Voorspellen van studiesucces in toelatingsprocedures

Niessen, Susan

Published in:
De Psycholoog

Publication date:
2019

VOORSPELLEN VAN STUDIESUCCES IN TOELATINGSPROCEDURES

SUSAN NIESSEN

Sinds kort worden studenten op basis van selectie- en matchings-procedures toegelaten tot het hoger onderwijs. Sociaal wenselijke antwoorden maken van scores op motivatie en persoonlijkheid echter slechte voorspellers van studiesucces. Proefstuderen zou de meest geschikte invulling van toelatingsprocedures zijn, concludeert Susan Niessen. *Proefstuderen had een hoge predictieve validiteit, werd positief gewaardeerd door kandidaten en liet weinig predictieve bias zien op basis van geslacht.*

SELECTEREN VOOR HET HOGER ONDERWIJS

VOORSPELLEN VAN STUDIESUCCES IN TOELATINGS-PROCEDURES

INTRODUCTIE

Tot voor kort waren bijna alle studieprogramma's in het hoger onderwijs toegankelijk voor iedereen die de juiste vooropleiding had voltooid en werd de toelating tot programma's met een *numerus fixus* bepaald door een gewogen loting, waarbij de cijfers uit het voortgezet onderwijs de kans op toelating bepaalden. Na enkele jaren van experimenteren worden alle kandidaten voor numerus fixus-studies nu geselecteerd via selectie, waarbij de onderwijsinstellingen de toelatingscriteria bepalen. De overige studieprogramma's organiseren een verplichte matchingsprocedure of 'studiekeuzecheck' die resulteert in een niet-bindend studiekeuzeadvies. Het doel van deze maatregelen is 'de juiste student op de juiste plaats' krijgen. Dat zou moeten leiden tot betere studiekeuze, minder uitval, snelle studievoortgang en betere studieprestaties. De vraag is welke instrumenten en procedures het beste ingezet kunnen worden om dit doel te bereiken.

In Europa is het meest voorkomende toelatingscriterium de prestatie in het voortgezet onderwijs, meestal in kaart gebracht via het gemiddelde middelbare schoolcijfer. Dat schoolcijfer wordt ook vaak genoemd als de beste voorspeller van prestaties in het hoger onderwijs (Robbins et al., 2004; Westrick et al., 2015). In Nederland mag dit wettelijk echter niet het enige toelatingscriterium zijn (Wet Kwaliteit in Verscheidenheid, 2013). Bovendien zijn middelbare schoolcijfers vaak lastig te vergelijken, omdat aspirant-studenten verschillende onderwijsachtergronden hebben en in toenemende mate uit verschillende landen komen. Veelvoorkomende onderdelen van selectie- en matchingsprocedures in Nederland zijn vaardigheden- en kennistests, motivatie- en persoonlijkheidsvragenlijsten, motivatiebrieven en interviews. Vaak worden deze instrumenten in verschillende combinaties gebruikt (Van den Broek et al., 2017; Warps et al., 2017). Voorbeelden zijn vragenlijsten over de Big Five persoonlijkheidskenmerken, studiegedrag, of de motivatie voor die studie, toetsen over aan de studie gerelateerde vaardigheden, zoals Engelstalige leesvaardigheid of kennis over biologie, en brieven of interviews waarin de kandidaat zijn of haar geschiktheid en motivatie toelicht. Een empirische onderbouwing van de validiteit van deze instrumenten ontbreekt echter in veel gevallen.

De centrale vraag in dit artikel is: Hoe kunnen we studenten het beste selecteren, rekening houdend met de context van het Nederlandse hoger onderwijs? In de meeste studies zijn gegevens gebruikt van (aspirant) studenten psychologie.

In Europa is het meest voorkomende toelatingscriterium de prestatie in het voortgezet onderwijs, meestal in kaart gebracht via het gemiddelde middelbare schoolcijfer

EFFECTIEVE TOELATINGSPROCEDURES Om deze onderzoeksvraag te beantwoorden, moeten we eerst definiëren wanneer er sprake is van een effectieve toelatingsprocedure. In dit onderzoek heb ik aangenomen dat het doel van selectie is om de beste studenten voor een studieprogramma te selecteren. De beste studenten zijn gedefinieerd als studenten die niet uitvallen, de meeste voortgang boeken en de hoogste cijfers behalen. Daarom is een goede predictieve validiteit voor toekomstige studieprestaties een eerste criterium voor een effectieve selectieprocedure. Ten tweede moeten selectieprocedures eerlijk zijn. Kandidaten moeten niet ten onrechte benadeeld worden op basis van bijvoorbeeld geslacht, etniciteit, of sociaaleconomische status. Ten derde moeten belanghebbenden zich eerlijk behandeld voelen en de procedures als wenselijk ervaren.

De belangrijkste belanghebbenden zijn de aspirant-studenten, aangezien een toelating of afwijzing tot een studie grote gevolgen kan hebben voor hun toekomstige loopbaan. Bovendien heeft selectie het oude systeem van gewogen loting vervangen. Daarom is een ander aandachtspunt of selectie de moeite waard is en inderdaad verbeterde studieresultaten oplevert.

STUDIESUCCES VOORSPELLEN

TRADITIONELE INSTRUMENTEN De beste voorspeller voor prestaties in het hoger onderwijs, prestaties uit het voortgezet onderwijs (Westrick et al., 2015), mag in Nederland niet als enig toelatingscriterium gebruikt worden. Andere instrumenten die goede voorspellingen opleveren (Sackett, et al., 2009), en die in veel landen zoals de Verenigde Staten vaak gebruikt worden, zijn tests voor cognitieve vaardigheden, zoals de SAT en ACT. Deze tests worden echter vooral gebruikt in landen waarin het voorgezet onderwijs in mindere mate gestratificeerd is dan in Nederland. Door de verregaande voorselectie op cognitieve vaardigheden die al in het voortgezet onderwijs plaatsvindt, zijn dit soort tests minder geschikt om onderscheid te maken tussen aspirant-studenten voor het Nederlandse hoger (wetenschappelijk) onderwijs (Resing & Drenth, 2007).

