JUDGING MEDICAL STUDENTS

W. DE LENG

This paper describes a PhD thesis examining the development and use of a Situational Judgement Test (SJT) for selecting medical students based on the noncognitive attribute of integrity. An SJT consists of scenarios that describe dilemma-like situations in a relevant context (e.g. medical school) followed by possible reactions to those situations. The task of the candidate is to judge the appropriateness of those response options. The five studies of the thesis investigate how several test characteristics influence a number of quality criteria of an SJT in measuring integrity among medical school applicants. The test characteristics under study are the development method, response format, scoring method, response appropriateness and response instructions. The quality criteria under examination are reliability, construct validity, criterion-related validity, subgroup differences, fakability and applicant perceptions. The study outcomes are described in this paper and may contribute to the development and use of integrity-based SJTs for medical school admissions.

Literatuur

- Barriga, A.Q. & Gibbs, J.C. (1996). Measuring cognitive distortion in antisocial youth: Development and preliminary validation of the "How I Think" questionnaire. *Aggressive Behavior*, 22, 333-343.
- Bergman, M.E., Drasgow, F., Donovan, M.A., Henning, J.B. & Juraska, S.E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14, 223-235.
- Catano, V.M., Brochu, A. & Lamerson, C.D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment*, 20, 333-346.
- Chan, D. & Schmitt, N. (2005). Situational Judgment Tests. In A. Evers, N. Anderson, & O. Voskuil (Eds.), *The Blackwell Handbook of Personnel Selection* (pp. 219-242). Malden, Verenigde Staten: Blackwell Publishing.
- Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S. & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, 82, 300-310.
- Christian, M.S., Edwards, B.D. & Bradley, J.C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83-117.
- Cook, M. (2016). *Personnel selection: Adding value through people - A changing picture*. Malden, Verenigde Staten: John Wiley & Sons Ltd.
- De Leng, W. (2019). *Measuring integrity for selection into medical school. Development of a Situational Judgement Test*. Academisch proefschrift, Erasmus Universiteit Rotterdam.
- De Visser, M. (2018). *Curriculum sample selection for entry to medical school* (Dissertatie). Geraadpleegd van: <http://hdl.handle.net/2066/195514>
- De Vries, R.E. & Born, M.P. (2013). De Vereenvoudigde HEXACO Persoonlijkheidsvragenlijst en een additioneel interstitieel Proactiviteitsfacet. *Gedrag & Organisatie*, 26, 223-245.
- Duffy, T.P. (2019). The Flexner report - 100 years later. *Yale Journal of Biology and Medicine*, 84, 269-276.
- Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin*, 51, 327-358.
- Frank, J.R., Snell, L. & Sherbino, J. (2015). *CanMEDS 2015 Physician competency framework*. Ottawa, Canada: The Royal College of Physicians and Surgeons of Canada.
- Lee, K., Ashton, M.C., Morrison, D.L., Cordery, J. & Dunlop, P.D. (2008). Predicting integrity with the HEXACO personality model: Use of self and observer reports. *Journal of Occupational and Organizational Psychology*, 81, 147-167.
- Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Medical Education*, 47, 182-189.
- Lievens, F., Patterson, F., Corstjens, J., Martin, S. & Nicholson, S. (2016). Widening access in selection using situational judgement tests: Evidence from the UKCAT. *Medical Education*, 50, 624-636.
- Lievens, F., Sackett, P.R. & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, 97, 1095-1101.
- Lucieer, S.M. (2016). *Selecting students for medical education: Exploring novel approaches* (Dissertatie). Geraadpleegd van https://repub.eur.nl/pub/79232/160108_Lucieer-Susanna-Magdalena.pdf
