

Maatwerk bij persoonlijkheidsmeting in de 21^e eeuw

verslag van de zevende netwerkbijeenkomst van het Nederlands-Vlaamse Netwerk voor Wervings- en Selectieonderzoek

dr. René Butter
Erasmus Universiteit Rotterdam

Inleiding

Op 26 oktober 2012 vond de zevende netwerkbijeenkomst van het Nederlands-Vlaamse Netwerk voor Wervings- en Selectieonderzoek plaats aan de Vrije Universiteit Amsterdam. Dit netwerk heeft tot doel om de onderzoekskwaliteit te versterken, de kloof tussen onderzoek en de praktijk te dichten en de samenwerking tussen onderzoekers te stimuleren. Bij deze bijeenkomst waren 25 deelnemers aanwezig. Het ging om wetenschappers van verschillende universiteiten, vertegenwoordigers uit het werkveld en personen die deze twee rollen combineren. Aansluitend vond het congres Assessment en Co-Creation in the Cloud plaats dat werd georganiseerd door de Nederlandse Stichting voor Psychotechniek.

De gastspreker op de netwerkbijeenkomst van dit jaar was dr. Don Kluemper, assistent professor aan het Management Department van de Northern Illinois University in DeKalb.

De bijeenkomst werd ingeleid door prof dr. Rob Meier, hoogleraar Psychometrische en Statistische Technieken aan de Rijksuniversiteit Groningen. Samen met prof dr. Marise Born, hoogleraar Personeelspsychologie aan de Erasmus Universiteit Rotterdam, is hij oprichter en coördinator van het netwerk.

Inhoudelijke deel

Don Kluemper hield een presentatie over zelfbeoordeling en beoordeling door derden van persoonlijkheid bij het selecteren van medewerkers en ging daarbij in op de beloften die dit thema biedt en op de beperkingen en mogelijkheden voor toekomstig onderzoek. In dit verslag gebruiken wij zijn presentatie als kapstok voor de andere bijdragen.

Don Kluemper, die overigens vòòr zijn wetenschappelijke carrière in de jeugdreclassering werkte, laat zien dat de voorspellende kracht van zelf-beschrijvende persoonlijkheidsvragenlijsten niet erg hoog is ondanks het uitvoerige onderzoek dat ernaar wordt gedaan. Dat wil zeggen dat deze instrumenten niet sterk gerelateerd zijn aan indicatoren van werksucces en dat hun bijdrage in selectiecontext dus voor verbetering vatbaar is. In de wetenschappelijke literatuur wordt daarom steeds meer aandacht besteed aan door kennissen of andere

observanten over een individu ingevulde vragenlijsten. Deze varianten zouden meer voorspellende waarde hebben voor een aantal maten van werksucces.

Don Kluemper vat de voor de personeelspsychologie belangrijkste persoonspredictoren op basis van zelfbeschrijving samen. Zoals bekend is consciëntieusheid de belangrijkste algemene predictor van werksucces. Deze is niet gerelateerd aan intelligentie en heeft een validiteitcoëfficiënt van .30. Op de tweede plaats komt Emotionele Stabiliteit. Daarnaast is Extraversie belangrijk voor verkoopfuncties, Altruïsme voor klantenservice functies en Openheid voor werk waarbij expertise belangrijk is. Opmerkelijk is dat door anderen beoordeelde Openheid veel vaker predictief is voor functiesucces dan zelf-beoordeelde Openheid. De presentatie van Don Kluemper ging over de "normale persoonlijkheid". Marise Born ging in haar presentatie die hieronder wordt besproken daarentegen in op meer psychiatrisch getinte persoonlijkheidspredictoren van functiesucces.

