RAPPORT ONDERZOEK

EVALUATIE VAN HET SELECTIEPROCES VOOR DE RIO-OPLEIDING

Dr. Marvin Neumann
Dr. A. Susan M. Niessen
Prof. dr. Rob R. Meijer

VU VRIJE UNIVERSITEIT AMSTERDAM

rijksuniversiteit groningen

Inhoudsopgave

1 Samenvatting

Het doel van dit rapport was om de huidige selectieprocedure voor rechters in opleiding (rio's) te evalueren. Uit deze evaluatie bleek dat de Rechtspraak zeer gemotiveerd is om goede en diverse kandidaten te selecteren. Er worden bijvoorbeeld veel verschillende selectie-instrumenten gebruikt op basis waarvan na een beraadslaging een selectiebeslissing genomen wordt. Om te evalueren in hoeverre dit leidt tot goede selectiebeslissingen bespreken wij in Hoofdstuk 2.4 de belangrijkste bevindingen uit de wetenschappelijke literatuur. De belangrijkste aandachtspunten op basis daarvan zijn dat 1) het gebruik van meer selectie-instrumenten niet altijd leidt tot betere selectiebeslissingen (d.w.z. tot het aannemen van beter presterende kandidaten), maar ook nadelig kan zijn, 2) *gestandaardiseerde* gesprekken een hogere voorspellende waarde hebben dan niet gestandaardiseerde gesprekken en ook in kleinere verschillen in beoordelingen tussen kandidaten met en zonder een etnisch-culturele achtergrond resulteren, en 3) betere selectiebeslissingen worden genomen als informatie volgens een vooraf opgestelde beslisregel wordt gecombineerd, in plaats van op basis van menselijke oordelen gevormd door een enkele beoordelaar of in een beraadslaging. In Hoofdstuk 3 bespreken wij de effectiviteit van de gebruikte selectie-instrumenten op basis van wetenschappelijke bevindingen en data uit het rio-volgsysteem. Daaruit volgen aanbevelingen over welke selectie-instrumenten wel of niet gebruikt moeten worden, en waarom. Uit Hoofdstuk 3 blijkt dat verschillende selectie-instrumenten, zoals de meerderheid van de zelfrapportagevragenlijsten, de Watson-Glaser test, de Matrix test, de Dilemmatest, en de simulatieopdracht ('Everest Games') het beste niet meer gebruikt kunnen worden. De scores op deze instrumenten hebben geen of een verwaarloosbaar klein verband met prestatie in de opleiding, of hebben wel een bruikbaar verband met prestatie in de opleiding, maar zijn overbodig omdat ze sterk overlappen met andere selectie-instrumenten. De analytische tests voorspelden zoals verwacht prestatie in de opleiding redelijk goed. Het assessment heeft in de huidige vorm geen toegevoegde waarde boven op de analytische tests. Uit Hoofdstuk 3 blijkt ook dat de standaardisatie in de selectiegesprekken ontbreekt en het soms niet voor iedereen duidelijk is wat in de gesprekken wordt gemeten. De selectiegesprekken moeten dus nog (verder) gestandaardiseerd worden. Ook moeten de lokale gesprekken en de eindgesprekken beter op elkaar afgestemd worden. Op dit moment worden in de lokale gesprekken soms ook competenties beoordeeld die pas in de eindgesprekken beoordeeld horen te worden, terwijl het doel van de lokale selectie is om de 'fit' met het gerecht te beoordelen. Het blijft vaak onduidelijk wat onder 'fit' met het gerecht wordt verstaan. Daarom moet per gerecht 'fit' concreet omschreven worden. Als alternatief kan overwogen worden om de lokale en landelijke gesprekken samen te voegen door bijvoorbeeld in de lokale selectie ook competenties te beoordelen. De LSR zou dan echter volledige standaardisatie in de lokale gesprekken moeten waarborgen. Een andere optie zou zijn de fit met het gerecht in de landelijke gesprekken te laten beoordelen.

Er zijn echter ook positieve punten. De eindgesprekken voorspellen prestaties in de opleiding al enigszins en kunnen door middel van relatief eenvoudige ingrepen aantoonbaar beter functioneren. *Onder de aanname dat selecteurs hun oordelen na de eindgesprekken zouden middelen*, zou het gemiddelde eindoordeel een redelijk hoge voorspellende waarde voor prestatie in de opleiding tonen. Wij benadrukken echter dat het eindoordeel in de huidige selectieprocedure nog niet op deze manier tot stand komt. Wij raden daarom aan de beraadslaging te laten vervallen en in plaats daarvan de oordelen van de selecteurs te middelen en een beslisregel te gebruiken om te bepalen wanneer een kandidaat wordt aangenomen. In de eindgesprekken die door de LSR gevoerd worden wordt al ten dele een beslisregel gebruikt om tot een oordeel per selecteur te komen, dus daar

bestaat al een basis voor die verder uitgewerkt kan worden. In Hoofdstuk 3.5 leggen wij uit hoe dit geoptimaliseerd kan worden. In Hoofdstuk 4 bespreken wij hoe de selectieprocedure als geheel door gebruik van beslisregels het beste ingericht kan worden om goede en diverse kandidaten te selecteren. In Hoofdstuk 5 geven wij een overzicht van onze aanbevelingen. Samengevat zijn de belangrijkste aanbevelingen dus om *minder* informatie te gebruiken, gesprekken (verder) te standaardiseren, beslisregels toe te passen en vooral voor consistent gebruik te zorgen. Wij concluderen in Hoofdstuk 6 dat deze aanbevelingen voor selecteurs relatief eenvoudig te implementeren zijn en de selectiebeslissingen aanzienlijk zullen verbeteren. Ten slotte willen wij benadrukken dat het gebruik van gestandaardiseerde gesprekken en beslisregels de selectieprocedure eerlijker en transparanter zou maken, wat ook toekomstige evaluaties vergemakkelijkt.