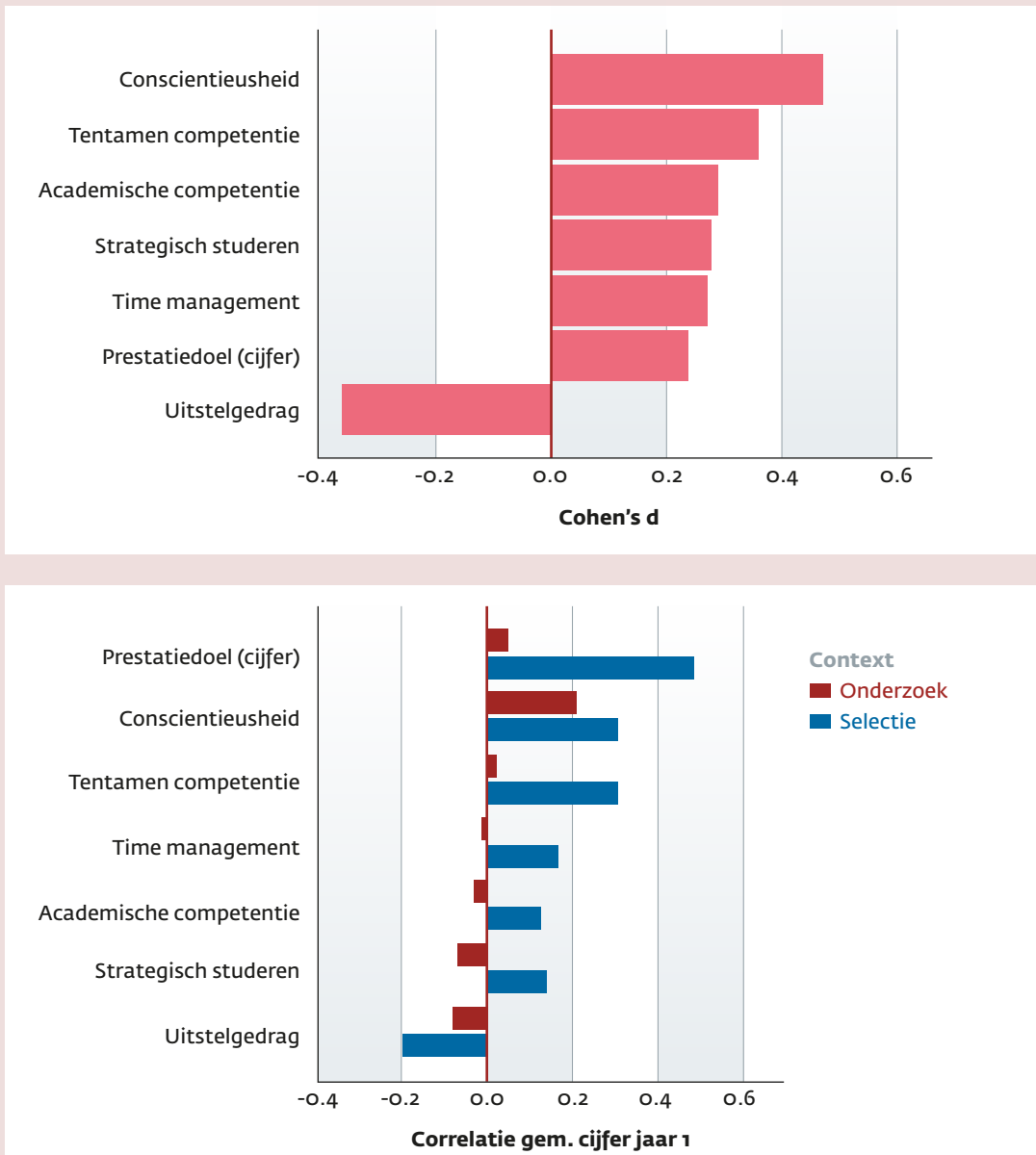
Andere veelgebruikte instrumenten zijn motivatiebrieven en interviews. De predictieve validiteit van deze instrumenten is echter klein (Goho & Blackman, 2006; Murphy et al., 2009; Patterson et al., 2016). Daaraan wordt hier verder geen aandacht besteed.

NONCOGNITIEVE EIGENSCHAPPEN Een alternatief dat recent in opkomst is geraakt, is het meten van 'noncognitieve' eigenschappen – een verzamelterm term die vaak gebruikt wordt om eigenschappen en vaardigheden zoals motivatie, persoonlijkheid, zelfvertrouwen, studievaardigheden en studiegerelateerde gedragstendenties te beschrijven. In een meta-analyse (Richardson et al., 2012) werden grote verbanden gevonden tussen studieprestaties in het hoger onderwijs en conscientieusheid, studiegerelateerd zelfvertrouwen, zelfregulatie, uitstelgedrag en strategisch studeergedrag. Daarbij komt dat dit soort eigenschappen vaak incrementele validiteit bovenop cognitieve vaardigheden laten zien (Credé & Kuncel, 2008; Richardson et al., 2012). Dit heeft ervoor gezorgd dat het opnemen van zulke 'noncognitieve' eigenschappen, na de ontwikkeling van gestandaardiseerd testen, omschreven wordt als de volgende grote stap in de ontwikkeling van selectie voor het hoger onderwijs (Hoover, 2013).

Een andere term voor deze eigenschappen is echter 'moeilijk te meten eigenschappen' (Kyllonen & Bertling, 2017). Deze eigenschappen worden meestal gemeten door middel van zelfrapportagevragenlijsten, wat sociaal wenselijk antwoorden mogelijk maakt (Birkeland et al., 2012). Sociaalwenselijk antwoorden kan bedoeld ('faken') of onbedoeld (bijvoorbeeld zelfdeceptie) zijn (Pauls & Crost, 2004). Of de mogelijkheid tot sociaal wenselijk antwoorden of 'faken' een bedreiging vormt voor de validiteit van dit soort instrumenten is een onderwerp van debat (Morgeson et al., 2007; Ones et al., 2007). Er is echter nog weinig onderzoek uitgevoerd dat een overtuigend antwoord oplevert. De meeste studies waarop de hierboven beschreven validiteitsbevindingen gebaseerd zijn, zijn niet uitgevoerd in de context van selectie of toelating.

PREDICTIEVE VALIDITEIT Om het effect van de mogelijkheid tot sociaal wenselijk antwoorden op de predictieve validiteit

FIGUUR 1. VERSCHILLEN IN SCORES EN IN PREDICTIEVE VALIDITEIT VAN SCORES, VERKREGEN IN EEN SELECTIE-CON TEXT EN EEN ONDERZOEKSCONTEXT.



te onderzoeken, hebben we 140 (aspirant-)studenten psychologie twee keer gevraagd om een vragenlijst in te vullen; één keer voor de toelating tot de studie (selectiecontext) en één keer na de start van de studie (onderzoekscontext). Beide keren hadden de scores geen consequenties. We hebben echter aangenomen dat aspirant-studenten die nog in de toelatingsfase zitten, toch sociaal wenselijk zouden antwoorden, zoals in eerder onderzoek is aangetoond (Griffin & Wilson, 2012).

