De comparatieve beoordelingsmethode voor een betrouwbare en valide cv-screening: een vergelijking tussen experts en studenten

Anneleen V. Mortier, Renske Bouwer, Liesje Coertjens, Ellen Volckaert, Amelie Vrijdags, Roos Van Gasse, Peter Vlerick & Sven De Maeyer[*]

In de praktijk en uit eerder empirisch onderzoek blijkt dat cv-screening niet altijd zorgt voor geschikte kandidaten voor een vacature. Verschillende zaken kunnen hiervan de oorzaak zijn: één beoordelaar voert de screening uit, waardoor cognitieve vertekeningen het selectieproces kunnen beïnvloeden; de beoordeling focust niet op alle relevante aspecten van de selectie of laat bepaalde criteria harder doorwegen dan andere; en/of de beoordelaar heeft onvoldoende expertise om de cv-screening uit te voeren. De huidige studie komt aan deze tekortkomingen tegemoet door de alternatieve beoordelingsmethode, comparatieve vergelijking, te beschrijven en de interbeoordelaarsbetrouwbaarheid en constructvaliditeit ervan voor cv-screening na te gaan. In deze studie is gebruikgemaakt van een bestaande vacature waarvoor 42 kandidaten hun cv hebben ingestuurd. Deze cv's zijn comparatief met elkaar vergeleken door ervaren (experts; N = 7) en minder ervaren beoordelaars (studenten; N = 57). De resultaten tonen aan dat comparatieve oordelen van ervaren beoordelaars samenhangen met een valide en betrouwbare cv-screening. De interbeoordelaarsbetrouwbaarheid van de oordelen van de studenten was lager dan de interbeoordelaarsbetrouwbaarheid van de oordelen van de experts. Hoewel er een sterke correlatie was tussen de rangorde van de ervaren beoordelaars en de rangorde van de studenten, lieten de studenten hun oordeel vaker afhangen van irrelevante aspecten.

1 Introductie

Het screenen van curricula vitae (cv's) helpt organisaties om snel potentieel geschikte kandidaten te selecteren voor een bepaalde vacature (Born, 2008; Burns, Christainsen, Morris, Periard & Coaster, 2014; Derous, Ryan & Serlie, 2014). In de

* Anneleen Mortier en Peter Vlerick zijn verbonden aan de Universiteit Gent, Vakgroep Personeelsbeleid, Arbeids- en Organisatiepsychologie. Correspondentieadres: Anneleen Mortier, Vakgroep Personeelsbeleid, Arbeids- en Organisatiepsychologie, Henri Dunantlaan 2, B-9000 Gent, e-mail: Anneleen.Mortier@UGent.be. Renske Brouwer is werkzaam aan de Vrije Universiteit Amsterdam, Pedagogische Onderwijswetenschappen. Op het moment van deze studie was zij werkzaam bij Universiteit Antwerpen, Afdeling Opleidings- en Onderwijswetenschappen. Roos Van Gasse en Sven De Maeyer waren op het moment van deze studie werkzaam bij Universiteit Antwerpen, Afdeling Opleidings- en Onderwijswetenschappen. Liesje Coertjens is werkzaam bij Université Catholique de Louvain, Psychological Sciences Research Institute. Ellen Volckaert en Amelie Vrijdags zijn werkzaam bij HR-adviesverlener Hudson België.

meerderheid van de gevallen gebeurt dit aan de hand van schriftelijke cv's, die onder andere persoonlijke informatie en informatie over de gevraagde vaardigheden en competenties bevatten. Een cv is meestal het eerste wat een recruiter te zien krijgt van een kandidaat. Als gevolg daarvan proberen kandidaten om in hun cv een zo goed mogelijke eerste indruk te maken op de recruiters of beoordelaars van hun cv. Eerder onderzoek laat echter zien dat beoordelaars niet altijd in staat zijn om de meest geschikte kandidaten te selecteren op grond van de informatie in schriftelijke cv's. Zo laten zij zich bij het beoordelen nog te veel beïnvloeden door oppervlakkige en irrelevante kenmerken in een cv, zoals de lay-out of spelfouten (Derous et al., 2014). Ook kunnen onbewuste (negatieve of positieve) associaties met bepaalde persoonseigenschappen van de kandidaat invloed hebben op de beoordeling (Burns et al., 2014). Bijvoorbeeld wanneer men aangeeft dat men lid is/was van een professionele vereniging, wordt dit geassocieerd met consciëntieuze kandidaten (Cole, Feild, Giles & Harris, 2009). Dit kan worden gezien als een positieve eigenschap, aangezien consciëntieusheid geassocieerd is met productievere werknemers (bijv. Hurtz & Donovan, 2000). Verder kan het ook zijn dat beoordelaars verschillen in de strategie die ze gebruiken om de cv-screening uit te voeren (Fritzsche & Brannick, 2002). Deze zaken worden nog problematischer wanneer de screening slechts door één persoon wordt uitgevoerd en de selectie van kandidaten afhankelijk is van deze ene beoordelaar.

Dit onderzoek bestudeert een alternatieve beoordelingsmethode voor cv-screening: comparatief beoordelen (Lesterhuis, Verhavert, Coertjens, Donche & De Maeyer, 2017). In deze methode worden de te beoordelen producten niet één voor één beoordeeld, maar comparatief. Dit comparatieve proces blijkt voor beoordelaars veel gemakkelijker te zijn (Thurstone, 1927). Ook voorkomt het direct vergelijken van producten met elkaar dat beoordelaars terugvallen op hun eigen, vaak onbewuste, standaarden (Laming, 2004). Cognitieve vertekeningen van individuele beoordelaars hebben zo een minder groot effect op de uiteindelijke oordelen. Voorgaand onderzoek onderzocht de comparatieve beoordelingsmethode in de onderwijscontext en biedt reeds evidentie dat deze methode leidt tot betrouwbare en valide beslissingen over de kwaliteit van verschillende onderwijsproducten (bijv. Lesterhuis et al., 2017; Van Daal, Lesterhuis, Coertjens, Donche & De Maeyer, 2017).

Het is echter nog onduidelijk in hoeverre comparatief beoordelen ook van meerwaarde is voor cv-screening. In dit onderzoek wordt daarom de interbeoordelaarsbetrouwbaarheid en constructvaliditeit van cv-screening nagegaan op basis van comparatieve beoordelingen.

2 Theoretische achtergrond

2.1 Belang van kwaliteitsvolle cv-screening

De eerste screening van kandidaten in het selectieproces vindt meestal plaats op basis van de beoordeling van schriftelijke cv's (Piotrowski & Armstrong, 2006; Roe, 1983; Steiner, 2012). Over het algemeen verloopt deze screening systematisch,

waarbij de functievereisten op voorhand zijn vastgelegd (Roe, 1983). Het voordeel van cv-screening is dat men kan vaststellen of een kandidaat de gevraagde kennis, ervaring en vaardigheden heeft vooraleer er meer tijdrovende en duurdere methodes worden gebruikt, zoals vaardigheidstesten en simulaties (Cole et al., 2009). Met andere woorden, cv-screening is een soort van eerste horde die kandidaten moeten nemen vooraleer ze een volgende stap in de selectieprocedure kunnen nemen. Doordat het een initiële screening is van de meest geschikte kandidaten, is zowel de betrouwbaarheid als de validiteit van deze screening van groot belang. Verkeerde beslissingen kunnen er immers voor zorgen dat potentieel geschikte kandidaten niet geselecteerd worden. Dit kan zowel voor de kandidaat als de organisatie negatieve gevolgen hebben (Sutherland & Wöcke, 2011).