2 Inleiding

2.1 Doelstelling

De laatste evaluatie van de selectieprocedure voor rio's resulteerde in het advies om de selectie-procedure te herzien en daarna opnieuw te laten evalueren (Niessen & Meijer, 2019). De doelstelling van dit rapport is daarom om de huidige selectieprocedure voor rio's te evalueren. Daarbij dient beoordeelt te worden in hoeverre de selectieprocedure geschikte kandidaten oplevert die in staat zijn om de rio-opleiding succesvol af te ronden. Andere doelen zijn de selectieprocedure zo in te richten dat zij door potentiële kandidaten als aantrekkelijk wordt gezien en dat er geen geschikte kandidaten onbedoeld afvallen op basis van onder andere hun geslacht, leeftijd, of afkomst. In eerste instantie dient dit onderzoek te voorzien in een onafhankelijk en objectief oordeel over de mate waarin de huidige selectieprocedure erin slaagt de beoogde doelstellingen te realiseren. Ten tweede worden in dit rapport concrete aanbevelingen gegeven op basis van de bevindingen uit het onder-zoek en de wetenschappelijke literatuur. Soms is een bijkomstig doel van een selectieprocedure op basis van de selectie-instrumenten uitspraken te doen over sterke en minder sterke punten van kandidaten, om de ontwikkeling te ondersteunen. Wij merken op dat instrumenten altijd van voldoende kwaliteit moeten zijn, ongeacht of het primaire doel selectie of ontwikkeling is.

2.2 Context en beschrijving van de selectieprocedure

Alle gerechten stellen jaarlijks een aantal vacatures open voor de korte en de lange opleiding, waarbij de rechtbanken minimaal twee vacatures openstellen voor de lange opleiding. Op dit moment geldt de afspraak om als Rechtspraak als geheel jaarlijks 140 rio's in opleiding te nemen. Het doel is dat kandidaten na het succesvol afronden van de opleiding ook bij het gerecht waar zij in opleiding waren, blijven werken. Kandidaten solliciteren online op een vacature bij één of twee specifieke gerechten. Nadat een kandidaat het digitale sollicitatieformulier heeft ingediend volgt een selectieprocedure op basis van het rechtersprofiel. De selectieprocedure bestaat uit landelijke en lokale elementen, zoals hieronder weergegeven:

Selectiemethode	Beschrijving
Briefselectie	Toetsing van formele vereisten (onder andere Nederlandse nationaliteit en vooropleiding) en beslissing voortzetting procedure op basis van de brief en CV. Kandidaten die aan de formele vereisten voldoen worden beoordeeld op hun opleiding, werkervaring, bijzondere kennis of vaardigheden, motivatie voor het rechterschap en de opleiding tot rechter, maatschappelijke oriëntatie, en taalgebruik en schrijfstijl.
Analytische tests	Tests afgenomen door adviesbureau LTP, met als doel om vast te stellen of er wordt voldaan aan het vereiste niveau van verbaal en kritisch redeneervermogen.
Lokale selectiegesprekken	Gesprekken met lokale selectiecommissies.
Referenties	Kandidaten moeten tijdens het solliciteren twee referenties aanleveren van referenten die zij zelf mogen kiezen.

Selectiemethode	Beschrijving
Assessment	Uitgevoerd door LTP, bestaand uit een gesprek met een psycholoog, een rollenspel, een selectiespel, en een aantal zelfrapportagevragenlijsten. Gebaseerd op deze instrumenten worden de competenties behorend bij het rechtersprofiel in kaart gebracht, resulterend in een assessmentrapport en een selectie- en ontwikkeladvies.
Eindgesprekken met de landelijke selecteurs	Drie eindgesprekken met steeds twee leden (een intern en extern lid) van de Landelijke Selectiecommissie Rechters (LSR).

Kandidaten doorlopen de volledige selectieprocedure, tenzij zij vallen onder 'bijzondere kandidaten' (overstappende kandidaten uit het Openbaar Ministerie die eerder raio zijn geweest, super ervaren kandidaten, super-specialisten en herintredende rechters). Voor deze groep geldt een aangepaste procedure. In dit rapport zijn zowel gewone als ook bijzondere kandidaten in de analyses meegenomen.

2.3 Werkwijze

Er zijn door de Rechtspraak en de onderzoekers specifieke onderzoeksvragen geformuleerd (zie Bijlage 1). Om de selectieprocedure aan de hand van deze onderzoeksvragen te evalueren zijn verschillende methodes ingezet. Er zijn 19 gesprekken met verschillende personen gevoerd die bij de selectie betrokken zijn om de onderzoeksvragen 1-4, 9, 10, en 17 te beantwoorden. Het doel van de gesprekken was om inzicht te krijgen in de meningen van betrokkenen over de onderdelen van de selectieprocedure, om in kaart te brengen hoe de selectie daadwerkelijk uitgevoerd wordt, en hoe men tegen eventuele veranderingen aan zou kijken. De semigestandaardiseerde gesprekken duurden ongeveer 45 minuten en zijn met toestemming van de respondenten opgenomen en getranscribeerd. De gesprekken zijn volgens thematische analyse geanalyseerd (Braun & Clarke, 2006). Daarnaast hebben wij de meningen van geselecteerde oud-kandidaten via een online vragenlijst in kaart gebracht om onderzoeksvraag 5 te beantwoorden. De vragenlijst is opgesteld op basis van de organizational justice theory (Steiner & Gilliland, 1996). Om te beantwoorden in hoeverre de selectie-instrumenten succes in de opleiding kunnen voorspellen (onderzoeksvragen 6, 11, 12, 13, 19, 20) zijn de verbanden berekend tussen scores en beoordelingen uit de selectie-instrumenten en prestatie in de rio-opleiding. Hiervoor zijn gegevens uit het rio-volgsysteem, van de opleiding, en van adviesbureau LTP ontvangen. Ten slotte zijn documenten van de LSR bestudeerd en de wetenschappelijke literatuur geraadpleegd om onderzoeksvragen 7, 8, 14, 15, 16, en 18 te beantwoorden.