De vragenlijst bestond uit verschillende 'noncognitieve' schalen die studiesucces in het hoger onderwijs zouden

moeten voorspellen, zoals conscientieusheid, doelstellingen en uitstelgedrag (Richardson et al., 2012). Voorbeelditems zijn 'Ik zie mezelf als iemand die geneigd is lui te zijn' (*Big Five Inventory*; John, et al., 1991) en 'Ik ben vaak bezig met dingen die ik al veel eerder had moeten doen' (*Lay's Procrastination Scale*; Lay, 1986). Uit de resultaten (zie figuur 1) bleek dat er kleine tot middelgrote verschillen waren tussen de scores verkregen in beide contexten. Zoals verwacht waren de scores uit de selectiecontext hoger voor positieve eigenschappen en lager voor negatieve eigenschappen.

Ook hebben we de relatie met het gemiddelde cijfer in het eerste jaar en studie-uitval vergeleken voor de scores uit beide contexten. Voor de scores uit de onderzoekcontext waren die verbanden zoals verwacht op basis van de literatuur (Richardson et al., 2012). Voor de scores verkregen uit de selectiecontext waren die verbanden echter voor de meeste schaalscores afwezig of aanzienlijk lager.

Deze resultaten laten zien dat aspirant-studenten, met opzet of onbewust, sociaal wenselijke antwoorden verstrekken op dit soort vragenlijsten en dat dit de predictieve validiteit van de scores op deze vragenlijsten negatief beïnvloedt of zelfs doet verdwijnen. Onderzoeksresultaten over zelfrapportagevragenlijsten gebaseerd op scores verkregen in 'low-stakes'-situaties kunnen dus niet zomaar gegeneraliseerd worden naar selectiesituaties. Op basis van deze resultaten concluderen we dat zelfrapportagevragenlijsten geen geschikte instrumenten zijn voor gebruik in selectieprocedures voor het hoger onderwijs.

PROEFSTUDEERTESTS

Een andere trend is het gebruik van instrumenten en procedures in de vorm van simulaties, ofwel de 'samples'-benadering. Psychologische tests zijn meestal ontworpen om 'signs', of theoretische constructen, zoals cognitieve capaciteiten of persoonlijkheid te meten. Simulaties of 'samples' zijn daar niet op gericht, maar proberen in plaats daarvan een zo representatief mogelijke 'steekproef' van het te voorspellen gedrag of de te voorspellen prestaties te leveren (Wernimont & Campbell, 1968). Deze aanpak wordt veel toegepast binnen de personeelsselectie in de vorm van work samples en assessment centers, die in die context een hoge predictieve validiteit opleveren (Schmidt & Hunter, 1998).

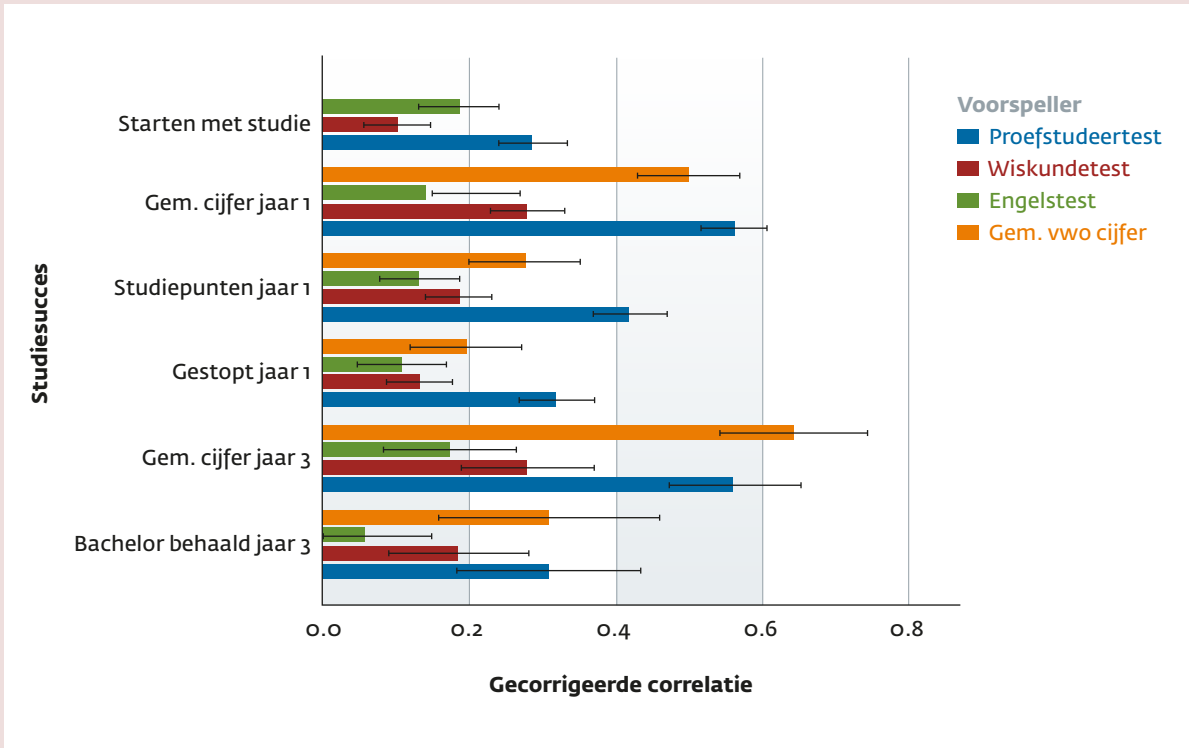
In het hoger onderwijs zijn proefstudeertests een voorbeeld van een 'samples'-aanpak van voorspellen. Proefstudeertests zijn tests die qua inhoud, vorm en voorbereiding zoveel mogelijk lijken op de inhoud van (het eerste jaar van) een studieprogramma en worden gebruikt in België, Oostenrijk, Finland en Nederland (De Visser et al., 2017; Lievens & Coetsier, 2002; Reibnegger et al., 2010; Visser et al., 2012; Vihavainen et al., 2013). Een verondersteld bijkomend voordeel van proefstuderen is dat het, naast een goede voorspellende waarde, de aspirant-student inzicht kan geven in de 'fit' met de studie; omdat de uit te voeren taken of tests erg lijken op wat er in het studieprogramma van studenten wordt gevraagd, biedt het de studenten een inkijkje in de studie en kunnen ze beslissen of het wel is wat ze ervan verwacht hadden en of het is wat ze willen. Er is echter nog weinig onderzoek uitgevoerd naar deze methode binnen de context van selectie in het hoger onderwijs.