2.2 Interbeoordelaarsbetrouwbaarheid van cv-screening

Bij een betrouwbare cv-screening kan worden verondersteld dat met andere, vergelijkbare beoordelaars dezelfde kandidaten geselecteerd zouden worden (interbeoordelaarsbetrouwbaarheid). Een manier om de interbeoordelaarsbetrouwbaarheid van cv-screening te waarborgen is door vooraf criteria te definiëren op basis waarvan een oordeel over elke kandidaat wordt geveld (Jonsson & Svingby, 2007). Deze criteria kunnen bestaan uit de minimumvereisten van een baan, zoals een relevante vooropleiding of werkervaring, maar ook uit de mate waarin cv's meer of minder aansluiten bij de functievereisten. In dit geval dienen beoordelaars verschillende aspecten van de functievereisten tegen elkaar af te wegen om tot een algemeen oordeel te komen. Doordat niet elke beoordelaar elk aspect van een functie even belangrijk vindt, kunnen er vooral bij deze meer subjectieve beslissingen grote verschillen tussen beoordelaars optreden (Knouse, 1994). Zo kunnen beoordelaars die bijvoorbeeld een voorkeur hebben voor kandidaten die in hun cv een volledig beeld weergeven van zichzelf, met zowel de positieve als negatieve kanten, tot een geheel andere selectie komen dan beoordelaars die een voorkeur hebben voor kandidaten die slechts een (positief) deel van zichzelf tonen (Highhouse & Hause, 1995). Wanneer er slechts één beoordelaar betrokken is bij de cv-screening, is de selectie sterk afhankelijk van de individuele voorkeuren van deze ene beoordelaar en kan deze mogelijk niet worden gegeneraliseerd naar de organisatie als geheel (Derous et al., 2014; Knouse, 1994; Seibert, Williams & Raymark, 2010).

2.3 Constructvaliditeit van cv-screening

Naast betrouwbaarheid is ook validiteit van groot belang voor kwaliteitsvolle cv-screening. Hierbij is het vooral de vraag of beoordelaars alle voor de vacature relevante aspecten van het cv in beschouwing nemen en zich daarbij niet laten leiden door irrelevante aspecten. Alleen in dit geval kan worden gesproken van een constructvalide screening van cv's (Cohen, Manion & Morrison, 2007).

De constructvaliditeit van cv-screening kan om verschillende redenen in het gedrang komen. Ten eerste proberen beoordelaars vaak meer dan enkel de feitelijke informatie af te leiden uit een cv om tot een beslissing te komen. Zo proberen beoordelaars bijvoorbeeld persoonlijkheidstrekken uit cv's te herleiden, aangezien

sommige persoonlijkheidstrekken (zoals bijvoorbeeld consciëntieusheid) een goede voorspeller zijn van latere werkprestaties (Sackett, Lievens, Van Iddekinge & Kuncel, 2017). Echter, beoordelaars zijn niet altijd in staat om deze persoonlijkheidstrekken accuraat uit cv's te halen (Cole et al., 2009). Sterker nog, ze blijken dit zelfs te doen wanneer een cv enkel informatie bevat over bijvoorbeeld opleiding en ervaring (Burns et al., 2014). Het kan bijvoorbeeld zijn dat wanneer een kandidaat een cv instuurt met verschillende kleuren en een atypische lay-out, deze kandidaat door de beoordelaar gezien wordt als een creatief persoon. Dit heeft een negatief effect op de constructvaliditeit van de uiteindelijke beslissingen in de cv-screening.

Ten tweede kunnen vooroordelen (onbewust) een rol spelen in het selectieproces (Webster, 1964). Zo kunnen enkele aspecten van een cv zoals naam, geslacht, ras en/of sociaaleconomische status een negatief effect hebben op het algemene oordeel over de kandidaat (Cotton, O'Neill & Griffin, 2008). Kandidaten met een naam kenmerkend voor een etnische minderheid worden bijvoorbeeld vaker afgewezen in vergelijking met kandidaten die een naam hebben kenmerkend voor een etnische meerderheid (Bertrand & Mullainathan, 2004; Riach & Rich, 2002; Zegers de Beijl, 2000), zelfs wanneer de cv's verder gelijkaardig zijn (Kaas & Manger, 2012). De kans om afgewezen te worden ligt ongeveer 2.7 tot 7.3 keer hoger voor cv's met een naam van een etnische minderheid (Derous, 2007). In lijn hiermee blijkt dat recruiters sneller kandidaten afwijzen met een zwarte huidskleur (Maddox, 2004). Ook worden vrouwelijke kandidaten die zichzelf promoten in video-cv's minder geschikt gevonden voor een baan (Waung, Hymes, Beatty & McAuslan, 2015).

Samenvattend laten bovenstaande studies zien dat recruiters vaak beslissingen nemen op basis van incorrecte aannames, irrelevante informatie en negatieve associaties met eigenschappen van de kandidaat. Een manier om hieraan te proberen tegemoet te komen is het anonimiseren van cv's om zo meer gelijke kansen aan te bieden voor iedereen (Voncken & Westendorp, 2007). Dit is echter geen volledige oplossing, aangezien er ook nog andere verwijzingen in het cv kunnen staan naar de etnische achtergrond van de kandidaat, zoals geboorteplaats of lidmaatschap van verenigingen (Derous, 2011; Derous, Nguyen & Ryan, 2009). Een andere oplossing is om een divers team van beoordelaars te betrekken bij de cv-screening (Barber, 1998). Hierdoor kunnen bepaalde vooroordelen, verschillende strategieën en aannames worden opgevangen door het team. In deze studie maken we daarom ten eerste gebruik van meerdere beoordelaars. Bovendien hanteren we een beoordelingsmethode waarbij deze beoordelaars cv's beoordelen door een reeks van paarsgewijze vergelijkingen. We gaan na in hoeverre deze aanpak leidt tot constructvalide beoordelingen.

2.4 Comparatieve beoordeling van cv's

Comparatief beoordelen is gebaseerd op het principe van Thurstone (1927) zoals beschreven in *the law of comparative judgment*. Hierbij worden producten niet op zichzelf beoordeeld maar altijd in vergelijking tot elkaar. Toegepast op de thema-

tiek die in dit artikel aan de orde komt, houdt dit in dat beoordelaars de cv's steeds in paren met elkaar vergelijken en voor elk paar beslissen welke van de twee cv's beter past bij een vacature. Comparatief beoordelen is een intuïtieve manier van beoordelen. Laming (1990) beschrijft hoe beoordelaars tijdens het proces van oordelen altijd een vergelijking maken. Dat kan een vergelijking zijn met producten die eerder zijn gezien, maar ook een vergelijking met een meer abstract referentiekader dat iemand in de loop van de tijd heeft opgebouwd (Laming, 1990). Waar dit soort vergelijkingen echter normaliter impliciet blijven, worden deze vergelijkingen bij comparatief beoordelen expliciet gemaakt. Bovendien worden bij comparatief beoordelen cv's niet slechts één keer beoordeeld, maar verschillende malen door verschillende beoordelaars, en steeds in vergelijking met een willekeurig ander cv.
Op basis van deze reeks van paarsgewijze vergelijkingen kan een rangorde worden gevormd van minst geschikte naar meest geschikte kandidaat volgens een Raschmodel (Bradley-Terry-Luce-Model; Bradley & Terry, 1952; Luce, 1959). Deze rangorde geeft de consensus van de beoordelaars weer over wat een geschikt cv is voor de vacature. Deze rangorde kan vervolgens als input dienen voor de uiteindelijke selectie.

In de onderwijscontext is reeds aangetoond dat deze methode leidt tot betrouwbare en valide oordelen wanneer complexe competenties moeten worden beoordeeld (Bramley, 2015; Pollitt, 2012; Van Daal et al., 2017; Verhavert, Bouwer, Donche & De Maeyer, 2019). De betrouwbaarheid bij comparatieve beoordelingen geeft weer hoe stabiel de rangorde is tussen beoordelaars (Coertjens, Lesterhuis, Verhavert, Van Gasse & De Maeyer, 2017).
Uit een recente meta-analyse met 49 verschillende paarsgewijze beoordelingsreeksen (met verschillende producten, zoals verslagen en portfolio's), bleek een gemiddelde betrouwbaarheid van 0.78. Ook is gebleken dat het proces van paarsgewijs vergelijken snel gaat, sneller dan tot het komen van analytische oordelen (Coertjens et al., 2017). De minimale betrouwbaarheid van 0.80 kan al behaald worden wanneer elk product ongeveer 17 keer wordt vergeleken met willekeurige andere producten (Verhavert et al., 2019. Wat de validiteit betreft, bestudeerde Lesterhuis (2018) het comparatief beoordelen bij argumentatieve teksten. Dit zijn teksten waarbij leerlingen een standpunt moeten innemen en waarbij ze dit standpunt moeten beargumenteren door middel van bronnen. Haar studie liet zien dat zowel ervaren als minder ervaren beoordelaars tijdens het vergelijken letten op relevante aspecten van argumentatieve teksten, zoals de kwaliteit van de argumentatie, de argumentatiestructuur en brongebruik. Er werd zelfs op veel meer relevante aspecten gelet dan wat in een beoordelingsschema met specifieke criteria omschreven kan worden. Ook bleek uit dit onderzoek dat er nuanceverschillen waren tussen beoordelaars in de aspecten waarop zij hadden gelet tijdens het vergelijken. Waar sommige beoordelaars bijvoorbeeld meer aandacht hadden voor de inhoud van de teksten, hadden andere beoordelaars meer aandacht voor het brongebruik. Doordat deze comparatieve beoordelingsmethode steunt op de vergelijkingen van verschillende beoordelaars, worden deze verschillende beoordelaarsperspectieven automatisch samengenomen in de uiteindelijke oordelen (Pollitt, 2012; Van Daal et al., 2017).