2.4 Uitleg van fundamentele concepten bij selectiebeslissingen

Standaardisatie en kwaliteit van beslissingen

Een robuuste bevinding in de literatuur is dat de voorspellende waarde van een selectie-instrument toeneemt naarmate het instrument meer gestandaardiseerd is (oftewel gestructureerd, Huffcutt & Arthur, 1994; Sackett et al., 2022). Dit geldt niet alleen voor selectiegesprekken, maar ook voor alle andere selectie-instrumenten die in een selectieprocedure ingezet worden (briefselectie, analytische tests, lokale gesprekken, en het assessment). Aangezien de rio-selectieprocedure echter onder andere uit een aantal selectiegesprekken bestaat (lokale gesprekken, eindgesprekken, een gesprek met een LTP-adviseur) leggen wij standaardisatie uit aan de hand van een selectiegesprek. Er bestaan verschillende componenten van standaardisatie binnen selectiegesprekken (Campion et al., 1997; Chapman & Zweig, 2005). Ten eerste moeten de vragen gerelateerd zijn aan de baan. Ten tweede

moeten de vragen zelf gestandaardiseerd zijn, dat wil zeggen, dezelfde vragen worden in dezelfde volgorde aan iedere kandidaat gesteld. Ten derde moet de manier waarop de antwoorden van de kandidaten beoordeeld wordt gestandaardiseerd zijn; de antwoorden van kandidaten worden *per vraag*, *onafhankelijk* (zonder overleg met andere selecteurs) en *numeriek* gescoord, bijvoorbeeld op een vijfpuntsschaal. Daarbij wordt idealiter gebruik gemaakt van geankerde beoordelingsschalen waarin per vraag vooraf gedefinieerd is hoe goede en minder goede antwoorden eruitzien. Ook is het belangrijk te voorkomen dat selecteurs voorafgaand aan een gesprek aanvullende informatie, zoals brieven, cv's, test scores en rapporten, bekijken. De informatie verkregen uit het gesprek is dan namelijk niet langer onafhankelijk van de andere informatie die in de procedure wordt gebruikt, wat maakt dat het gesprek minder toegevoegde waarde heeft. Natuurlijk kan aanvullende informatie meegenomen worden in een selectiebeslissing, maar pas nadat het oordeel op basis van het gesprek is tot stand gekomen. De laatste stap is het combineren van de beoordelingen op vraagniveau tot een eindoordeel op het gesprek. In de literatuur worden twee vormen onderscheiden hoe informatie kan worden gecombineerd. Dit betreft zowel het combineren van beoordelingen binnen een selectie-instrument (bijvoorbeeld, hoe worden beoordelingen in een gesprek tot een eindoordeel gecombineerd) als ook het combineren van scores of beoordelingen die uit meerde selectie-instrumenten voortkomen (bijvoorbeeld, hoe wordt het advies uit het assessmentrapport gecombineerd met het oordeel uit de eindgesprekken).

Holistische combinatie betekent dat informatie door 'nadenken' naar een oordeel vertaald wordt (Meehl, 1954; Meijer et al., 2020). Dit kan ook in groepsverband gebeuren, waar meerdere selecteurs kandidaten bespreken en zo een oordeel vormen, zoals in de huidige beraadslagingen na de eindgesprekken. Alternatief kan informatie via een vooraf opgestelde beslisregel of formule op een consistente manier tot een oordeel gecombineerd worden. Dit wordt ook *mechanische combinatie* genoemd (Meehl, 1954; Meijer et al., 2020). Mechanische combinatie vereist dat informatie zoals indrukken uit gesprekken numeriek gescoord worden. Het numeriek scoren van informatie impliceert andersom echter niet per definitie dat er mechanische combinatie gebruikt wordt. Een selecteur die antwoorden van een kandidaat numeriek scoort, maar vervolgens een oordeel vormt door over die scores na te denken, wellicht in overleg met andere selecteurs, gebruikt holistische combinatie. Zelfs als men een 'mentale' beslisregel gebruikt, door tijdens het nadenken bewust meer of minder gewicht te geven aan sommige informatie, is dit nog steeds holistische combinatie. Pas als scores via een expliciete beslisregel of formule consistent op dezelfde manier voor iedere kandidaat worden gecombineerd is er sprake van mechanische combinatie. Een belangrijk onderscheid is dus het *verzamelen* van informatie (meer of minder gestandaardiseerd) en het *combineren* van informatie (holistisch of mechanisch).

Een robuuste bevinding is dat het consistent gebruiken van beslisregels tot betere oordelen en selectiebeslissingen leidt dan holistische combinatie (Grove & Meehl, 1996; Kuncel et al., 2013; Meehl, 1954; Yu & Kuncel, 2020, 2022). Die beslisregels kunnen simpel en transparant zijn. Bij selectiegesprekken kunnen numerieke beoordelingen van antwoorden bijvoorbeeld gemiddeld worden (Arkes et al., 2010). Als het wenselijk is kunnen beoordelingen ook eerst per competentie (bijvoorbeeld besluitvaardigheid) gemiddeld worden en als meerdere selecteurs betrokken zijn bij gesprekken kan een gemiddelde over de selecteurs heen berekend worden. Het gebruik van meerdere selecteurs bij gesprekken, zoals dat nu ook gebeurt, is wenselijk en leidt tot betere voorspellingen dan het gebruik van een enkele selecteur, mits de selecteurs de antwoorden *onafhankelijk* scoren en vervolgens hun oordelen volgens een beslisregel combineren (Armstrong, 2001). Het (gewogen)