PREDICTIEVE VALIDITEIT Om de predictieve validiteit van proefstuderen te onderzoeken, hebben we gegevens gebruikt van drie cohorten (aspirant) studenten psychologie die in de selectieprocedure een proefstudeertest hebben gemaakt en vervolgens aan de opleiding begonnen zijn (de gemiddelde leeftijd was M=20). De proefstudeertest hield in dat kandidaten thuis zelfstandig twee hoofdstukken uit het boek van het vak *Introductie in de Psychologie* moesten bestuderen en hier vervolgens een tentamen over maakten. Deze test is dus een kleine 'steekproef' van wat studenten in het eerste jaar van de studie moeten doen en kunnen.

Het verband tussen deze testscores en studieprestaties in onderzocht op basis van gegevens van 1804 studenten. Geen van de kandidaten die hebben deelgenomen aan de selectieprocedure is afgewezen voor de studie. Wel koos in ieder cohort ongeveer 20% van de kandidaten er zelf voor om toch niet aan die studie te beginnen. Deze kandidaten hadden bovendien lagere scores op de proefstudeertest, wat erop wijst dat de test wellicht inderdaad zelfselectie bevordert. Deze zelfselectie kan echter ook zorgen voor restrictie in range, wat tot onderschattingen van de predictieve validiteit van de test kan leiden. Daarom hebben we een correctie voor indirecte range-restrictie (Hunter et al., 2006) toegepast op de validiteitscoëfficiënten van de verschillende selectieonderdelen. Ter vergelijking hebben we ook gekeken naar de predictieve validiteit van het gemiddelde vwo-cijfer en van scores op een wiskundetest en een test voor Engelse leesvaardigheid, twee relevante vaardigheden voor de studie psychologie.

De resultaten in figuur 2 laten zien dat de proefstudeertest een goede voorspeller was voor studiesucces in het eerste jaar en na drie jaar. De proefstudeertest was bovendien een betere voorspeller dan de scores op tests voor wiskunde en Engels en ongeveer een even goede voorspeller als het gemid-

FIGUUR 2. PREDICTIEVE VALIDITEIT VAN VERSCHILLENDE VOORSPELLERS VOOR STUDIESUCCES, MET 95% BETROUWBAARHEIDSINTERVALLEN.

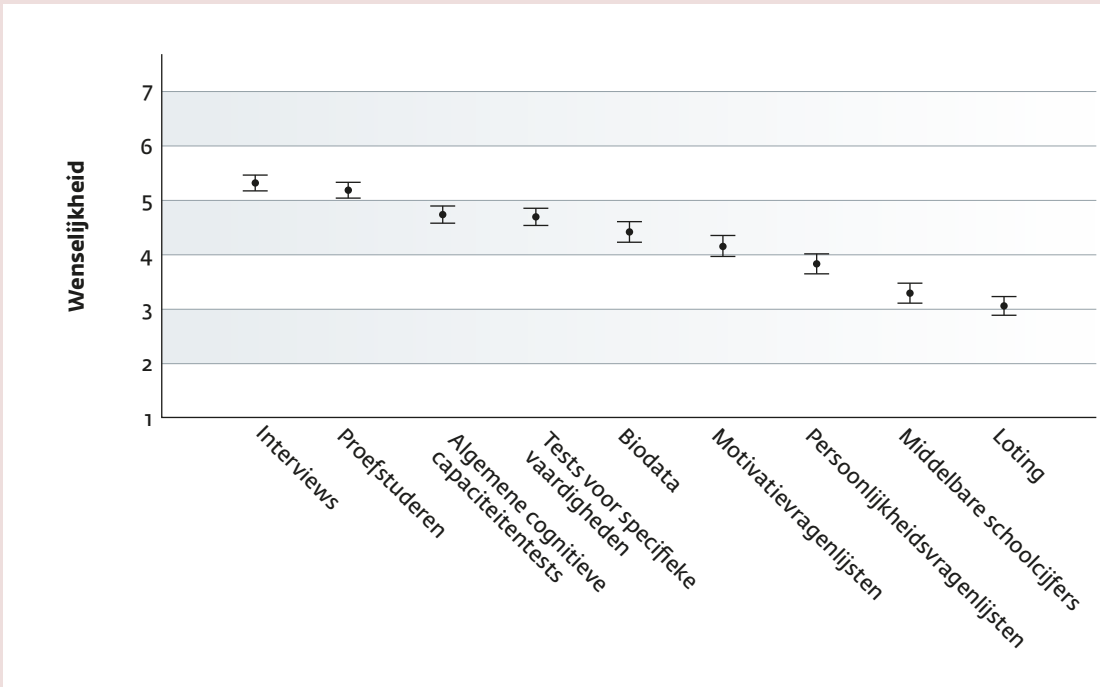


delde vwo-cijfer. Als we bedenken dat het gemiddelde vwo-cijfer een samenvatting is van gegevens die over een periode van drie jaar verzameld zijn, is dit een indrukwekkend resultaat voor een enkele testscore. De scores op de proefstudeertest hadden bovendien incrementele validiteit bovenop het gemiddelde vwo-cijfer ($\Delta R^2 = .12$ na correctie voor range-restrictie) en kan dus ook aanvullend gebruikt worden.

BIAS Een veelbesproken probleem bij toelatingstests is *bias* op basis van geslacht. Op traditionele gestandaardiseerde toelatingstests, zoals de SAT en de ACT, scoren vrouwelijke kandidaten lager dan mannelijke kandidaten. We spreken in deze context echter pas van bias als er sprake is van differentiële predictie, ofwel een systematisch verschil in studieprestaties voor mannen en vrouwen die dezelfde testscores behalen (Cleary, 1968). Bij deze tests is echter ook sprake van onderpredictie van vrouwelijke kandidaten; zij behalen systematisch een hoger studiesucces dan voorspeld op basis van de testscores, wat leidt tot onjuiste en oneerlijke selectiebeslissingen. Een veelgenoemde verklaring hiervoor is *selection system bias*: het ontbreken van andere relevante voorspellers die ook gerelateerd zijn aan geslacht. Uit verschillende studies is bijvoorbeeld gebleken dat vrouwelijke onderpredictie afneemt wanneer scores op conscientieusheid of zelfdiscipline aan het voorspellingsmodel worden toegevoegd (Keiser et al., 2016; Mattern et al., 2017). Dit zijn echter nu juist eigenschapen die moeilijk te meten zijn in selectiesituaties.