Door meerdere beoordelaars in het beoordelingsproces te betrekken geeft de uiteindelijke beslissing de consensus over de beoordelaars weer (Lesterhuis, 2018). Mogelijke effecten van vooroordelen, strategieën en aannames van individuele beoordelaars worden zo opgevangen (Bramley, 2007; Whitehouse, 2012; Whitehouse & Pollitt, 2012). Waar het betrekken van meerdere beoordelaars normaal gesproken altijd veel organisatie vergt in termen van het vooraf afstemmen van de beoordelingscriteria, eventueel tussentijds overleg over individuele kandidaten, en het komen tot een uiteindelijke selectie op basis van soms zeer uiteenlopende oordelen, is het betrekken van beoordelaars bij de comparatieve beoordeling relatief eenvoudig. Paren kunnen gemakkelijk worden verspreid over verschillende beoordelaars en niet iedereen hoeft elk paar te beoordelen om toch generaliseerbare resultaten te verkrijgen (Whitehouse & Pollitt, 2012). Digitale comparatieve beoordelingstools zoals D-PAC (www.d-pac.be) helpen om deze paren automatisch te verspreiden over beoordelaars en om op basis van individuele vergelijkingen een rangorde samen te stellen. Op basis van de rangorde en de interbeoordelaarsbetrouwbaarheid in oordelen kan vervolgens een omslagpunt (bijv. cut-off) worden bepaald, waarboven cv's voldoende geschikt worden geacht voor de vacature. Dit geeft meer transparantie in de uiteindelijke selectie van kandidaten.

Hoewel tot op heden nog niet bestudeerd, lijkt de comparatieve beoordelingsmethode daarmee ook in de context van cv-screening een oplossing te bieden voor de problemen die reeds genoemd zijn met betrekking tot de constructvaliditeit en interbeoordelaarsbetrouwbaarheid van cv-screening. Of comparatief beoordelen inderdaad leidt tot betrouwbare en valide oordelen, wordt onderzocht in deze studie.

2.5 Rol van expertise

Omdat comparatief beoordelen een meer intuïtieve methode is, kan de vraag worden gesteld of de expertise van de beoordelaar een effect heeft op de interbeoordelaarsbetrouwbaarheid en de constructvaliditeit van de cv-screening. In een aantal eerdere studies naar cv-screening werd gebruikgemaakt van studenten als beoordelaars (bijv. Blommaert, Van Tubergen & Coenders, 2012; Ziegert & Hanges, 2005). Het is echter onduidelijk wat de impact is van deze expertise (student-novice versus expert) op de interbeoordelaarsbetrouwbaarheid en constructvaliditeit van de beoordeling bij comparatief beoordelen.

Eerder onderzoek (Verhavert, 2018) geeft aan dat vergelijkingen maken, wat men bij het comparatief beoordelen doet, mogelijk meer cognitief belastend is voor beginners (novices). Dit lijkt zeker het geval wanneer een complexe beslissing gemaakt moet worden op basis van de afweging van verschillende aspecten in een cv, zoals de vooropleiding, eerdere werkervaring en andere relevante verworven competenties. Zo gebruiken experts in cv-screening hiervoor andere beslissingsstrategieën: ze gebruiken nieuwe informatie om hun initiële beslissing aan te passen, ze leren van eerdere ervaringen en ze gebruiken strategieën om bepaalde vertekeningen te onderdrukken (Shanteau, 1988). Daarnaast detecteren ze andere

informatie dan studenten, weten ze beter wat relevante en irrelevante informatie is en hebben ze meer domeinkennis waardoor ze waarschijnlijk een juistere beslissing maken (Shanteau, 1988). Bovendien geven expert-beoordelaars meer aandacht aan informatie die de toekomstige werkprestaties voorspelt (zoals ervaring en specifieke kennis), in vergelijking met minder ervaren beoordelaars (Hanak, Sirota & Juanchich, 2013), ook wanneer ze onder tijdsdruk staan (Cavojova & Hanak, 2014). Dit zou komen omdat het zoeken naar relevante informatie in het cv voor experts minder cognitief belastend is dan voor student-beoordelaars (Chase & Simon, 1973). Student-beoordelaars, daarentegen, proberen alle informatie mee te laten wegen in hun oordeel, zelfs de minder relevante (Shanteau, 1992).

Gezien de vrijheid bij comparatief beoordelen beslissen beoordelaars zelf welke kenmerken van de twee cv's in een vergelijking de doorslag geven. Bijgevolg kunnen de verschillen in beslissingsstrategieën tussen expert- en student-beoordelaars tot uitdrukking komen in de beslissingen die worden gemaakt. Dit zou betekenen dat experts en studenten niet alleen verschillen in de mate van overeenstemming over geschikte en minder geschikte cv's, maar ook dat ze tot een geheel andere selectie kunnen komen. Om dit te onderzoeken stellen we de exploratieve vraag of de expertise van beoordelaars een effect heeft op respectievelijk de interbeoordelaarsbetrouwbaarheid en de constructvaliditeit van de cv-selectie via comparatief beoordelen.

2.6 Het doel van deze studie

Tot nu toe zijn er geen studies beschikbaar over de psychometrische karakteristieken van het comparatief beoordelen van cv's. Daarom is het eerste doel van de huidige studie om de interbeoordelaarsbetrouwbaarheid en constructvaliditeit van comparatief beoordelen voor cv-screening te onderzoeken. Met andere woorden, de huidige studie gaat na in welke mate men op grond van een reeks comparatieve beoordelingen uitgevoerd door verschillende beoordelaars tot een betrouwbare en valide selectie van geschikte cv's kan komen.

Het tweede doel van deze studie is om te onderzoeken of de interbeoordelaarsbetrouwbaarheid en de constructvaliditeit van comparatieve beoordelingen van cv's verschillen tussen expert- en student-beoordelaars. We formuleren dan ook twee onderzoeksvragen: (1) In hoeverre is de interbeoordelaarsbetrouwbaarheid van de oordelen van studenten over de geschiktheid van cv's vergelijkbaar met de interbeoordelaarsbetrouwbaarheid van expert-beoordelaars? En: (2) In hoeverre komen de oordelen van de experts en studenten overeen: letten ze tijdens het vergelijken op dezelfde aspecten en komen ze tot dezelfde cv-selectie (constructvaliditeit)?

3 Methode

3.1 Participanten

Zeven experts hebben de cv's beoordeeld; zij waren allen werkzaam bij een Belgische HR-adviesverlener, gespecialiseerd in werving, selectie en training, waar de vacature ook was uitgezet. Deze beoordelaars hadden vanuit hun professionele context dus allen expertise op het gebied van de gevraagde competenties en met het uitvoeren van cv-screening. De studenten hadden geen ervaring met het uitvoeren van cv-screening, aangezien dit nog niet aan bod was gekomen in hun opleidingscurriculum op het moment van de dataverzameling. Alle expert-beoordelaars waren vrouwen; zij waren gemiddeld 4.5 jaar werkzaam bij hun huidige werkgever (variërend van een half jaar tot 12 jaar).

De groep student-beoordelaars bestond uit 57 studenten uit het derde jaar van de bacheloropleiding Arbeids- en Organisatiepsychologie aan een Vlaamse Universiteit. Hiervan was 81% vrouw. De modus van de leeftijd van de studenten was 21 jaar.

3.2 Materiaal

In dit onderzoek is gebruikgemaakt van 42 schriftelijke cv's die door kandidaten zijn opgestuurd naar aanleiding van een reële vacature voor een Talent Acquisition Officer. Hierbij werd gezocht naar kandidaten die in het bezit waren van een masterdiploma, met minstens vijf jaar ervaring in Human Resources. Verder werd vermeld dat de kandidaten dynamisch, flexibel en tweetalig (Nederlands en Frans) dienden te zijn (zie de bijlage voor de volledige vacatureomschrijving). In de vacaturetekst waren geen specifieke eisen gesteld aan de vorm en inhoud van de cv's. Bij alle cv's vermeldden de kandidaten de volgende aspecten: personalia, opleiding, ervaring, vaardigheden, interesses. Soms was er informatie beschikbaar over hobby's en persoonlijke interesses. Er waren 30 vrouwelijke kandidaten en 12 mannelijke. De gemiddelde leeftijd van de kandidaten was 34 jaar en hun leeftijd varieerde van 22 jaar tot en met 58 jaar.