- Sociaal wenselijk antwoordgedrag

Een van de voordelen van door anderen ingevulde persoonlijkheidsvragenlijsten is dat deze de kans op sociaal wenselijk antwoordgedrag verminderen. Sociaal wenselijk antwoordgedrag is een vaak voorkomend probleem als sollicitanten de persoonlijkheidsvragenlijst zelf invullen. Vijftig procent van de sollicitanten zou volgens Don Kluemper de werkelijkheid verdraaien bij het invullen van persoonlijkheidsvragenlijsten. Dit leidt ertoe dat scores op belangrijke persoonlijkheidsdimensies als consciëntieusheid onbetrouwbaar worden, wat er toe kan bijdragen dat mensen op verkeerde plekken terechtkomen. Een aantal persoonlijkheidsvragenlijsten heeft dan ook zogenaamde "fakeschalen". Deze bestaan uit controlevragen als "ik lieg nooit" die tezamen een beeld geven van sociaal wenselijk antwoordgedrag. Deze schalen zijn vaak nogal doorzichtig en kunnen daarnaast ook kunnen correleren met persoonlijkheidstrekken als consciëntieusheid en dus ook met functiesucces. Dat is een probleem, want daardoor zijn ze moeilijk te gebruiken als controlevariabele om de relatie tussen de belangrijke persoonspredictor consciëntieusheid en werksucces te zuiveren van het effect van sociaal wenselijk antwoordgedrag. Dat is van belang omdat je met een assessment een zo zuiver mogelijke inschatting van de geschiktheid van mensen voor functies wilt maken. Als alternatieve methode bespreekt Don Kluemper daarom een methode die gebaseerd is op zogenaamd "overclaring". Daarmee bedoelt hij dat mensen zichzelf vaak meer kennis toedichten dan ze in werkelijkheid hebben. Een en ander meet hij door onderwerpen of personen uit bijvoorbeeld de geschiedenis te gebruiken. De respondent moet dan aangegeven of hij deze stimuli al dan niet kent. Een aantal van de items zijn fictief. Bijvoorbeeld niet-bestaande figuren die tussen echte figuren uit de geschiedenis in staan. De somscore op de fictieve items geeft een maat van sociaal wenselijk antwoordgedrag dat niet gerelateerd is aan consciëntieusheid en daarom gebruikt kan worden als controlevariabele. Deze aanpak werkt in het lab, maar moet nog bevestigd worden in veldstudies. Uit het

publiek kwam de suggestie om naast sociaal wenselijk antwoordgedrag als bedreiger van de validiteit van persoonlijkheidsvragen ook te kijken naar het verhogen van validiteit door de vragen meer aan te laten sluiten bij de leefwereld van de respondent. Bijvoorbeeld door deze aan te passen aan de specifieke context.

- Thema's voor toekomstig onderzoek

We kunnen meer inzicht krijgen in sociaal wenselijk antwoordgedrag door mensen echte kennisvragen voor te leggen die betrekking hebben op de organisatie in kwestie. Het gaat hier dus om instrumenten die zijn afgestemd op de omgeving waarin ze worden toegepast, wat een vorm van contextualisering is. Dit is overigens een thema dat, als het gaat om het verbeteren van de voorspellende kracht van persoonlijkheidsinstrumenten, ook aan bod komt in de presentatie van Djurre Holtrop. Volgens Don Kluemper moet ook aandacht worden besteed aan de relatie tussen sociaal wenselijk antwoordgedrag en intelligentie. Als de score op sociaal wenselijk antwoordgedrag hoger wordt dan neemt de correlatie tussen consciëntieusheid en intelligentie toe. Zoals ook werd uitgelegd bij de vorige bullet, wordt door deze overlap tussen predictoren de bijdrage van consciëntieusheid aan de predictie kleiner. De predictieve kracht van intelligentie en consciëntieusheid tezamen neemt dan af, wat de validiteit van de gehele testbatterij aantast. Het is dus van belang om meer grip te krijgen op sociaal wenselijk antwoordgedrag.