Wij veronderstelden dat proefstudeertests, gezien de sterke representativiteit voor de te voorspellen prestaties, indirect zowel relevante cognitieve- en noncognitieve eigenschappen meten (Callinan & Robertson, 2000). Dit zou, naast een hoge predictieve validiteit, ook kunnen resulteren in weinig of geen differentiële predictie op basis van geslacht. We hebben deze hypothese onderzocht op basis van Bayesiaanse gemodereerde multipele regressieanalyses. Op basis van deze analyses kan de evidentie voor de nulhypothese (geen differentiële predictie) gekwantificeerd worden en kunnen 95% credible intervals rondom de effectgroottes

FIGUUR 3. PERCEPTIES VAN ASPIRANT-STUDENTEN OVER DE WENSELIJKHEID VAN VERSCHILLENDE SELECTIEMETHODEN.



berekend worden, die meest waarschijnlijke waarden voor de effectgroottes weergeven gegeven de data (zie bijvoorbeeld Kruschke et al., 2012; Kruschke & Liddell, 2018).

In twee van de drie onderzochte cohorten was er evidentie voor het ontbreken van differentiële predictie van het gemiddelde cijfer in het eerste jaar op basis van de scores op de proefstudeertest, en in één cohort was er evidentie voor onderpredictie van vrouwelijke kandidaten, maar waren de effecten slechts klein. Voor het voorspellen van het gemiddelde cijfer na het derde jaar vonden we ook evidentie voor het ontbreken van differentiële predictie op basis van de scores op de proefstudeertest.

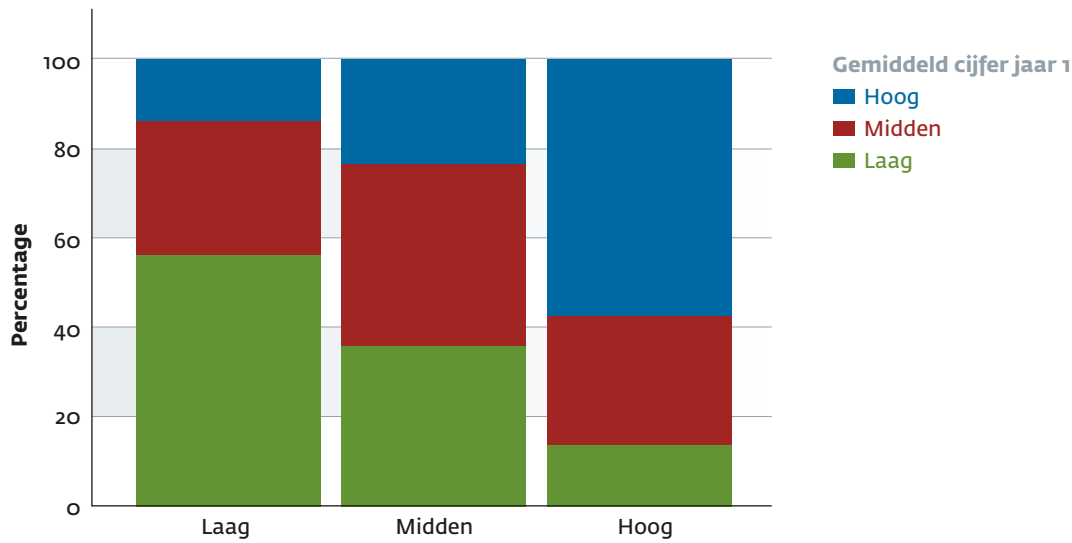
Het inzetten van representatieve proefstudeertests lijkt dus niet of nauwelijks tot een bias op basis van geslacht te leiden.

PERCEPTIES VAN ASPIRANT-STUDENTEN

We hebben ook de belangrijkste belanghebbenden, aspirant-studenten, naar hun mening gevraagd over verschillende methoden die veel worden gebruikt in selectie voor het hoger onderwijs. We hebben daarvoor een vragenlijst gebruikt gebaseerd op die van Steiner en Gilliland (1996), ontwikkeld om percepties van sollicitanten in de context van personeelselectie te meten aan de hand van *organizational justice theory*. Voorbeeldvragen zijn: 'Hoe zou je de effectiviteit van [methode] om gekwalificeerde personen voor de studie psychologie te identificeren beoordelen?' en 'Als je zou worden afgewezen op basis van een [methode], wat zou je dan van de eerlijkheid van deze procedure vinden?'

De vragenlijst is ingevuld door 220 aspirant-studenten psychologie. De aspirant-studenten vonden interviews en proefstuderen de meeste wenselijke selectiemethoden en vonden het gebruik van middelbare schoolcijfers en loting het minst wenselijk (zie figuur 3). Vooral de lage wenselijkheid van het gebruik van middelbare schoolcijfers is opvallend, gezien het vele gebruik in selectieprocedures en de hoge predictieve validiteit. De respondenten vonden dat het gebruik van middelbare schoolcijfers weinig mogelijkheden bood om de eigen capaciteiten te laten zien of om onderscheid te maken tussen studenten onderling en

FIGUUR 4. STUDENTEN UIT COHORT 2013, OPGEDEELD IN DRIE GELIJKE GROEPEN OP BASIS VAN SCORES OP DE PROEFSTUDEERTEST EN OP BASIS VAN HET GEMIDDELDE CIJFER IN HET EERSTE JAAR.



vonden de relevantie voor studiesucces in de psychologieopleiding niet sterk.

Wat eveneens opvalt, is de relatief lage wenselijkheid van motivatie- en persoonlijkheidsvragenlijsten. Respondenten gaven bovendien aan dat deze vragenlijsten gemakkelijk te 'faken' zijn, wat negatief samenhing met de beoordeelde wenselijkheid van deze methoden.

Verder bleek dat de wenselijkheid van methoden vooral samenhing met percepties van de predictieve validiteit, de relatie met de studie, het onderscheidend vermogen, de kans om capaciteiten te tonen, de indruk dat de methode wetenschappelijke onderbouwd is en de indruk dat de methode veel gebruikt wordt. Er waren slechts kleine of geen verbanden tussen de wenselijkheid van de methoden en percepties over het recht om de methode te gebruiken, onpersoonlijkheid, de schending van privacy en de invloed van voorbereiding op testprestaties.

In de context van personeelsselectie is bovendien eerder gevonden dat percepties van kandidaten samenhangen met gedrag, zoals de kans om een aanbod voor een baan af te wijzen (Hausknecht et al., 2004). Daarom verwachtten we dat de percepties van aspirant-studenten over de gebruikte selectiemethoden (proefstuderen en tests voor specifieke vaardigheden) zou samenhangen met de beslissingen om wel of niet aan de studie te beginnen. De relaties tussen percepties over de gehanteerde methoden en de beslissing om aan de studie te beginnen waren echter zeer klein en niet statistisch significant.