3.3 Procedure

Zowel expert- als student-beoordelaars kregen de instructie om de cv's te screenen met als criterium de geschiktheid van de kandidaat voor de vacature 'talent acquisition officer' (zie bijlage). Dit deden ze aan de hand van de comparatieve beoordelingsmethode, die werd ondersteund door D-PAC (www.d-pac.be). In dit digitale platform zijn alle 42 cv's allereerst anoniem en versleuteld opgeslagen, waarna door middel van een voorgeprogrammeerd algoritme de cv's automatisch in random paren aan de beoordelaars werden toegewezen. Hiervoor zijn twee aparte assessments in D-PAC opgezet, één voor de studenten en één voor de experts. Met een persoonlijke login voor een van de assessments kregen de beoordelaars toegang tot het digitale platform, waar zij voor elk van de toegewezen paren dienden aan te geven welke van de twee kandidaten (cv's) zij beter vonden voor de functie van 'talent acquisition officer'. De beoordelaars zijn hierbij vrijgelaten in hoe ze de specifieke informatie in de cv's met elkaar vergeleken om tot een oordeel te

komen. De beoordelaars konden gedurende de reeks vergelijkingen te allen tijde de vacaturetekst raadplegen om te weten wat de relevante criteria voor de functie waren (o.a. opleiding, werkervaring, taalvaardigheid, zie ook de bijlage). Om meer inzicht te krijgen in de aspecten waarop de beoordelaars tijdens het vergelijken hadden gelet, dienden de beoordelaars na elke vergelijking expliciet aan te geven waarom ze de ene cv beter vonden dan de andere. Zij waren hierin vrij om te bepalen op welke manier deze toelichting werd gegeven. De instructies waren dezelfde voor zowel de experts als de studenten.

Doordat het random algoritme in D-PAC rekening houdt met het aantal keer dat een cv reeds is vergeleken, zijn de cv's binnen elk assessment ongeveer even vaak vergeleken met een willekeurig ander cv. Ook zijn alle cv's random verspreid over de verschillende beoordelaars, waardoor de uiteindelijke beoordelingen gegeneraliseerd kunnen worden over deze groep beoordelaars, ook al hebben ze niet elke vergelijking gemaakt (Van Daal et al., 2017). Aangezien er meer studenten dan experts deelnamen aan het onderzoek, waren er gemiddeld meer vergelijkingen in de studentengroep gemaakt dan in de expertgroep. Bij de experts is ieder cv tussen de 21 en 23 keer beoordeeld, waarbij één van de experts 43 vergelijkingen heeft gemaakt en de andere zes experts elk 70 vergelijkingen. In totaal resulteerde dit in 463 vergelijkingen. Dit is 53.8% van alle mogelijke paren van de 42 cv's (463/861 vergelijkingen = 53.7%). Bij de studenten is elk cv 28 tot 30 keer beoordeeld, met in totaal 613 vergelijkingen (71.2% van het totaal aantal mogelijke vergelijkingen). De studenten maakten ieder 11 beoordelingen, met uitzondering van twee studenten die respectievelijk 3 en 5 vergelijkingen hebben gemaakt.

3.4 Analyse

Het Rasch Model (zie hiervoor; Bradley & Terry, 1952; Luce, 1959) is gebruikt om de cv's te rangschikken van minst geschikt tot meest geschikt voor de functie van talent acquisition officer. Hierbij zijn twee verschillende rangordes opgesteld aan de hand van vergelijkingen: één voor de experts en één voor de student-beoordelaars. De interbeoordelaarsbetrouwbaarheid van deze rangordes wordt uitgedrukt in de Rasch Separation Reliability of Scale Separation Reliability (SSR) (Lesterhuis et al., 2017). De SSR is een maat voor interbeoordelaarsbetrouwbaarheid (Verhavert et al., 2018). Een hoge SSR in deze studie betekent dus dat er een hoge mate van zekerheid is over de positie van de individuele cv's in de totale rangorde, en dat wanneer een willekeurige andere (vergelijkbare) groep beoordelaars de cv's zou beoordelen, zij tot een vergelijkbare rangorde zou komen.

Omdat het aantal vergelijkingen verschilt tussen de studenten en de experts, is naast het maximale niveau van de interbeoordelaarsbetrouwbaarheid van elk van de rangordes ook gekeken naar het aantal vergelijkingen dat nodig is voor een minimaal niveau van interbeoordelaarsbetrouwbaarheid. In navolging van Jonsson en Svingby (2007) beschouwen we een interbeoordelaarsbetrouwbaarheidsniveau van .80 als minimum voor beslissingen waar veel vanaf hangt, zoals het selecteren van kandidaten voor de volgende ronde in het selectieproces.

De constructvaliditeit van de oordelen is op twee manieren geanalyseerd. Er is gekeken naar de overeenkomst van de uiteindelijke cv-selectie tussen studenten en experts, door de rangorde zoals verkregen op grond van de beoordelingen door de studenten te correleren met de rangorde op basis van de beoordelingen door de experts. Daarnaast is de constructvaliditeit van de oordelen vastgesteld middels een analyse van de toelichtingen voor de keuzes die de student- en expertbeoordelaars gaven na elke vergelijking. Hierbij is gekeken naar de mate waarin beoordelaars hun keuze baseerden op aspecten die relevant zijn voor de functie zoals vermeld in de vacaturetekst. De eerste auteur voerde een eerste codering uit op deze tekstdata. Aangezien de meeste argumentaties erg kort waren (bijv. 'ervaring'), kon deze codering voor de meeste argumentaties eenvoudig worden gemaakt. De eerste en derde auteur bespraken vervolgens de tekststukjes die twijfel gaven en hoe de verschillende categorieën konden worden gegroepeerd (Braun & Clarke, 2006). Op basis van deze discussie vervolledigde de eerste auteur de codering. Daarna werd de codering besproken met de vierde auteur, een expert die ook betrokken was bij het beoordelen in D-PAC.

4 Resultaten

4.1 Interbeoordelaarsbetrouwbaarheid

De groep experts maakte per cv tussen de 21 en 23 beoordelingen. De rangorde van cv's op basis van deze beoordelingen had een hoge interbeoordelaarsbetrouwbaarheid (SSR = .88). Dit betekent dat wanneer een andere groep van experts hetzelfde aantal vergelijkingen had gemaakt, zij met een hoge mate van zekerheid tot dezelfde rangorde van kandidaten zou komen.

Om een afdoend niveau van interbeoordelaarsbetrouwbaarheid te bereiken volstaan minder dan 21 beoordelingen. Wanneer we kijken hoe de interbeoordelaarsbetrouwbaarheid evolueert naarmate de experts meer beoordelingen per cv afwerkten, zien we namelijk dat 13 vergelijkingen per cv voldoende zijn om het betrouwbaarheidsniveau van .80 te bereiken.

De uiteindelijke rangorde van de studenten, op basis van 30 beoordelingen per cv, bleek ook betrouwbaar te zijn (SSR = .81). Wanneer de evolutie in de betrouwbaarheid naar het aantal beoordelingen per cv wordt bestudeerd, blijkt dat 29 beoordelingen nodig zijn om een interbeoordelaarsbetrouwbaarheid van .80 te bereiken.