middelen van scores betekent dat informatie compensatorisch wordt gebruikt, omdat hoge scores lage scores kunnen compenseren. Het is ook mogelijk om informatie via een beslisregel noncompensatorisch te combineren. Een voorbeeld zou zijn: Wanneer de score op twee of meer competenties matig of slechter is (<= 2 uit 5), krijgt de kandidaat een negatieve beoordeling. Anders krijgt de kandidaat een positieve beoordeling. Ook is het mogelijk beslisregels op te stellen waarin informatie noncompensatorisch en compensatorisch wordt gebruikt. Een voorbeeld zou zijn: Wanneer de score op de competentie 'besluitvaardigheid' matig of slechter is (<= 2 uit 5) krijgt de kandidaat een negatieve beoordeling. Is de score op de competentie besluitvaardigheid voldoende of hoger (>= 3 uit 5) worden de competentiescores vervolgens gemiddeld en worden de kandidaten met de hoogste scores aangenomen.

Kortom, beslisregels moeten ingezet worden om informatie binnen een selectie-instrument te combineren, maar ook om informatie uit meerdere selectie-instrumenten te combineren om een selectiebeslissing te nemen. De net beschreven principes zijn niet alleen van toepassing bij de selectieprocedure zoals die wordt uitgevoerd door de LSR, maar ook bij de lokale selectie en het adviesbureau LTP. Het is belangrijk te benadrukken dat zelfs simpele beslisregels tot betere selectiebeslissingen en selectieadviezen leiden dan het holistisch oordeel van ervaren professionals (Kuncel et al., 2013; Morris et al., 2015; Yu & Kuncel, 2020, 2022). De belangrijkste standaardisatiecomponenten worden in Tabel 1 weergegeven en uitgelegd.

Tabel 1 *Standaardisatiecomponenten in selectiegesprekken*

Component	Betekenis	Voorbeeld	Te vermijden
1. Vragen zijn gerelateerd aan de baan	Vragen zijn gericht op competenties die relevant voor de taak zijn. Via een *taakanalyse* kunnen experts middels gesprekken of vragenlijsten gevraagd worden kritische gebeurtenissen te benoemen. Vragen zijn gericht op hoe men is omgegaan met situaties uit het verleden (gedragsvragen), hypothetische situaties die zich in de toekomst voor kunnen doen (situationele vragen), of kennis die relevant is voor de baan.	*"Beschrijf een situatie waarin je hebt samengewerkt met diverse belanghebbenden met verschillende perspectieven en belangen. Hoe zorgde je voor een goede samenwerking?"* (Gedragsvraag)	Vragen stellen die geen duidelijke directe relatie tot de baan hebben zoals "*Wat zijn je sterke en zwakke punten?*"
2. Standaardisatie van vragen	Aan alle kandidaten worden dezelfde vragen gesteld, in dezelfde volgorde.		Selecteurs vragen volledig vrij te laten kiezen, alleen een lijst met globale onderwerpen gebruiken die zullen worden besproken, de vragen laten afhangen van de kandidaat.
3. Evaluatie-standaardisatie	De antwoorden worden *per vraag* numeriek en *onafhankelijk* (indien meerdere selecteurs een gesprek bijwonen) beoordeeld, bijvoorbeeld op een vijfpuntsschaal. Daarbij is gedefinieerd wat goede en minder goede antwoorden zijn, door gebruik te maken van geankerde beoordelingsschalen. Er vindt geen overleg tussen selecteurs plaats.	Zie de noot[1)] voor voorbeelden, uitgaand van de voorbeeldvraag hierboven (component 1).	Slechts per gesprek of per thema scoren, antwoorden niet numeriek scoren. Hierdoor kan ook de mechanische methode (zie component 5) niet gebruikt worden.
4. Controleer aanvullende informatie	Gesprekken zijn betrouwbaarder en meer valide wanneer aanvullende informatie zoals motivatiebrieven, cv's, test scores, en rapporten van assessments niet vooraf beschikbaar zijn voor selecteurs, omdat verschillende selecteurs informatie verschillend en inconsistent gebruiken (Dipboye et al., 1984; McDaniel et al., 1994). Bovendien is de informatie verkregen uit het gesprek dan niet langer onafhankelijk van de andere informatie die in de procedure wordt gebruikt.		Selecteurs *voorafgaand aan het gesprek* aanvullende informatie te laten bekijken, verschillende aanvullende informatie verstrekken voor verschillende kandidaten, verschil in het wel/niet bestuderen van aanvullende informatie tussen selecteurs.
5. Numerieke beoordelingen worden via een beslisregel gecombineerd	Numerieke beoordelingen per vraag worden via een beslisregel tot een oordeel gecombineerd (mechanische combinatie). Dit voorkomt dat verschillende selecteurs de beoordelingen verschillend wegen. Dit is het geval wanneer selecteurs beoordelingen 'door nadenken', gebaseerd op hun oordeel, of via groepsoverleg combineren (holistische combinatie). De mechanische methode resulteert in tot 50% betere beslissingen dan de holistische methode (Kuncel et al., 2013).	Numerieke beoordelingen op de antwoorden worden per thema gemiddeld. Als meerdere selecteurs een gesprek bijwonen kunnen de gemiddeldes per thema over selecteurs heen gemiddeld worden. Vervolgens worden de scores per thema gemiddeld. Ook noncompensatorische beslisregels waarbij op kritische thema's voldoende moet worden gescoord zijn mogelijk.	Beoordelingen holistisch te combineren.

1) *Noot bij tabel 1*

1 (onvoldoende): Kandidaat heeft moeite om een voorbeeld te geven of geeft een vaag, algemeen antwoord. Noemt geen specifieke belanghebbenden of hun verschillende perspectieven. Legt geen strategieën uit voor samenwerking of hoe conflicten werden opgelost. Geen bewijs van initiatief om met belanghebbenden in gesprek te gaan.