DE TOEGEVOEGDE WAARDE VAN DECENTRALE SELECTIE

Er zijn verschillende instrumenten en methoden waarvan de scores een verband met studiesucces laten zien. De vraag is of die verbanden sterk genoeg zijn om aanzienlijk betere beslissingen te maken en tot een verhoging van het studiesucces te leiden. Sommigen beargumenteren dat, gezien de verre van perfecte predictieve validiteit van de instrumenten, selectie hooguit tot een kleine toename van het studiesucces kan leiden. In de praktijk gebruiken we selectie-instrumenten om aspirant-studenten te rangordenen en op basis daarvan

Zelfrapportagevragenlijsten zijn geen geschikte instrumenten voor gebruik in selectieprocedures voor het hoger onderwijs

te besluiten wie wel en niet wordt aangenomen. Critici van selectie beargumenteren dat deze rangordening echter niet zó valide is dat we mogen concluderen dat, bijvoorbeeld, de kandidaat met rangnummer 30 een grotere kans heeft op succes dan de kandidaat met rangnummer 31, zelfs niet als we erg betrouwbare en valide instrumenten gebruiken (Van der Maas en Visser, 2017). Het klopt dat we niet zo precies kunnen meten of voorspellen dat we een dergelijk onderscheid kunnen maken. Dat betekent echter niet dat we *helemaal* geen onderscheid kunnen maken tussen aspirant-studenten onderling.

De praktische betekenis van de correlatiecoëfficiënten, die meestal worden gebruikt om uit te drukken in hoeverre een onderscheid mogelijk is, zijn echter vaak lastig in te schatten. Daarom staat ter illustratie van de mate waarin we onderscheid kunnen maken op basis van een testscore met een ongecorrigeerde predictieve validiteit van r=.49 een voorbeeld in figuur 4. In deze figuur zijn de studenten uit één cohort opgedeeld in drie gelijke groepen op basis van de prestatie op de proefstudeertest en op basis van het gemiddelde cijfer in het eerste jaar. Te zien is dat, hoewel verre van perfect, op basis van deze testscore wel onderscheid te maken is; 56% van de studenten die laag scoorden op de proefstudeertest behoorden ook tot de groep met de laagste cijfers, terwijl maar 14% bij de groep met de hoogste cijfers behoorde, en 58% van de studenten die hoog scoorden op de proefstudeertest hoorde ook bij de groep met de hoogste cijfers, terwijl maar 14% bij de groep met de laagste cijfers hoorde.

Een dergelijk voorbeeld beantwoordt echter nog niet de vraag of decentrale selectie, bijvoorbeeld op basis van een proefstudeertest, leidt tot hoger studiesucces. Om die vraag te kunnen beantwoorden is informatie over de predictieve validiteit niet voldoende, maar moeten we ook rekening houden met de base rate (het percentage kandidaten dat geschikt is) en de selectieratio (het percentage kandidaten dat wordt aangenomen). Daarvoor kunnen we verschillende utiliteitsmodellen gebruiken, zoals die van Taylor en Russell (1939), gebruikt om de toename in het percentage succesvolle kandidaten onder de geselecteerden te schatten, of die van Naylor en Shine (1965), waarmee de gemiddelde toename van de criteriumprestatie geschat kan worden.

Om deze modellen te gebruiken hebben we eerst schattingen nodig van de base rate en de selectieratio. Deze gegevens zullen per opleiding en uitkomstmaat verschillen. We gebruiken in onderstaand voorbeeld (tabel 1) studie-uitval en het gemiddelde cijfers als uitkomstmaten. De base rates hebben we geschat op basis van informatie van studenten psychologie. Aangezien er zelfselectie heeft plaatsgevonden, zijn dit echter waarschijnlijk overschattingen. We gebruiken daarnaast twee hypothetische selectieratio's: een 'milde' selectieratio van .80 en een 'strenge' selectieratio van .30. Daarnaast gaan we ervan uit dat we een test gebruiken met een hoge predictieve validiteit voor het gemiddelde cijfer (r=.50) en een middelgrote predictieve validiteit voor studieuitval (r=.30). De resultaten staan in tabel 1, waar duidelijk te zien is dat het studiesucces maar beperkt toeneemt als de selectieratio hoog is. Het studiesuc-

TABEL 1. VOORBEELD VAN HET RESULTAAT VAN SELECTIE BIJ VERSCHILLENDE SELECTIERATIO'S

CRITERIUM	BASE RATE	RESULTAAT	
		SELECTIERATIO	
		.80	.30
Studieuitval[a]	.80	.85	.92
Gemiddeld cijfer jaar 1[b]	6.6	6.8	7.4

[a]Taylor-Russell model aangepast voor dichotome uitkomstmaten (Abrahams et al., 1971) [b]Naylor/Shine model.

ces neemt pas behoorlijk toe als er veel aspirant-studenten worden afgewezen.

Uit een inventarisatie van de selectieratio's bij alle studieprogramma's die voor het studiejaar 2014-2015 een numerus fixus hadden (Dienst Uitvoering Onderwijs, 2014) bleek echter dat de meeste programma's uiteindelijk helemaal geen kandidaten afwezen, omdat het aantal gegadigden uiteindelijk niet groter was dan het aantal beschikbare plaatsen. Slechts een klein aantal studies, zoals (dier)geneeskunde en tandheelkunde (wo) en fysiotherapie en mondzorgkunde (hbo), hadden een selectieratio kleiner dan .80.

Bij de meeste studieprogramma's zal het studiesucces dus waarschijnlijk slechts in beperkte mate toenemen door de invoering van selectie, zelfs als er selectie-instrumenten met een behoorlijke predictieve validiteit worden ingezet. Een kanttekening bij deze conclusie is dat we in deze modellen aannemen dat de base rate constant blijft, ongeacht welke selectieprocedure er wordt gehanteerd. Het is goed denkbaar dat dat niet het geval is en dat verschillende selectieprocedures (bijvoorbeeld loting of proefstuderen) verschillende kandidaten aantrekken.

DISCUSSIE

De centrale vraag in dit onderzoeksproject was: Hoe kunnen we studenten het beste selecteren, rekening houdend met de context van het Nederlandse hoger onderwijs? Effectieve toelatingsprocedures waren daarbij gedefinieerd als procedures die (1) een goede predictieve validiteit voor studiesucces hebben, (2) kandidaten niet ten onrechte benadelen op basis van geslacht, etniciteit, of sociaaleconomische status en (3) als wenselijk en rechtvaardig ervaren worden door aspirant-studenten.