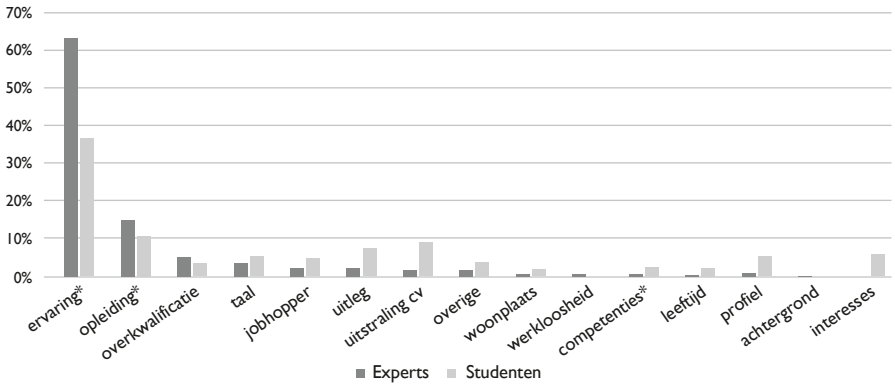
4.2 Constructvaliditeit

De oordelen van experts en studenten komen voor een groot deel overeen, met een correlatie van .77 ($p < .05$, tweezijdig). Dit laat zien dat experts en studenten de cv's op vergelijkbare wijze inschatten: cv's die volgens de experts beter passen bij de functieomschrijving, werden ook door de studenten hoger op de rangorde geplaatst. Met een correlatie van .77 is de verklaarde variantie (R^2) in de oordelen 59%. Dit impliceert dat de oordelen nog voor 41% verschillend zijn. Het is daarom van belang om verder te kijken naar de constructvaliditeit van de oordelen: zijn

er verschillen in de aspecten waarop experts en studenten hebben gelet tijdens het beoordelen van cv's?

Experts. De expert-beoordelaars gaven in totaal 836 argumenten voor hun keuzes, verdeeld over 14 categorieën. Deze 14 aangehaalde categorieën bevatten twee grote categorieën: aspecten die te maken hebben met werk-gerelateerde ervaring (64%) en aspecten die te maken hebben met opleiding (15%). De categorie werk-gerelateerde ervaring kan verder worden opgesplitst in vijf kleinere thema's: relevante ervaring (37% van het totale aantal argumenten), hoeveelheid ervaring (17%), algemene ervaring (6%), vereiste ervaring voor de baan (2%) en uitleg die kandidaten geven over hun ervaring (1%).

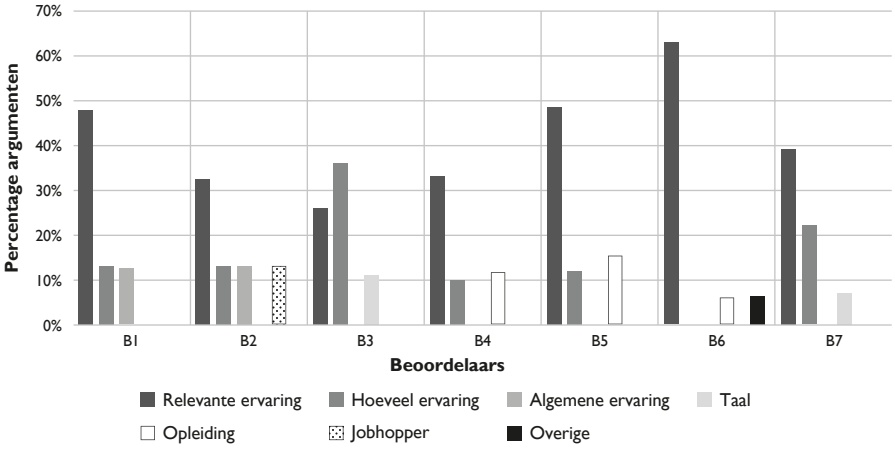
Daarnaast haalden de experts nog 13 kleinere categorieën aan: overkwalificatie (5%), taal (4%), jobhoppen (2%), uitleg over cv (2%), uitstraling van het cv (2%), overige (2%), woonplaats (1%), werkloosheid (1%), competenties (1%), globaal profiel van de kandidaat (1%), leeftijd en achtergrond. Van deze categorieën waren enkel 'taal' en 'competenties' relevant met betrekking tot de vacaturetekst. Met andere woorden, 16% van alle gegeven argumenten voor het selecteren van een cv als meest geschikt kwam niet overeen met de functieomschrijving in de vacature. De categorieën 'overkwalificatie' en 'jobhoppen' werden daarbij beschouwd als negatieve aspecten van een cv. Met andere woorden, als overkwalificatie en/of jobhoppen uit een cv werden afgeleid, verminderde de kans dat dit cv werd gekozen als meer geschikt in een paarsgewijze vergelijking. Figuur 1 geeft de percentages van de gegeven argumenten weer van zowel de expert-beoordelaars als de student-beoordelaars.



Figuur 1 ***Percentage gegeven argumenten per categorie tijdens het vergelijken van cv's door expert- en student- beoordelaars. Categorieën met een asterisk zijn de argumenten die aansluiten bij de vacature.***

Aangezien er tussen de experts onderling nog wel verschillen waren in de aspecten waarop cv's werden beoordeeld, toont Figuur 2 de meest genoemde categorieën per beoordelaar. Hieruit blijkt dat elke expert heeft gelet op relevante ervaring

van de sollicitanten, een criterium dat ook expliciet was omschreven in de functieomschrijving. Er was maar één expert die bij het selecteren van cv's ook in grote mate heeft gelet op een criterium dat niet in de functieomschrijving was vermeld, namelijk jobhopping (zie Figuur 2, beoordelaar 2).



Figuur 2 ***Percentage gegeven argumenten per categorie tijdens het vergelijken van cv's per expert-beoordelaar (N = 7)***

Student-beoordelaars. Student-beoordelaars gaven in totaal 415 argumenten voor hun oordelen. Dit was de helft van het aantal argumenten dat de experts gaven, ook al waren er ruim meer studenten dan experts. Net als de experts keken de student-beoordelaars voornamelijk naar de werk-gerelateerde ervaring (37% van de argumenten) en opleiding van de kandidaten (11% van de argumenten). Gerelateerd aan de vacature keken student-beoordelaars ook nog naar taal (5%) en competenties (3%). Wel gaven de student-beoordelaars relatief meer aandacht aan categorieën die niet in de vacaturetekst voorkwamen, dan de experts (44% van alle argumenten). Zo keken zij meer naar de uitstraling van het cv (9% van de totale argumenten), uitleg die kandidaten gaven over hun ervaring (7%), persoonlijke interesses (6%), jobhoppen (5%), globaal profiel van de kandidaat (5%) overkwalificatie (3%), woonplaats (2%), leeftijd (2%) en overige (4%) (zie Figuur 1).

Om na te gaan of de gevonden verschillen tussen de studenten en de expertgroep significant zijn, werd de z-score berekend voor de proportie van relevante argumenten in beide groepen. De resultaten geven weer dat student-beoordelaars meer aandacht gaven aan niet-relevante categorieën dan expert-beoordelaars ($z = 10.26$; $p < .001$).

5 Discussie

In de huidige studie is gebruikgemaakt van een alternatieve manier om cv's te screenen. Voorgaand onderzoek in de onderwijscontext gaf reeds evidentie voor de interbeoordelaarsbetrouwbaarheid (bijv. Verhavert et al., 2018) en construct-validiteit (bijv. Van Daal et al., 2017) van comparatieve beoordelingen. Deze studie bouwt hierop verder door voor het eerst te onderzoeken of dit ook zo is bij cv-screening door zowel expert- als student-beoordelaars.

De resultaten laten allereerst een hoge interbeoordelaarsbetrouwbaarheid zien voor de comparatieve oordelen van cv's door expert-beoordelaars (SSR = .88; 21 tot 23 beoordelingen per cv). De oordelen van de student-beoordelaars waren minder betrouwbaar in vergelijking met die van de expert-beoordelaars, hoewel zij meer beoordelingen per cv maakten (SSR = .81; 30 beoordelingen per cv). Ook hadden de studenten meer vergelijkingen nodig om een .80 betrouwbaarheidsniveau te bereiken op basis waarvan robuuste beslissingen gemaakt kunnen worden (respectievelijk 13 beoordelingen voor de experts en 29 voor de studenten).

Daarnaast laten de resultaten zien dat de oordelen van experts en studenten voor bijna 60% met elkaar overeenkomen. De verschillen tussen experts en studenten kunnen worden verklaard doordat studenten zich meer door irrelevante aspecten laten beïnvloeden dan experts. Expert-beoordelaars bleken voornamelijk op relevante aspecten te letten die overeenkomen met de functie-eisen in de vacature. Zo hebben ze de cv's voornamelijk vergeleken op relevante werkervaring en opleiding. Wel bleek dat er verschillen waren tussen beoordelaars: sommige beoordelaars waren meer gefocust op de hoeveelheid ervaring die de kandidaat had, terwijl anderen meer keken naar het taalcriterium. Studenten bleken net als experts werkervaring en opleiding de belangrijkste aspecten te vinden voor het bepalen van de geschiktheid van een cv. Daarnaast lieten studenten zich echter ook meer beïnvloeden door irrelevante zaken, zoals de lay-out van de cv's.