2 (matig): Kandidaat geeft een minimaal of onduidelijk voorbeeld dat weinig relevant is. Noemt belanghebbenden, maar de uitleg over hun verschillen is niet duidelijk. Zwakke of onderontwikkelde strategieën om samenwerking te waarborgen. Beperkte of ineffectieve communicatie, zonder probleemoplossend vermogen te tonen.

3 (voldoende): Kandidaat geeft een duidelijk en relevant voorbeeld van samenwerking met belanghebbenden. Identificeert verschillende perspectieven en beschrijft enkele strategieën om hiermee om te gaan. Toont basisvaardigheden in communicatie en samenwerking. Enig bewijs van het oplossen van kleine conflicten of het balanceren van de belangen van belanghebbenden.

4 (ruim voldoende): Kandidaat geeft een gedetailleerd, specifiek en relevant voorbeeld. Benoemt duidelijk belanghebbenden met uiteenlopende perspectieven en conflicterende belangen. Beschrijft effectieve strategieën voor samenwerking, zoals het opbouwen van relaties, het faciliteren van open communicatie, of het onderhandelen over oplossingen. Toont proactief probleemoplossend vermogen en bevordert samenwerking, zelfs in uitdagende situaties.

5 (sterk): Kandidaat geeft een zeer gedetailleerd, inzichtelijk en complex voorbeeld met grote relevantie. Herkent en begrijpt diverse belanghebbenden met complexe en conflicterende perspectieven. Toont leiderschap in samenwerking door geavanceerde strategieën te gebruiken, zoals het opbouwen van vertrouwen, het afstemmen van gezamenlijke doelen, en het oplossen van conflicten met creatieve oplossingen. Heeft een diep begrip van de dynamiek tussen belanghebbenden, met bewijs van succesvolle, langdurige samenwerking en positieve resultaten voor alle betrokken partijen.

Standaardisatie en diversiteit

Het standaardiseren van selectie-instrumenten zoals gesprekken zorgt naast betere selectiebeslissingen (d.w.z. het selecteren van kandidaten die goed zullen presteren in de baan) ook voor kleinere verschillen in scores tussen kandidaten met en zonder etnisch-culturele achtergrond (Dahlke & Sackett, 2017). Ook zorgt standaardisatie voor grotere kansengelijkheid onder kandidaten en kan dus *onterechte* verschillen op basis van bijvoorbeeld etnisch-culturele achtergrond grotendeels voorkomen (Odijk et al., 2024). Het standaardiseren van de selectie-instrumenten, zoals de verschillende gesprekken, draagt dus ook bij aan de doelstelling om divers talent aan te trekken. Vervolgens moet echter ook ervoor gezorgd worden dat *de beslissing* op een gestandaardiseerde wijze via een beslisregel wordt genomen. Zelfs als dus gestandaardiseerde selectie-instrumenten van goede kwaliteit ingezet worden, maar de informatie uit deze instrumenten vervolgens holistisch, op basis van het menselijk oordeel gecombineerd wordt, kunnen deze verschillen weer geherintroduceerd worden (van Andel et al., 2019). Het is dus geen oplossing om selectie-instrumenten waarop bepaalde groepen kandidaten lager op scoren dan maar holistisch te gebruiken. Verschillen in scores tussen groepen worden daardoor niet 'opgelost'. Het maakt het slechts onduidelijk hoe selectie-instrumenten gebruikt worden. Het gebruik van gestandaardiseerde instrumenten en beslisregels zorgt dus ook ervoor dat alle kandidaten gelijk behandeld worden, de selectieprocedure transparant is en maakt een evaluatie in de toekomst makkelijker.

Meer informatie kan nadelig zijn

Het is ook belangrijk te realiseren dat geen enkel selectie-instrument of combinatie van instrumenten resulteert in 'perfecte' voorspellingen voor latere prestaties en gedrag. Deze maatstaf van perfectie wordt echter impliciet vaak gebruikt (Highhouse, 2008), en wordt vaak in het kader van analytische tests genoemd. Zo rapporteerden selecteurs van de LSR soms dat zij rechters kennen die laag scoorden op de analytische tests maar (ogenschijnlijk) een uitstekende rechter zijn, en daarom de analytische tests niet bruikbaar zouden zijn. Vergelijkbare uitzonderingen zullen voor ieder denkbaar selectie-instrument te vinden zijn. De vraag moet niet zijn of een selectie-instrument of procedure foutloze beslissingen oplevert, maar of een alternatief *betere* beslissingen zou opleveren. In de praktijk wordt dat betere alternatief vaak nagestreefd door meer selectie-instrumenten toe te voegen, om meer informatie te verzamelen. Daarbij is het echter heel belangrijk te benadrukken dat het beschikken over meer informatie over een kandidaat niet altijd tot betere selectiebeslissingen leidt. Sterker nog, het toevoegen van selectie-instrumenten kan de voorspellende waarde van de hele selectieprocedure zelfs *verminderen* (Murphy, 2019; Sackett, Dahlke, et al., 2017). Kausel et al. (2016) lieten zien dat selecteurs die alleen scores uit twee valide, gestandaardiseerde tests zagen betere selectiebeslissingen namen dan selecteurs die naast de testscores ook nog een beoordeling uit een niet gestandaardiseerd gesprek zagen. Het toevoegen van minder of irrelevante selectie-instrumenten, zoals niet gestandaardiseerde gesprekken, kan dus schadelijk zijn, en een gedachte zoals 'baat het niet dan schaadt het niet' gaat hier vaak niet op (Niessen et al., 2019). Zelfs als alleen maar valide selectie-instrumenten gebruikt worden kan het toevoegen van selectie-instrumenten de kwaliteit van selectiebeslissingen *verminderen,* als de toegevoegde instrumenten sterker samenhangen met de al gebruikte instrumenten dan met prestatie in de opleiding (Murphy, 2019). Een voorbeeld zou zijn een vragenlijst over persoonlijkheid welke een klein verband met prestatie in de opleiding toont toe te voegen aan een procedure waar al een vragenlijst over persoonlijkheid (of vergelijkbare aspecten) inzit die ook een klein verband met prestatie toont, maar deze twee vragenlijsten onderling zeer sterk samenhangen (zoals nu het geval is in het assessment). Daarom is een goed advies dat als instrumenten onderling sterker samenhangen dan zij samenhangen met prestatie in de opleiding, slechts één van de twee instrumenten te gebruiken. Hierbij kan dan het instrument gekozen worden die betere psychometrische kwaliteiten en theoretische onderbouwing laat zien, of tot betere uitkomsten in diversiteit leidt. Meer informatie geeft ons als selecteurs meer vertrouwen in onze beslissing, maar hoeft de *kwaliteit* van de beslissingen niet per se te verbeteren en kan deze dus soms zelfs verminderen (Dana et al., 2013; Kausel et al., 2016). In Hoofdstuk 3 geven wij advies over welke selectie-instrumenten behouden kunnen worden.