Hoewel al vaak is aangetoond dat scores op schalen voor noncognitieve eigenschappen, zoals persoonlijkheidskenmerken, gedragstendenties en motivatie, samenhangen met studiesucces (Credé & Kuncel, 2008; Richardson et al., 2012), bleek uit onze resultaten dat scores op dit soort schalen verzameld met zelfrapportagevragenlijsten geen goede voorspellende waarde hebben wanneer ze in een selectiesituatie worden afgenomen. Ook vinden aspirant-studenten deze vragenlijsten geen geschikte selectie-instrumenten. Om noncognitieve eigenschappen in selectiesituaties te kunnen meten, zullen dus andere instrumenten ontwikkeld en onderzocht moeten worden. Voorbeelden van zulke initiatieven zijn *situational judgment*-tests of *forced choice*-vragenlijsten, waarmee tot nu toe wisselend succes is behaald wat betreft de mate waarin de instrumenten bestand zijn tegen sociaal wenselijk antwoorden (Christiansen et al., 2005; Lievens, 2013).

Proefstudeertests hadden wel een hoge predictieve validiteit voor studiesucces wanneer ze gebruikt werden in een selectiesituatie. Deze tests lieten geen, of slechts in kleine mate, differentiële predictie op basis van geslacht zien en werden door aspirant-studenten als een wenselijke selectiemethode gezien. Proefstuderen lijkt dus de meest geschikte methode voor gebruik in selectieprocedures in het Nederlandse hoger onderwijs, in plaats van of naast het gebruik van middelbare schoolcijfers. Ook voor matching-procedures of studiekeuzechecks lijkt deze methode geschikt, gezien de aanwijzingen voor zelfselectie (lager scorende kandidaten kozen er vaker voor om toch niet aan de studie te beginnen). Bovendien past deze methode, door de inhoudelijke overeenkomst met het studieprogramma waarvoor wordt geselecteerd of gematcht, beter bij het doel van matching en selectie in Nederland dan het gebruik van algemene voorspellers zoals persoonlijkheidseigenschappen, cognitieve capaciteiten of middelbare schoolcijfers. Met de doelstelling: 'De juiste student op de juiste plaats' selecteren of matchen we immers vooral voor een *onderwijsprogramma*, niet zozeer voor een *onderwijsniveau*. Het is bijvoorbeeld waarschijnlijk dat een kandidaat die laag scoort op conscientieusheid of een laag gemiddeld vwo-cijfer heeft voor verschillende onderwijsprogramma's zou worden afgewezen of een negatief matchingsadvies zou krijgen. Dat is niet de doelstelling van deze toelatingsprocedures.

De conclusie is dus dat, als er wordt geselecteerd, proefstuderen de beste methode lijkt te zijn. Gezien de hoge selectieratio's van de meeste studieprogramma's en de daarmee gepaard gaande geringe toename van studiesucces als gevolg van selectie, is het echter de vraag of decentrale selectie überhaupt de moeite waard is. Het antwoord op die vraag is deels subjectief; over de vraag of, bijvoorbeeld, een afname van vijf procent in studie-uitval de moeite waard is, zijn de meningen waarschijnlijk verdeeld. Daarbij moet ook rekening worden gehouden met de inspanningen die selectieprocedures vragen van onderwijsinstellingen en kandidaten; loting is bijvoorbeeld veel eenvoudiger te organiseren dan proefstuderen.

Hoe dan ook zullen er bij studieprogramma's die meer aanmeldingen dan plaatsen hebben kandidaten moeten worden afgewezen. Ondanks de vaak geringe toename in studiesucces is proefstuderen naar onze mening toch de meest geschikte manier om toelatingsbeslissingen te maken.

Met de doelstelling 'de juiste student op de juiste plaats' selecteren we vooral voor een onderwijsprogramma, niet zozeer voor een onderwijsniveau

Studenten geven hier, zeker in vergelijking met loting, de voorkeur aan en het biedt een mogelijkheid tot zelfselectie en inzicht in de 'fit' met de studie. Dit zou er in de praktijk voor kunnen zorgen dat studie-uitval en -switch afneemt, wat positief zou zijn voor onderwijsinstellingen én studenten.

De effecten van verschillende selectie- en matchingsprocedures op het maken van een juiste studiekeuze moeten in de toekomst echter verder onderzocht worden. Tevens merken we op dat het ontwikkelen van proefstudeertests voor meer praktijk- of beroepsgerichte onderwijsprogramma's complexer en arbeidsintensiever zal zijn dan het geval was voor een theoretische studie zoals psychologie, waarbij de meeste cursussen bestaan uit het volgen van colleges, het bestuderen van boeken en artikelen en het maken van een tentamen. Ook is differentiële predictie van proefstudeertests alleen onderzocht op basis van geslacht. Dergelijk onderzoek zou ook uitgevoerd moeten worden op basis van etniciteit en sociaaleconomische status.

Concluderend heeft dit onderzoek laten zien dat een 'samples'-benadering van voorspellen met succes kan worden toegepast in de context van het hoger onderwijs. Toekomstig onderzoek zal meer inzicht moeten geven in waarom deze aanpak goede resultaten oplevert en welke kenmerken van proefstudeertests bijdragen aan de predictieve validiteit, positieve percepties van stakeholders en het verminderen van bias.

OVER DE AUTEUR

Susan Niessen is werkzaam als universitair docent aan de afdeling Psychometrie en Statistiek (Psychologie) van de faculteit Gedrags- en Maatschappijwetenschappen, Rijksuniversiteit Groningen. Dit artikel is geschreven naar aanleiding van haar proefschrift New Rules, New Tools: Predicting academic achievement in college admissions, onder supervisie van prof. dr. Rob R. Meijer, dr. Jorge N. Tendeiro en mr. dr. Jaap J. Dijkstra. Correspondentie over dit artikel kan via Susan Niessen, e-mail: a.s.m.niessen@rug.nl.

Literatuur

Birkeland, S.A., Manson, T.M., Kisamore, J.L., Brannick, M.T. & Smith, M.A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317–335. doi:10.1111/j.1468–2389.2006.00354.x.

Callinan, M. & Robertson, I.T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8*, 248-260. doi:10.1111/1468-2389.00154

Cleary, T.A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. doi:10.1111/j.1745-3984.1968.tb00613.x

Credé, M. & Kuncel, N.R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate performance. *Perspectives on Psychological Science, 3*, 425-453. doi:10.1111/j.1745-6924.2008.00089.x

Cremonini, L., Leisyte, L., Weyer, E. & Vossensteyn, J. J. (2011). *Selection and matching in higher education: An international comparative study.* Enschede, the Netherlands: Center for Higher Education Policy Studies (CHEPS).