Op grond van deze studie kan worden geconcludeerd dat de comparatieve beoordelingsmethode een betrouwbare en valide cv-screening biedt. Wel lijkt het beter om deze comparatieve beoordelingsmethode eerder te laten gebruiken door experts dan door studenten of andere onervaren beoordelaars. Dit sluit aan bij eerdere onderzoeksbevindingen waaruit blijkt dat studenten tot minder betrouwbare en minder valide oordelen komen (zie bijv. Cole et al., 2009; De Meijer, Born, Van Zielst & Van der Molen, 2007).

5.1 Theoretische en praktische implicaties

In de huidige studie vonden we een relatief hoge interbeoordelaarsbetrouwbaarheid voor de cv-screening. Hierdoor kunnen we dus stellen dat er tussen onze beoordelaars consensus was over de rangorde van de cv's. Echter, er kunnen situaties zijn waarin er geen sprake is van consensus (Lesterhuis et al., 2017). Dit is dan een signaal dat de beoordelaars het onderling niet eens zijn over de volgorde van de geschiktheid van de betreffende cv's. Een lage(re) interbeoordelaarsbetrouwbaar-

heid kan het gevolg zijn van één of meer beoordelaar(s) die systematisch anders beoordelen (Lesterhuis et al., 2017). Sommige beoordelaars kunnen bijvoorbeeld een groter belang hechten aan aspecten die door andere beoordelaars juist als minder belangrijk worden gezien. Dit zal binnen deze beoordelingsmethode tot uiting komen in afwijkende beslissingen gedurende de comparatieve beoordelingen. Als dit patroon systematisch afwijkt van de consensus van andere beoordelaars, kan/kunnen de afwijkende beoordelaar(s) geïdentificeerd worden en kan een overleg gepland worden om toch tot een consensus te komen, waarbij de gegeven toelichting bij de oordelen meegenomen kan worden als startpunt voor het overleg. Het kan ook zijn dat verschillen van mening tussen beoordelaars uitsluitend voorkomen bij bepaalde cv's. Ook deze specifieke patronen kunnen gedetecteerd en nader bekeken worden in de resultaten van een comparatieve beoordelingsmethode.

De huidige studie toonde ook aan dat het merendeel van de beoordelaars let op aspecten die gerelateerd zijn aan de vacature. Echter, de beoordelaars verschilden ook van elkaar in de aspecten waarop ze de cv's beoordeelden. Dit kan worden veroorzaakt door verschillen in informatieverwerking tussen beoordelaars (Kinicki, Lockwood, Hom & Griffeth, 1990). Door bijvoorbeeld de comparatieve beoordelingsmethode met een team van beoordelaars uit te voeren, stijgt de kans dat meerdere relevante aspecten meegenomen worden in de beoordeling, wat de constructvaliditeit van de cv-screening sterk ten goede komt (Messick, 1989; Van Daal et al., 2017). Verder onderzoek is nodig om na te gaan hoeveel beoordelaars precies nodig zijn om zo'n valide oordeel per cv te garanderen.

De huidige studie onderzocht ook de rol van expertise van de beoordelaars. De resultaten tonen aan dat zowel de experts als de student-beoordelaars naar relevante zaken kijken bij de cv-screening. Ondanks het feit dat de instructie voor beide groepen exact hetzelfde was, keken student-beoordelaars meer naar irrelevante zaken wanneer ze cv's comparatief beoordeelden. Dit is in lijn met voorgaand onderzoek naar de kwaliteit van cv-screening bij studenten (Hanak, Sirota & Juanchich, 2013) en kan toe te schrijven zijn aan het feit dat student-beoordelaars nog geen goed onderscheid kunnen maken tussen relevante en irrelevante aspecten voor een bepaalde functie en daarom proberen alle beschikbare informatie mee te nemen bij het beoordelen (Shanteau, 1992). Dit kan ook het verschil in interbeoordelaarsbetrouwbaarheid verklaren tussen experts en student-beoordelaars. Een praktijkimplicatie hiervan is dat een beoordelaar ook in de comparatieve beoordelingsmethode enige expertise dient te hebben vooraleer deze cv's gaat screenen.

5.2 Beperkingen en verder onderzoek

Een eerste beperking voor het interpreteren van de resultaten in deze studie is dat experts en studenten een verschillend aantal vergelijkingen hebben gemaakt. Hierdoor is het niet mogelijk om verschillen in verkregen interbeoordelaarsbetrouwbaarheid eenvoudig te vergelijken tussen de twee groepen en eventuele verschillen te wijten aan expertise. De resultaten geven aan dat, ondanks dat er in de groep van student-beoordelaars in totaal meer vergelijkingen zijn gemaakt dan in de groep van experts, de totale interbeoordelaarsbetrouwbaarheid van de stu-

denten lager is. Dat is in strijd met de resultaten van een recente meta-analyse die aangaf dat de mate van expertise (novice-expert) niet samenhangt met de verkregen interbeoordelaarsbetrouwbaarheid (Verhavert et al., 2018). Uit dezelfde meta-analyse blijkt echter dat er ook een klein, maar significant effect is van het aantal vergelijkingen per beoordelaar. Nu was het in de huidige studie zo dat de studenten per persoon veel minder vergelijkingen hebben gemaakt dan de experts. Het zou kunnen dat dit de lagere interbeoordelaarsbetrouwbaarheid verklaart. Tegelijkertijd lijken de resultaten van de studenten te laten zien dat in deze studie de maximale interbeoordelaarsbetrouwbaarheid al bereikt was, dus dat er ook met meer vergelijkingen waarschijnlijk geen hoge interbeoordelaarsbetrouwbaarheid bereikt zou worden. Dit zou erop wijzen dat studenten toch minder betrouwbaar beoordelen dan experts (Verhavert, 2018). Verder onderzoek met een gelijk aantal beoordelingen per persoon overheen de twee groepen is daarom nodig om na te gaan of studenten (of breder: novices) minder betrouwbaar beoordelen dan experts. Op grond daarvan kan worden uitgemaakt of het verschil in interbeoordelaarsbetrouwbaarheid veroorzaakt wordt door het aantal vergelijkingen of door het verschil in expertise. Hierbij aansluitend: de huidige studie heeft enkel gebruikgemaakt van externe experts (namelijk van een extern HR-adviesverlener). Toekomstig onderzoek zou de rol van expertise ook kunnen nagaan door ook interne experts (namelijk experts gebonden aan de werkgever) erbij te betrekken. Op die manier kan het effect van de leer- en socialisatie-ervaringen in kaart worden gebracht en kan worden nagegaan of het zinvol is om externe experts te betrekken bij het beoordelen van cv's.

Een tweede beperking van de huidige studie is dat we de constructvaliditeit hebben bestudeerd in termen van de al dan niet relevante argumenten die beoordelaars gaven. Aan de beoordelaars werd niet expliciet gevraagd waarom ze juist dit specifieke argument (of argumenten) gaven bij elke vergelijking. We kunnen dus niet precies hun beslissingsproces in kaart brengen. Om dit verder te bestuderen zou een vervolgstudie een 'think-aloud'-procedure kunnen opzetten waarbij aan de beoordelaars wordt gevraagd om hun gedachtegang te vocaliseren, waardoor deze nauwkeurig kan worden bestudeerd. Hierdoor kan worden nagegaan waarom men op dat moment die beslissing maakt op basis van de gegeven argumentatie.

Een derde beperking van de huidige studie is dat we geen directe vergelijking kunnen maken tussen de comparatieve methode en de klassieke methode. Hierdoor kunnen we op basis van onze resultaten niet stellen dat de hoge mate van interbeoordelaarsbetrouwbaarheid en constructvaliditeit te danken zijn aan (a) de comparatieve methode op zich, of (b) omdat expert-beoordelaars cv's op redelijk soortgelijke wijze beoordelen of screenen, of (c) omdat de vorm en/of de inhoud van de cv's onderling sterk van elkaar verschilden en cv's snel te onderscheiden waren (waardoor sterke cv's gemakkelijk te identificeren zijn door experts). Verder onderzoek is nodig om te weten te komen of de methode van comparatief beoordelen ook tot een meer betrouwbare en valide selectie leidt in vergelijking met bijvoorbeeld een klassieke niet-comparatieve manier van beoordelen.