- Door evaluatoren beoordeelde persoonlijkheid

Persoonsfactoren kunnen door anderen bijvoorbeeld worden afgeleid uit foto's, kleding, de mate van opgeruimdheid van het bureau, emailverkeer et cetera. Het kan daarbij ook gaan om zogenaamde "gedragmatige residuen". Dat zijn sporen die mensen nalaten waarvan zij zich niet altijd bewust zijn en die iets zeggen over wie ze zijn. Denk aan via een bonuskaart geregistreerd aankoopgedrag in een supermarkt. Belangrijk uitgangspunt hierbij is het zogenoemde Realistic Accuracy Model (Funder, 1999). Dit gaat ervan uit dat persoonlijkheidstrekken reële menselijke attributen zijn die zich manifesteren in het concrete gedrag van mensen en die gemeten kunnen worden op basis van beschikbare relevante gedragmatige informatie.

- *Facebook studies*

Hier gaat het om de vraag of je uit iemands *facebook*profiel zijn persoonlijkheid kan afleiden en mogelijk daarmee voor de organisatie relevant gedrag kunt voorspellen. Het blijkt zo te zijn dat beoordelaars op basis van algemene instructies in staat zijn om een redelijk betrouwbaar persoonsbeeld af te leiden uit een facebookpagina. Dit profiel blijkt volgens Kluemper ook behoorlijk te correleren met door de betrokkene

zelf ingevulde persoonlijkheidsvragen en met het door hen opgeven cijfergemiddelde op de middelbare school (Grade Point Average). Vervolgonderzoek moet zich ook richten op toekomstige werkprestaties. Naast het gebruik van sociale media blijft uiteraard ook face-to-face contact van belang bij het werven van kandidaten voor functies. Speciale aandacht gaat daarom uit naar de invloed van persoonlijkheid op sollicitantengedrag in interviews. Naast intelligentie blijken extraversie en emotionele intelligentie van belang blijken te zijn. Door kennissen beoordeelde persoonlijkheid blijkt hierbij over het algemeen meer voorspellende kracht te hebben dan zelf-beoordeelde persoonlijkheid.

Iris Egberink van de Rijksuniversiteit Groningen hield een presentatie over het onderzoek naar meetinvariantie in een echte selectiecontext.

Het gaat bij dit onderwerp om het vraagstuk of testitems en -schalen zich hetzelfde gedragen voor verschillende groepen. Het is de bedoeling dat personen die eenzelfde positie op de te meten dimensie innemen een gelijke kans hebben om een bepaald antwoord op een vraag te geven. De groep waartoe ze behoren mag hierop verder geen invloed hebben. Het zoveel mogelijk uitsluiten van discriminatie door het psychologisch instrumentarium is een van de onderwerpen waarmee de Commissie Testaangelegenheden Nederland (COTAN) van het NIP zich bezighoudt. Iris is naast haar werk in Groningen eindredacteur van deze commissie. Het verhaal van Iris ging over discriminatoire effecten die onbedoeld kunnen voortvloeien uit het gebruik van bepaalde instrumenten. Er zijn in ons vakgebied ook andere vormen van discriminatie. Zo gaat Jeroen Decoster in zijn presentatie in op discriminatie in sollicitatiesituaties.

Iris laat zien hoe het bovenbeschreven "differential item functioning" (of *dif*) met een *psychometrische aanpak* voorkomen kan worden. Om dit te detecteren zijn er statistische toetsen, zogenaamde "*dif* tests". Er zijn twee onderzoeksbenaderingen: 1) confirmatieve factoranalyse en 2) item respons theorie modellen. Iris paste de tweede methode toe. Het gaat om technische materie die niet gemakkelijk in eenvoudige taal is uit te leggen. Hieronder doen wij een poging. De lezer die daaraan geen behoefte heeft kan dit overslaan.