2.5 Overname advies uit de laatste evaluatie (Niessen & Meijer, 2019)

Om inzicht te geven in de verbetering van de selectieprocedure zoals die sinds het verschijnen van het vorige evaluatierapport (Niessen & Meijer, 2019) in gang is gezet hebben we eerst gekeken welke verbeteringen al tot stand zijn gekomen. In Tabel 2 is een overzicht weergegeven van de aanbevelingen uit het vorige evaluatierapport, reacties vanuit LSR, en een waardering van in hoeverre de aanbevelingen daadwerkelijk zijn overgenomen (zie de kolom 'Implementatie zoals beoordeeld door onderzoekers'). Daarbij bleek dat er enige aanbevelingen niet of onvoldoende zijn overgenomen en er wat misverstanden lijken te bestaan over of een aanbeveling wel of niet is overgenomen. Dit betreft voornamelijk het standaardiseren van (lokale) gesprekken en het gebruik van beslisregels. Wij herhalen daarom de belangrijkste nog niet geïmplementeerde aanbevelingen uit de laatste evaluatie in dit rapport.

Tabel 2 *Overzicht van aanbevelingen uit de laatste evaluatie en in hoeverre deze daadwerkelijk zijn overgenomen*

Aanbevelingen uit de laatste evaluatie (Niessen & Meijer, 2019)	Besluit & Opmerkingen LSR 2022	Implementatie zoals beoordeeld door onderzoekers
Selectie leden LSR		
De selectiegesprekken met kandidaat-leden van de LSR te standaardiseren, aangezien dit tot een hogere voorspellende waarde en accuratere selectiebeslissingen leidt.	**De aanbeveling wordt in zijn geheel overgenomen:** selectiegesprekken met kandidaat-leden LSR zullen meer gestandaardiseerd moeten plaatsvinden.	Op basis van informatie ontvangen door de LSR lijkt deze aanbeveling overgenomen te zijn.
Te overwegen om in de selectie van kandidaat-leden aandacht besteden aan kenmerken die positief samenhangen met het maken van accurate selectiebeslissingen, met name: empathisch vermogen, sociale vaardigheden, cognitieve capaciteiten en dispositional reasoning, eventueel in de vorm van een test.	**De aanbeveling wordt in zijn geheel overgenomen**: het profiel van de leden LSR wordt aangepast, en een eventuele test wordt onderzocht. Opmerking LSR: Tijdens de selectie vanaf 2023 zal dit worden gedaan dmv inzet LTP testen.	Uit gesprekken met medewerkers van de LSR en interne en externe leden bleek dat dit advies overgenomen is.
Training leden selectiecommissie		
Nieuwe leden meer mogelijkheden te bieden om zich te verdiepen in het beroep van rechter (externe leden) en op de selectieprocedure, bijvoorbeeld door zittingen bij te wonen en deel te nemen aan verschillende tests en assessments uit de selectieprocedure.	**De aanbeveling wordt in zijn geheel overgenomen,** met daarbij bijzondere aandacht voor verschillende type rechters (raadsheren, fiscalisten, etc.). Opmerking LSR: Uitgevoerd.	Kon niet beoordeeld worden.
Op kennismakingsdagen en terugkomdagen voor leden van de LSR meer tijd aan oefenen, rollenspelen en casussen te besteden, aangezien de leden hier behoefte aan hebben.	**De aanbeveling wordt in zijn geheel overgenomen**. Opmerking LSR: Uitgevoerd.	Uit gesprekken met medewerkers van de LSR en interne en externe leden bleek dat dit advies overgenomen is.
Een frame-of-reference training te (laten) ontwikkelen en aan te bieden aan selecteurs, aangezien dit de meest effectieve vorm van interviewtraining is. In een dergelijke training wordt veel aandacht besteed aan een gemeenschappelijke conceptualisatie van de verschillende competenties en eigenschappen die beoordeeld worden in interviews.	**De aanbeveling wordt in zijn geheel overgenomen.** Opmerking LSR: Uitgevoerd.	Uit gesprekken met medewerkers van de LSR en interne en externe leden bleek dat dit advies overgenomen is. De inhoud van de trainingen kon niet beoordeeld worden.