Aandachtspunten voor dergelijk onderzoek zijn enerzijds het maken van een random verdeling van beoordelaars over de condities en anderzijds het inbouwen van een groepsfase voor de klassieke aanpak. Elke beoordelaar in de conditie van de klassieke aanpak kan bijvoorbeeld eerst individueel de cv's beoordelen om daarna in overleg met andere beoordelaars tot een gedeelde rangschikking te komen. Dit zorgt ervoor dat de opzet van deze klassieke-aanpak-conditie beter vergelijkbaar is met de opzet in de conditie waarin comparatieve oordelen worden gemaakt. Hierbij aansluitend kan de huidige studie geen uitspraken doen over de balans tussen de tijdsinvestering van beide methoden en de interbeoordelaarsbetrouwbaarheid en constructvaliditeit. Bij de klassieke methode wordt de cv-screening in het algemeen uitgevoerd door één beoordelaar. In deze studie voerden zeven expert-beoordelaars de screening uit, wat misschien in de praktijk niet altijd praktisch haalbaar of te kostbaar is. De huidige studie vond dat de interbeoordelaarsbetrouwbaarheid bij de zeven expertbeoordelaars al na 13 vergelijkingen de .80 bereikte. Tot op heden is er echter nog geen onderzoek gepubliceerd met richtlijnen over het minimale en optimale aantal vereiste beoordelaars en te maken beoordelingen om voldoende psychometrische kwaliteit van uitgevoerde of nog uit te voeren comparatieve beoordelingen te kunnen garanderen (Verhavert, 2018). Verder onderzoek hiernaar is dan ook meer dan wenselijk.

De tijdsinvestering van de comparatieve methode kan bijvoorbeeld worden gereduceerd door een aanpassing te maken in de algoritmes die gebruikt worden in de D-PAC-tool om de paren van cv's te vormen en om het aantal aan te bieden paren van cv's te bepalen (Lesterhuis et al., 2017; Verhavert et al., 2018). Zo kan de aanpassing inhouden dat beoordelaars direct kunnen opgeven welk cv zeker niet in aanmerking komt voor de vacature. Indien een cv door bijvoorbeeld twee beoordelaars als absoluut ongeschikt wordt aangevinkt, zou dit cv verder niet meer kunnen worden aangeboden, wat het aantal te maken vergelijkingen zou verminderen. Verder onderzoek is nodig om te bepalen hoeveel oordelen er nodig zijn om een cv uit de vergelijkingen te weren (twee of meer?), hoeveel tijdswinst dit mogelijk zou opleveren en wat de invloed hiervan is op de interbeoordelaarsbetrouwbaarheid en constructvaliditeit.

De betrokken expert-beoordelaars hadden nog een bedenking over de inzet van de comparatieve methode in cv-screening bij een laag aantal ingezonden cv's. In deze situatie kan het minder interessant zijn om de comparatieve methode te gebruiken. Echter, bij terugkerende gelijke vacatures (bijv. voor generieke functies zoals 'adviseur') kan gebruik worden gemaakt van reeds bestaande rangordes op basis van de vorige cv-screening over dezelfde vacature. De nieuwe cv's kunnen dan ook vergeleken worden met reeds beoordeelde en geanonimiseerde cv's uit een bestaande rangorde of cv's die op basis van deze voorbeelden door de jury zelf zijn samengesteld. Zo'n cv kan een bepaalde grens aanduiden wanneer nieuwe cv's moeten worden beoordeeld: alles boven deze cv mag door naar de volgende ronde. Hierdoor kunnen de nieuwe cv's sneller worden gepositioneerd in de rangorde en hoeven er minder vergelijkingen te worden gemaakt.

6 Conclusie

We kunnen concluderen dat het gebruik een comparatieve beoordelingsmethode in de context van cv-screening het mogelijk maakt om tot valide en betrouwbare oordelen over cv's van beoordelaars te komen. Er moet wel rekening worden gehouden met de expertise van de beoordelaars die deze screening uitvoeren, zowel in de praktijk als in verder empirisch onderzoek. Onze resultaten bieden empirische evidentie voor de stelling dat zowel de interbeoordelaarsbetrouwbaarheid als de constructvaliditeit van comparatief gemaakte oordelen over cv's hoger is bij expert-beoordelaars dan bij studenten die dezelfde methode gebruiken.

Praktijkbox

Wat betekenen de resultaten voor de praktijk?

- De interbeoordelaarsbetrouwbaarheid en constructvaliditeit van cv-screening mogen niet als evident worden beschouwd: tijdens de selectie zijn cognitieve vertekening en beoordelingsfouten mogelijk.
- Het comparatief beoordelen van cv's (waarbij beoordelaars de cv's steeds in paren met elkaar vergelijken en voor elk paar beslissen welke van de twee cv's beter past bij een vacature) maakt het mogelijk om valide en betrouwbare oordelen over cv's te verkrijgen (met name constructvaliditeit en interbeoordelaarsbetrouwbaarheid).
- De expertise van de beoordelaars speelt een belangrijke rol in het comparatief beoordelen van cv's.
- Expert-beoordelaars maken meer valide en betrouwbaardere comparatieve oordelen over cv's dan student-beoordelaars.
- Student-beoordelaars letten bij comparatieve oordelen over cv's meer op irrelevante informatie dan expert-beoordelaars.

Literatuur

Barber, A.E. (1998). *Recruiting employees: Individual and organizational perspectives*. London: Sage Publications.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*, 991-1013.

Blommaert, L., Van Tubergen, F., & Coenders, M. (2012). Implicit and explicit interethnic attitudes and ethnic discrimination in hiring. *Social Science Research, 41*, 61-73.

Born, M.Ph. (2008). De selecteur, de sollicitant, de samenleving en de expert: Drijfveren bij de selectie van personen. *Gedrag & Organisatie, 21*, 150-169.

Bradley, R.A., & Terry, M.E. (1952). Rank analysis of incomplete block designs, the method of paired comparisons. *Biometrika 39*, 324-345.

Bramley, T. (2007). Paired comparisons methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.). *Techniques for monitoring the comparability of examination standards* (pp. 246-294). London: Qualification and Authority.

Bramley, T. (2015). I*nvestigating the reliability of Adaptive Comparative Judgment*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77-101.
Burns, G.N., Christiansen, N.D., Morris, M.B., Periard, D.A., & Coaster, J.A. (2014). Effects of applicant personality on resume evaluations. *Journal of Business Psychology, 29*, 573-591.
Cavojova, V., & Hanak, R. (2014). How much information do you need? Interaction of intuitive processing with expertise. *Studia Psychologica, 56*, 83-97.
Chase, W.G., & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.
Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., & De Maeyer, S. (2017). Judging texts with rubrics and comparative judgement: Taking into account reliability and time investment. [Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: Een afweging van betrouwbaarheid en tijdsinvestering]. *Pedagogische Studiën, 94*, 283-303.
Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. New York: Routledge.
Cole, M.S., Feild, H.S., Giles, W.F., & Harris, S.G. (2009). Recruiters' inferences of applicant personality based on resume screening: Do paper people have a personality? *Journal of Business Psychology, 24*, 5-18.
Cotton, J.L., O'Neill, B.S., & Griffin, A. (2008). The 'name game': Affective and hiring reactions to first names. *Journal of Managerial Psychology, 23*,18-39.
De Meijer, L.A.L., Born, M.Ph., Van Zielst, J., & Van der Molen, H.T. (2007). Analyzing judgments of ethnically diverse applicants during personnel selection: A study at the Dutch police. *International Journal of Selection and Assessment, 15*, 139-152.
Derous, E. (2007). Naamdiscriminatie bij CV-screening. *Tijdschrift voor Arbeidsvraagstukken, 23*, 366-380.
Derous, E. (2011). Geen baan voor een Marokkaan? Discriminatie bij CV-screening nader bekeken. *Gedrag & Organisatie, 24*, 139-164.
Derous, E., Nguyen, H.H., & Ryan, A.M. (2009). Hiring discrimination against Arab minorities: Interactions between prejudice and job characteristics. *Human Performance, 22*, 297-320 .
Derous, E., Ryan, A.-M., & Serlie, A.E. (2014). Double jeopardy upon resumé screening: When Achmed is less employable than Aïsha. *Personnel Psychology, 68*, 659-696.
Fritzsche, B.A., & Brannick, M.T. (2002). The importance of representative design in judgment tasks: The case of résumé screening. *Journal of Occupational and Organizational Psychology, 75*, 163-169.
Hanak, R., Sirota, M., & Juanchich, M. (2013). Experts use compensatory strategies more often than novices in hiring decisions. *Studia Psychologica, 55*, 251-263.
Highhouse, S., & Hause, E.L. (1995). Missing information in selection: An application of the Einhorn-Hogarth ambiguity model. *Journal of Applied Psychology, 80*, 86-83.
Hurtz, G.M., & Donovan, J.J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869-879.
Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130-144.
Kaas, L., & Manger, C. (2012). Ethnic discrimination in Germany's labor market: A field experiment. *German Economic Review, 13*, 1-20.
Kinicki, A.J., Lockwood, C.A., Hom, P.W., & Griffeth, R.W. (1990). Interview prediction of applicant qualifications and interviewer validity: Aggregate and individual analyses. *Journal of Applied Psychology, 75*, 477-486.