Christiansen, N.D., Burns, G.N. & Montgomery, G.E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267-307. doi:10.1207/s15327043hup1803_4

de Visser, M., Fluit, C., Fransen, J., Latijnhouwers, M., Cohen-Schotanus, J. & Laan, R. (2017). The effect of curriculum sample selection for medical school. *Advances in Health Science Education, 22*, 43–56. doi:10.1007/s10459-016-9681-x

Dienst Uitvoering Onderwijs (2014). *Jaarverslag numerus fixus-opleidingen.* Opgevraagd via https://tinyurl.com/y94c8ef3

Goho, J. & Blackman, A. (2006). The effectiveness of academic admission interviews: An exploratory meta-analysis. *Medical Teacher, 28*, 335-340. doi:10.1080/01421590600603418

Griffin, B. & Wilson, I.G. (2012). Faking good: Self-enhancement in medical school applicants. *Medical Education, 46*, 485–490. doi:10.1111/j.1365-2923.2011.04208.x.

Hausknecht, J.P., Day, D.V. & Thomas, S.C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639-683. doi:10.1111/j.1744-6570.2004.00003.x

Hunter, J.E., Schmidt, F.L. & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594-612. doi:10.1037/0021-9010.91.3.594

Hoover, E. (2013). Noncognitive measures: The next frontier in college admissions. *Chronicle of Higher Education*. Opgevraagd via: https://www.collegesuccessfoundation.org/document.doc?id=851

Keiser, H.N., Sackett, P.R., Kuncel, N.R. & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology, 101*, 569-581. doi:10.1037/apl0000069

Summary

NEW RULES, NEW TOOLS: PREDICTING ACADEMIC ACHIEVEMENT IN COLLEGE ADMISSIONS

A. S. M. NIESSEN

Since 2017, admission to higher education programs in the Netherlands happens through selective admission for programs with a fixed number of places, and through non-binding 'matching' procedures for open admission programs. This research project aimed to answer the question on what methods to use to predict academic performance within such admission procedures. Scales for noncognitive characteristics such as motivation and personality traits yielded moderate predictive validity when administered in low-stakes conditions, but predictive validity was attenuated when they were administered in a high-stakes context, and were not perceived favourably by applicants. Scores on a curriculum-sampling test, a test that mimics the study program in preparation and content, analogous to work samples in personnel selection, showed high predictive validity, favourable applicant perceptions, and little or no predictive bias based on gender. Hence, curriculum-sampling tests seem to be the best fitting method to use in admission procedures to higher education in the Netherlands.

Kruschke, J.K., Aguinis, H. & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*, 722-752. doi:10.1177/1094428112457829.

Kruschke, J.K. & Liddell, T.M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*, 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kyllonen, P. & Bertling, J. (2017, April). *Interpersonal and intrapersonal skills assessment: Design, development, scoring, and reporting*. Workshop provided at the Annual Meeting of the National Council on Measurement in Education, San Antonio, Texas.

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Medical Education, 47*, 182-189. doi:10.1111/medu.12089

Lievens, F. & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*,10, 245–257. doi:10.1111/1468-2389.00215

Mattern, K. Sanchez, E. & Ndum, E. (2017). Why do achievement measures underpredict female academic performance? *Educational Measurement: Issues and Practice, 36*, 47-57. doi:10.1111/emip.12138

Morgeson, F.P., Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K. & Schmitt, N. (2007b). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729. doi:10.1111/j.1744-6570.2007.00089.x.

Murphy, S.C., Klieger, D.M., Borneman, M.J. & Kuncel, N.R. (2009). The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. *College and University, 84*, 83-86.

Naylor, J.C. & Shine, L.C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology, 3*, 33-42.

Ones, D.S., Dilchert, S., Viswesvaran, C. & Judge, T.A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027. doi:10.1111/j.1744-6570.2007.00099.x

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F. & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*, 36–60. doi:10.1111/medu.12817

Pauls, C.A. & Crost, N.W. (2004). Effects of faking on self-deception and impressionmanagement scales. *Personality and Individual Differences, 37*, 1137–1151. doi:10.1016/j.paid.2003.11.018.

Reibnegger, G., Caluba, H.C., Ithaler, D., Manhal, S., Neges, H.M. & Smolle, J. (2010). Progress of medical students after open admission or admission based on knowledge tests. *Medical Education, 44*, 205–214. doi:10.1111/j.1365-2923.2009.03576.x

Resing, W.C.M. & Drenth, P.J.D. (2007). *Intelligentie: Weten en meten*. Amsterdam: Uitgeverij Nieuwezijds.

Richardson, M., Abrahams, C. & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353–387. doi:10.1037/a0026838.

Robbins, S.B., Lauver, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288. doi:10.1037/0033-2909.130.2.261

Sackett, P.R., Kuncel, N.R., Arneson, J.J., Cooper, S.R. & Waters, S.D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin, 135*, 1–22. doi:10.1037/a0013978

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 research findings. *Psychological Bulletin, 124*, 262–274. doi:10.1037/0033-2909.124.2.262

Steiner, D.D. & Gilliland, S.W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology, 81*, 134-141. doi:10.1037/0021-9010.81.2.134

Taylor, H.C. & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23*, 565-578. doi:10.1037/h0057079

van den Broek, A., Nooij, J., van Essen, M. & Duysak, S. (2017). *Selectie & plaatsing bij numerusfixusopleidingen*. Nijmegen, the Netherlands: ResearchNed. Opgevraagd via: https://tinyurl.com/y9449rhr

van der Maas, H. & Visser, K. (2017, June 8). Wet selectie studenten is niet uitvoerbaar. *De Volkskrant*. Opgevraagd via: https://tinyurl.com/yak8fxft

Vihavainen, A., Luukkainen, M., & Kurhila, J. (2013, October). *MOOC as semester-long entrance exam*. Paper presented at the 14th Annual ACM SIGITE Conference on Information Technology Education, Orlando, Florida, United States.

Visser, K., van der Maas, H., Engels-Freeke, M. & Vorst, H. (2012). Het effect op studiesucces van decentrale selectie middels proefstuderen aan de poort. *Tijdschrift voor Hoger Onderwijs, 30*, 161–173.

Warps, J., Nooij, N., Muskens, M., Kurver, B. & van den Broek, A. (2017). *De studiekeuzecheck*. Nijmegen, the Netherlands: ResearchNed. Opgevraagd via: https://tinyurl.com/y8dnvn3s

Wernimont, P.F. & Campbell, J.P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372-376. doi:10.1037/h0026244

Westrick, P.A., Le, H., Robbins, S.B., Radunzel, J.R. & Schmidt, F.L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT scores, high school grades, and SES. *Educational Assessment, 20*, 23-45. doi:10.1080/10627197.2015.997614

Wet kwaliteit in verscheidenheid. *Memorie van Toelichting*. Opgevraagd via: https://tinyurl.com/y7ekn87z