Aanbevelingen uit de laatste evaluatie (Niessen & Meijer, 2019)	Besluit & Opmerkingen LSR 2022	Implementatie zoals beoordeeld door onderzoekers
Organisatie selectieprocedure		
De kwaliteit en beschikbaarheid van digitale middelen voor het organiseren en registreren van de selectieprocedure en de doorlooptijd van de selectieprocedure ook in de toekomst een aandachtspunt moeten blijven, zeker in het geval van veranderingen in de procedure. Er lijken momenteel echter geen dringende aandachtspunten te zijn op dit gebied.	**Bijzondere aandachtspunten:** • Zorg voor een toegankelijk digitaal aanmeldformat inclusief informatievoorziening op rechtspraak.nl dat alle doelgroepen (i.h.k.v. diversiteit) aanspreekt en heldere informatie geeft. De informatie op de site lijkt verouderd. Houdt deze beter bij. Opmerking LSR: Uitgevoerd > recruitmentsysteem Ubeeo geimplementeerd + proces optimalisatie in 2022.	Kon niet beoordeeld worden.
Briefselectie		
Te expliciteren wat er verstaan wordt onder relevante juridische werkervaring en wanneer er voldaan is aan voldoende maatschappelijke ervaring. Daar worden momenteel geen formele definities voor gehanteerd.	**De aanbeveling overnemen.** Opmerking LSR: In de Handreiking LSR is aanvullende informatie opgenomen over wat wordt verstaan onder externe werkervaring.	De Handreiking LSR bevat inderdaad informatie over wat wordt verstaan onder externe werkervaring. Voor maatschappelijke nevenactiviteiten blijft het minder duidelijk. Wij bevelen aan om dit in de scorelijst te expliciteren (zie 3.1 Briefselectie).
Ook expliciete definities te hanteren van andere zaken waarop kandidaten tijdens de briefselectie beoordeeld worden. Deze definities op te nemen in een beoordelingsformulier dat gebruikt kan worden voor de briefselectie, zodat de briefselectie gestandaardiseerd kan worden uitgevoerd.	**De aanbeveling wordt in zijn geheel overgenomen,** de briefselectie dient verder te worden gestandaardiseerd. Er moet echter ruimte zijn voor maatwerk. Opmerking LSR: Uitgevoerd, format voor briefselectie is gemaakt.	Deze aanbeveling is gedeeltelijk maar niet voldoende overgenomen. Er bestaat inmiddels wel een scorelijst, maar daarin ontbreken definities en wordt de scorelijst niet consistent gebruikt (zie 3.1 Briefselectie). Het gebruik van een scorelijst dient juist om maatwerk te voorkomen, omdat maatwerk tot slechtere en minder diverse selectiebeslissingen leidt dan het consistente gebruik van een scorelijst (Kuncel et al., 2013; Odijk et al., 2024).
Analytische test		
De Taalgevoeltest in de toekomst niet meer op te nemen in de selectieprocedure, aangezien deze onvoldoende betrouwbare scores oplevert. Indien het testen van taalgevoel wenselijk wordt gevonden kan een ander instrument van voldoende kwaliteit worden gezocht.	**De aanbeveling wordt in zijn geheel overgenomen:** de taalgevoeltest zal vanaf 2020 niet meer worden afgenomen. Opmerking LSR: Uitgevoerd.	Deze aanbeveling is uitgevoerd.

Aanbevelingen uit de laatste evaluatie (Niessen & Meijer, 2019)	Besluit & Opmerkingen LSR 2022	Implementatie zoals beoordeeld door onderzoekers
Tests te gebruiken die in voldoende mate zijn onderzocht op psychometrische kwaliteit (betrouwbaarheid en validiteit) en waarvan de resultaten laten zien dat de tests aan de wetenschappelijke standaard voldoen. Bij voorkeur tests te gebruiken die op relevante punten minstens als 'voldoende' zijn beoordeeld door de COTAN.	**De aanbeveling wordt in zijn geheel overgenomen:** vanaf 2020 worden tests afgenomen die aan de standaard voldoen. Opmerking LSR: Matrix en verbale analogieën zijn COTAN gecertificeerd. Watson Glaser niet, wordt ingekocht door LTP.	Deze aanbeveling is niet voldoende overgenomen. Wij merken op dat de Matrix test en verbale test (TVRA) door de COTAN zijn *beoordeeld*. Dat oordeel was echter niet 'voldoende' op alle relevante punten. De COTAN verstrekt geen certificering. Wij bespreken de kwaliteit van de analytische tests en de andere tests die worden ingezet in Hoofdstuk 3.
Wat betreft de gebruikte tests: • De TVRA of een gelijksoortig instrument te handhaven, aangezien de scores op deze test samenhangen met beoordelingen uit de rio-opleiding. • Bij het gebruik van meerdere tests te overwegen om de tests compensatorisch te scoren, zodat een lage score gecompenseerd kan worden door een hoge score op een andere test, om het aantal vals-negatieve beslissingen terug te dringen. • Te overwegen om de Matrix test te vervangen door de Watson-Glaser test, of een ander instrument van goede kwaliteit, dat kritisch redeneervermogen meet, aangezien de scores op deze test samenhangen met beoordelingen uit de rio-opleiding en positief werd gewaardeerd door kandidaten.	**De aanbeveling wordt gedeeltelijk overgenomen.** Vanaf 2020 worden tests afgenomen die het volgende meten: • Verbaal-redeneervermogen • Kritisch-redeneervermogen • Abstract-redeneervermogen Kandidaten dienen bij de eerste twee tests de normscore te behalen. Bij de derde is dit niet noodzakelijk. De tests worden compensatorisch gescoord. Opmerking LSR: Uitgevoerd.	Deze aanbeveling is deels overgenomen. Zoals vermeld moeten kandidaten voor *ieder* van de eerste twee tests de normscore behalen, en is compensatie beperkt mogelijk. Deze twee tests worden dus zowel noncompensatorisch als compensatorisch gebruikt. De aanbeveling was bovendien de Watson-Glaser test *in plaats van* de Matrix test gebruiken en niet als toevoeging. Op dit moment wordt de Matrix test holistisch in het assessment gebruikt (zie 3.4.5 Competentie-oordelen).
Zeer voorzichtig om te gaan met uitzonderingen op de gehanteerde grensscores op basis van leeftijd of ervaring. Het verband tussen leeftijd en de scores op de analytische tests is terecht een punt van aandacht. Wat de oorzaak van dit verschil is, is echter niet duidelijk en het verband tussen het aantal jaren relevante werkervaring en latere baanprestaties is gering.	**De aanbeveling wordt in zijn geheel overgenomen:** er worden in beginsel geen uitzonderingen toegestaan. Opmerking LSR: Gerede kandidaatprocedure is een uitzondering hierop.	Deze aanbeveling is slechts gedeeltelijk overgenomen, daar er voor gerede kandidaten wel uitzonderingen worden gemaakt. Begin 2025 vervalt de gerede kandidaatprocedure en is de analytische test het startpunt van de selectieprocedure.
De mogelijkheid tot herkansing niet meer aan te bieden, aangezien dit een negatief effect heeft op de voorspellende waarde van de testscores en het aantal vals-positieve beslissingen doet toenemen. Indien het toch wenselijk wordt gevonden om een herkansing aan te bieden, kandidaten alle onderdelen te laten herkansen en de gemiddelde scores over de pogingen te gebruiken om het aantal vals-positieve beslissingen te verminderen.	**De aanbeveling wordt in zijn geheel overgenomen:** er worden geen herkansingen meer aangeboden – enkel in geval van overmacht (storing IT, geluidsoverlast). Opmerking LSR: Uitgevoerd, geen herkansing meer mogelijk.	Deze aanbeveling is uitgevoerd.