Knouse, S.B. (1994). Impressions of the résumé: The effects of applicant education, experience, and impression management. *Journal of Business and Psychology, 9*, 33-45.
Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgements. *The Quarterly Journal of Experimental Psychology, 42,* 239-254.
Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson Learning.
Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality: An assessor's perspective*. Proefschrift, Universiteit Antwerpen.
Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative practices for higher education assessment and measurement* (pp. 119-139). Hershey, PA: IGI Global.
Luce, R.D. (1959). On the possible psychophysical laws. *Psychological Review, 66*, 81-95.
Maddox, K.B. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review, 8*, 383-401.
Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5-11.
Piotrowski, C., & Armstrong, T. (2006). Current recruitment and selection practices: A national survey of fortune 1000 firms. *North American Journal of Psychology, 8*, 489-496.
Pollitt, A. (2012). The method of adaptive comparative judgment. Assessment in education: Principles. *Policy & Practice, 19*, 281-300.
Riach, P.A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal, 112*, 480-518.
Roe, R.A. (1983). *Grondslagen der personeelsselectie*. Assen: Van Gorcum.
Sackett, P.R., Lievens, F., Van Iddekinge, C.H., & Kuncel, N.R. (2017). Individual differences and their measurement: A review of 100 years of research. *Journal of Applied Psychology, 102*, 254-273.
Seibert, M., Williams, K., & Raymark, P. (2010). *Résumé screening: A policy capturing study of recruiter judgments*. Paper presented at 25th Annual Conference of the Society for Industrial & Organizational Psychology, Atlanta, GA.
Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica, 68*, 203-215.
Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica, 81*, 75-86.
Steiner, D.D. (2012). Personnel selection across the globe. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 740-767). Oxford Library of Psychology.
Sutherland, M., & Wöcke, A. (2011). The symptoms of and consequences to selection errors in recruitment decisions. *South African Journal of Business Management, 42*(2), 23-32.
Thurstone, L.L. (1927). The law of comparative judgment. *Psychological Review, 34*, 273-286.
Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice, 26*, 59-74.
Verhavert, S. (2018). *Beyond a mere rank order: The method, the reliability and the efficiency of comparative judgement*. Proefschrift, Universiteit Antwerpen.
Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy, & Practice, published online 12 April 2019*.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Comparative judgement and scale separation reliability: What does it mean? *Applied Psychological Measurement, 42*, 428-445.

Voncken, V., & Westendorp, M. (2007). *Anoniem solliciteren: zinvol en wenselijk? Rapportage van onderzoek onder werkgevers en consumenten*. Amsterdam: TNS NIPO.

Waung, M., Hymes, R., Beatty, J.E., & McAuslan, P. (2015). Self-promotion statements in video resumes: Frequency, intensity, and gender effects on job applicant evaluation. *International Journal of Selection and Assessment, 23*, 345-360.

Webster, E.C. (1964). *Decision making in the employment interview*. Montreal: Industrial Relations Center, McGill University.

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method*. Manchester: AQA Centre for Education Research and Policy.

Whitehouse, C., & Pollitt, A. (2012). *Using adaptive judgment to obtain a highly reliable rank order in summative assessment*. Manchester: AQA Centre for Education Research and Policy.

Zegers de Beijl, R. (Ed.). (2000). *Documenting discrimination against migrant workers in the labor market: A comparative study of four European countries*. Geneva, Switzerland: International Labor Office.

Ziegert, J.C., & Hanges, P.J. (2005). Employment discrimination: The role of implicit attitudes, motivation and a climate for racial bias. *Journal of Applied Psychology, 90*, 553-562.

Bijlage: TALENT ACQUISITION OFFICER

Creatieve professional met expertise in rekrutering en selectie

JOUW FUNCTIE

Binnen het HR team ben jij verantwoordelijk voor de kwalitatieve rekrutering en selectie van bedienden- en managementprofielen binnen <naam van de werkgever>.

- Je beheert de volledige rekruteringscyclus: van de bepaling van de skills met de lijnmanagers, cv screening, interviews, advisering naar het management om zo de bedrijfsdoelstellingen te realiseren.
- Om de perfecte match te vinden en een netwerk op te bouwen van topkandidaten in het kader van een proactieve aanwerving gebruik je alle mogelijke kanalen: rekruteringsdatabanken, job advertenties, externe partijen, referrals, social media,…
- Je staat in voor het arbeidsmarktcommunicatieplan en het positioneren van het bedrijf in de arbeidsmarkt en hebt daarbij oog voor diversiteit en inclusie.
- Als ambassadeur van het bedrijf sta je in nauw contact met universiteiten en hogescholen, neem je deel aan jobbeurzen en organiseer je wervingsevenementen.
- Je werkt in overleg met het lijnmanagement aan proactief talent management en successieplanning die tevens de interne mobiliteit van talenten bevordert.
- Je blijft continu op de hoogte van de evoluties en ontwikkelingen op het gebied van HR, zodat je je geloofwaardigheid kan verstevigen en een service kan verlenen die gebaseerd is op geactualiseerde kennis.

JOUW PROFIEL

Je genoot een masteropleiding en hebt een vijftal jaar ervaring in human resources. Je wenst je te specialiseren in het aantrekken van talent en talentontwikkeling. Je bent vertrouwd met diverse wervings- en selectietechnieken, testings, assessments en kan recruitment op een authentieke en professionele manier neer zetten, waardoor je intern als klankbord en adviseur ageert.

- Je hebt oog voor talent, je munt uit in het inschatten van mensen en je vraagt op een natuurlijke manier door.
- Je bent een dynamische persoon met een stevige dosis overtuigingskracht die initiatief neemt en doorzet wanneer nodig.
- Je bent flexibel om je naar verschillende vestigingen van <naam van de werkgever> in België te verplaatsen.
- Je drukt je vlot uit in het Nederlands en Frans, je hebt een goede kennis van de courante office pakketten en je bent vertrouwd met sociale media.

ONS AANBOD

<naam van de werkgever> biedt je een interessante functie in een leerrijke omgeving met ruimte voor creativiteit en eigen initiatief. Er heerst een ongedwongen werksfeer waar open communicatie en professionalisme voorop staan. Je ont-

vangt een aantrekkelijk salarispakket aangevuld met een aantal extralegale voordelen waaronder een firmawagen. <naam van de werkgever> gelooft niet enkel dat mensen het verschil maken, maar heeft dit geloof ook diep verankerd in zijn gehele bedrijfsvoering.

Comparative Judgment as a reliable and valid assessment method in resume screening: a comparison between experts and novices

A. Mortier, R. Bouwer, L. Coertjens, E. Volckaert, A. Vrijdag, R. van Gasse, P. Vlerick & S. De Maeyer, Gedrag & Organisatie, volume 32, June 2019, nr. 2, pp. 86-107.

In practice and in earlier empirical research, it is indicated that resume screening does not always provide suitable candidates for a vacancy. This might be due to several issues: one assessor carries out the screening, resulting in cognitive distortions influencing the selection process; the assessment does not focus on all aspects of the selection, nor does it allow for certain criteria to weigh more heavily than others; and/or the assessor is insufficiently trained to carry out the resume screening. The current study captures these issues by offering an alternative assessment method (Comparative Judgment) in which the inter-rater reliability and construct validity of the resume screening is studied: several assessors with different levels of expertise assess resumes comparatively. In this study, resumes from 42 candidates applying for an existing vacancy were used. These resumes were directly compared by experienced (N = 7; experts), and less experienced assessors (N = 57; students). Results show that the comparative judgements of experienced assessors are linked to valid and reliable resume screening. The inter-rater reliability of the student assessments was lower than that of the experts. Even though the final rank ordering of the resumes correlated, students often relied on irrelevant aspects in the resumes.

Key words: Comparative Judgment, construct validity, expertise, interrater reliability, resume screening