Technisch intermezzo

Het probleem is dat als zo'n *dif* test significant is, je nog niet weet hoe groot de afwijking precies is. Pas als de effectgrootte (vergelijkbaar met bijvoorbeeld de verklaarde variantie van een lineair regressiemodel) groot genoeg is, kan je zeggen dat een gevonden resultaat naast statistisch significant ook praktisch relevant is. Bij grote steekproeven geven bijna alle items statistisch gezien blijk van *dif*. De gebruikte likelihood ratio test (LRT) is namelijk erg gevoelig voor de steekproefgrootte. De LRT vergelijkt items die invariant zijn over de groepen met de andere items. De invariante items zijn de zogenaamde "ankeritems". Het probleem is dat niet van tevoren bekend is welke items ankeritems zijn, dus dat moet empirisch worden vastgesteld. Er zijn twee methoden om dit te doen. Bij de eerste methode wordt een LRT gebruikt met alle items als voorlopige ankers. Vervolgens worden de items waarvan de LRT resultaten niet significant zijn en die de grootste discriminatieparameter (vergelijkbaar met de item-totaalcorrelatie uit de klassieke testtheorie) hebben als ankers gebruikt, waarbij het aantal ankeritems gevarieerd wordt. Bij de tweede methode vindt eerst een rangordening van de LRT-statistieken plaats en worden de items met de laagste waarden geselecteerd om als ankers te gebruiken. Het bleek daarbij het beste om (bij tests vanaf 20 items) uit te gaan van drie en vijf ankeritems. De eerste methode is geschikt voor kleinere steekproeven (dan zijn immers niet a priori alle itemstatistieken significant) en de rangordemethode is meer geschikt voor grotere steekproeven. Voor de geselecteerde items wordt vervolgens de effectgrootte bepaald.

Iris gebruikte twee empirische steekproeven met consciëntieusheidscores. Een grote steekproef met als groepsindeling selectie versus ontwikkelingsgerichte toepassing. Daarnaast een kleinere steekproef met als groepsindeling allochtoon versus autochtoon. De analyses laten zien dat de effectgroottes, als *dif* optreedt, klein zijn. Dat betekent dat de items in dit geval praktisch gezien niet discrimineren tussen de verschillende groepen en er dus geen noodzaak is om aparte testvragen te ontwikkelen. Een en ander onderstreept dat het als algemeen methodologisch principe zeer belangrijk is om niet alleen naar statistische significantie te kijken, maar ook naar effectgroottes.

Djurre Holtrop van de Vrije Universiteit Amsterdam hield een presentatie over het voorspellen van prestaties met gecontextualiseerde persoonlijkheidsvragen.

Een recente metastudie (Shaffer & Postlethwaith, 2012) toont dat gecontextualiseerde persoonlijkheidsvragen beter prestaties voorspellen dan niet-gecontextualiseerde vragen. Djurre vergelijkt in zijn onderzoek twee manieren van contextualiseren met elkaar. "Tagged contextualization" wil zeggen dat er een referentiekader, bijvoorbeeld "op het werk" aan bestaande testitems wordt toegevoegd. Volledige contextualisering betekent in dit geval dat de bestaande testitems geheel worden aangepast aan de context. Djurre voerde een veldstudie uit onder apothekersassistenten. Deze vulden naast de twee gecontextualiseerde vragenlijsten ook een standaardvragenlijst in. Tevens waardeerden zij alle instrumenten op een aantal aspecten en beoordeelde hun leidinggevende hun prestatie. Uit de resultaten bleek dat de respondenten de volledig gecontextualiseerde vragenlijst het meest waardeerden, vervolgens de tagged vragenlijst en de standaardvragenlijst het minst. In afwijking van de metastudie bleek echter dat beide gecontextualiseerde vragenlijsten lager correleerden met de

prestatie-maten. Het is op dit moment nog onduidelijk hoe deze bevinding kan worden verklaard. Door de toehoorders werd de suggestie gedaan om te kijken naar het type situatie. Het zou kunnen dat de apotheekcontext in bepaalde opzichten die we nu nog niet kennen afwijkt van situaties die in de metastudie en andere recente literatuur worden vermeld (zie bijvoorbeeld Butter & Born, 2012).

Jeroen Decoster van de Universiteit Gent hield een presentatie over etnische discriminatie in het screenen van cv's.