- Luschin-Ebengreuth, M., Dimai, H.P., Ithaler, D., Neges, H.M. & Reibnegger, G. (2015). Situational judgment test as an additional tool in a medical admission test: An observational investigation. *BMC Research Notes*, 8, 81.
- McDaniel, M.A., Psotka, J., Legree, P.J., Yost, A.P. & Weekley, J.A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, 96, 327-336.
- Mercaldo, N.D., Lau, K.F. & Zhou, X. H. (2007). Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in Medicine*, 26, 2170-2183.
- Nguyen, N.T., Biderman, M.D. & McDaniel, M.A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment*, 13, 250-260.
- Niessen, A.S.M. (2018). *New rules, new tools: Predicting academic achievement in college admissions* (Dissertatie). Geraadpleegd van <http://hdl.handle.net/11370/57c4d08f-7ca2-4348-8aa1-69f09a499005>
- Patterson, F., Ferguson, E. & Thomas, S. (2008). Using job analysis to identify core and specific competencies: Implications for selection and recruitment. *Medical Education*, 42, 1195-1204.
- Robins, R.W., Fraley, R.C., Roberts, B.W. & Trzesniewski, K.H. (2001). A longitudinal study of personality change in young adulthood. *Journal of Personality*, 69, 617-640.
- Schmidt, F.L., Oh, I. & Shaffer, J.A. (2016). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research findings*. Geraadpleegd van <http://home.ubalt.edu/tmitch/645/session%204/Schmidt%20%20Oh%20MKUP%20validity%20and%20util%20100%20yrs%20of%20research%20Wk%20PPR%202016.pdf>
- Schmitt, N. & Chan, D. (1999). The status of research on applicant reactions to selection tests and its implications for managers. *International Journal of Management Reviews*, 1, 45-62.
- Schreurs, S. (2020). *Selection for medical school: The quest for validity* (Dissertatie). Geraadpleegd van <https://cris.maastrichtuniversity.nl/en/publications/selection-for-medical-school-the-quest-for-validity>
- Schripsema, N.R. (2017). *Medical student selection: Effects of different admissions processes* (Dissertatie). Geraadpleegd van <http://hdl.handle.net/11370/d1c85157-2d48-4f54-ae46-6f598b73e4f2>
- Siu, E. & Reiter, H.I. (2009). Overview: What's worked and what hasn't as a guide towards predictive admissions tool development. *Advances in Health Sciences Education*, 14, 759-775.
- Stemler, S.E., Aggarwal, V. & Nithyanand, S. (2016). Knowing what NOT to do is a critical job skill: Evidence from 10 different scoring methods. *International Journal of Selection and Assessment*, 24, 229-245.
- Uggerslev, K.L., Fassina, N.E. & Kraichy, D. (2012). Recruiting through the stages: A meta-analytic test of predictors of applicant attraction at different stages of the recruiting process. *Personnel Psychology*, 65, 597-660.
- Urlings-Strop, L. (2018). *Developing a tool for selection for medical school: A search for academic and non-academic parameters to predict future medical school performance* (Dissertatie). Geraadpleegd van <https://repub.eur.nl/pub/109919/>
- Van Hooft, E.A.J. & Born, M.P. (2012). Intentional response distortion on personality tests: Using eye-tracking to understand response processes when faking. *Journal of Applied Psychology*, 97, 301-316.
- Verhagen, L. (2020). *De sollicitatie robot moet kandidaten eerlijker beoordelen dan mensen. Lukt dat al? De Volkskrant*. Geraadpleegd van <https://www.volkskrant.nl/wetenschap/de-sollicitatie-robot-moet-kandidaten-eerlijker-beoordelen-dan-mensen-lukt-dat-al-be737994/>
- Weekley, J.A. & Ployhart, R.E. (2006). *Situational judgment tests: Theory, measurement, and application*. Mahwah, Verenigde Staten: Lawrence Erlbaum Associates, Inc.
- Witten, I.H., Frank, E. & Hall, M.A. (2011). *Data mining: Practical machine learning tools and techniques*. San Francisco, Verenigde Staten: Morgan Kaufmann.
- Wouters, A. (2017). *Effects of medical school selection: On the motivation of the student population and applicant pool* (Dissertatie). Geraadpleegd van <https://research.vu.nl/en/publications/effects-of-medical-school-selection-on-the-motivation-of-the-stud>