Jeroen gebruikte de methode van "correspondentie onderzoek". Dit houdt in dat de onderzoekers gefingeerde brieven sturen naar aanleiding van echte vacatures. Omdat deze onderzoeksmethode uitgaat van reëel bestaande vacatures is de ecologische validiteit relatief hoog. De brieven werden experimenteel gemanipuleerd. In dit geval varieerden de onderzoekers de naam (Belgisch versus Arabisch) en de affiliatie (jeugdvereniging versus Marokkaanse vereniging). Alle andere informatie bleef constant. De afhankelijke variabele was de reactie (ja, nee, geen antwoord). Verder keken de onderzoekers ook naar het verschil in type recruteerder (personeelsafdeling versus recruiteringsbureau). Als analysemethode werd logistische regressieanalyse gebruikt. De resultaten bevestigden eerder onderzoek, namelijk dat recruteerders Arabische cv's vaker afwijzen dan Belgische. Er was echter geen effect van affiliatie. Verder bleek dat discriminatie voornamelijk voorkomt bij recruiteringsbureaus. Dit kan onder andere te maken hebben met specifiek beleid van deze organisaties en de klantrelatie met de opdrachtgever. Door anoniem te solliciteren, recruteerders beter te trainen op ethische aspecten, sancties op te leggen vanuit de overheid of mystery sollicitanten in te zetten kan men deze problemen wellicht beter aanpakken.

Vanuit het publiek kwam een kritische vraag over de methode. Deze laat immers de echte, niet door de onderzoekers verstuurd reacties op de vacatures, die ook invloed kunnen hebben op de succeskans van de gefingeerde sollicitaties, buiten beschouwing. De onderzoekers gaven aan dat de door hen gehanteerde methode hiervoor correcties inbouwt. Bijvoorbeeld door de gefingeerde brieven steeds per vacature in batches te versturen en alleen de batches te analyseren waarop bij hen reacties zijn binnengekomen.

Marise Born van de Erasmus Universiteit Rotterdam hield een presentatie over de wijze waarop assessoren risicopersoonlijkheid wegen

De laatste jaren is er meer aandacht voor de donkere kant van de persoonlijkheid bij selectievraagstukken. O' Boyle, Jr., Forsythe, Banks en McDaniel hebben over dit thema in 2012 een meta-analyse gepubliceerd. Het gaat hierbij om de meer psychiatrische kant van de persoonlijkheid. Aspecten hiervan zijn bijvoorbeeld psychopathie, machiavellisme en narcisme. Recente schandalen in bijvoorbeeld de bankwereld laten zien dat er op dit vlak in organisaties het nodige mis lijkt te gaan. Andere onderzoekers onderscheiden afwijkend gedrag dat de eigen onzekerheid

vermindert door mensen te intimideren, te vermijden (te wantrouwen) dan wel te "pleasen".

De onderzoeksvraag was hoe assessoren risicopersoonlijkheid meewegen in hun uiteindelijke beoordeling van sollicitanten. Het ging hierbij om de dimensies persoonlijke kracht, interpersoonlijke vaardigheden, cognitieve vaardigheden en taakgerelateerde vaardigheden (resultaat- en kwaliteitsgerichtheid). De onderzoekers vonden alleen een negatieve correlatie voor vermijdend, wantrouwend gedrag, dat persoonlijke, interpersoonlijke en cognitieve vaardigheden negatief voorspelde. De overige correlaties waren niet significant. Vervolgonderzoek zal zich onder andere richten op andere "dark side predictoren", het meewegen van de resultaten uit rollenspellen en interviews en specifieke functies waarvoor onaangepast gedrag een groot risico is.

Recente artikelen

Butter, R. & Born, M.Ph. (2012). Enhancing criterion-related validity through bottom-up contextualization of personality inventories: Construction of an ecological conscientiousness scale for Ph.D. candidates. *Human Performance*, 25:4, 303-317.

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception*. San Diego, CA, US: Academic Press, Inc.

O' Boyle, E. H. Forsythe, D. R. Banks, G. C. & McDaniel, M. A. (2012). A meta-analysis of the dark triad and work behavior: A social exchange perspective. *Journal of Applied Psychology*, 97, 557-579.

Shaffer, A. & Postlethwaith, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, 65, 445-494.