Aanbevelingen uit de laatste evaluatie (Niessen & Meijer, 2019)	Besluit & Opmerkingen LSR 2022	Implementatie zoals beoordeeld door onderzoekers
Voorgesprek en lokale selectie		
Te overwegen om het voorgesprek te laten vervallen, aangezien het geen doorslaggevend onderdeel is en de betrouwbaarheid en de voorspellende waarde van de beoordelingen uit het voorgesprek gering lijken.	**De aanbeveling wordt gedeeltelijk overgenomen:** het voorgesprek in huidige vorm komt te vervallen. De lokale selectieprocedure wordt uitgevoerd ná de testfase en voor het assessment. Opmerking LSR: Uitgevoerd.	Deze aanbeveling is uitgevoerd.
Het komt voor dat kandidaten die succesvol door de procedure van de LSR zijn gekomen lokaal geen rio-plaats verwerven. Om het aantal succesvolle kandidaten dat lokaal geen rio-plaats kan vinden terug te dringen, kan overwogen worden om evenveel r(h)io's te selecteren als er plaatsen zijn. De lokale selectie kan dan ofwel: • Komen te vervallen en worden vervangen door plaatsing. • Al tijdens de briefselectie en/of het voorgesprek plaatsvinden.	Opmerking LSR: Uitgevoerd (2022 afgesproken in PRO alle groen licht kandidaten een opleidingsplek aan te bieden).	Deze aanbeveling is uitgevoerd.
Indien de lokale selectie gehandhaafd blijft bevelen we aan om: • De lokale selectieprocedure te standaardiseren. Dat lijkt op basis van de gevoerde gesprekken nu niet het geval te zijn. • Expliciet geoorloofde afwijzingsgronden af te spreken. • De doelstelling van de lokale selectie duidelijk te communiceren naar kandidaten. • Te zorgen voor meer bekendheid bij de lokale gerechten over de landelijke procedure, om het vertrouwen in de procedure gehanteerd door de LSR te vergroten.	**De aanbeveling wordt gedeeltelijk overgenomen:** • training van de verschillende selecteurs (landelijk en lokaal), zodat zij conform de gestandaardiseerde werkwijze kunnen werken. Aan hen kan bijvoorbeeld de frame-of- reference-training aangeboden worden. • Structureer de gesprekken, werk conform het format, maar niet dusdanig dat het gehele gesprek wordt dichtgeregeld. Er moet ruimte blijven voor creativiteit, vrijheid en flexibiliteit tijdens het gesprek, alsmede tijdens het overleg tussen de selecteurs na afloop van de gesprekken. • Alle selecteurs krijgen een training om gestandaardiseerd te kunnen selecteren. Opmerking LSR: Deels uitgevoerd, er wordt niet gewerkt met een landelijk format voor de lokale gesprekken.	Deze aanbeveling werd *gedeeltelijk* overgenomen. Er worden trainingen aangeboden, maar de lokale gesprekken zijn nauwelijks gestandaardiseerd (zie 3.3 Lokale gesprekken). Ook hier is besloten maatwerk mogelijk te maken tijdens het beoordelen en het de beraadslaging. Dit vermindert de kwaliteit van selectiebeslissingen (Kuncel et al., 2013) en kan worden voorkomen door het standaardiseren van gesprekken en het gebruik van expliciete beslis-regels, zoals het ook al in de laatste evaluatie aanbevolen werd (Niessen & Meijer, 2019).

Aanbevelingen uit de laatste evaluatie (Niessen & Meijer, 2019)	Besluit & Opmerkingen LSR 2022	Implementatie zoals beoordeeld door onderzoekers
Referenties		
De procedure rondom het opvragen van referenties te formaliseren en te standaardiseren door aan referenten duidelijk te maken wat voor informatie de LSR wel en niet wenst te ontvangen.	**Aanbeveling wordt geheel overgenomen:** kandidaten kunnen maximaal 2 referenties indienen, conform een vast format. Opmerking LSR: Ambtshalve referentie is in 2022 door LSR bekeken en informatie bij de uitvraag verduidelijkt.	Er bestaat een formulier met vragen die referenten beantwoorden. De antwoorden worden echter in de vorm van tekst gegeven en zijn niet gekwantificeerd (noch door de referent noch door de LSR).
Assessment		
Vragenlijsten op basis van zelfrapportage niet, of alleen voor 'select- out' beslissingen te gebruiken, gezien de mogelijkheid tot sociaal- wenselijk antwoorden en de lage criteriumvaliditeit in selectiesituaties.	**De aanbeveling wordt geheel overgenomen:** vanaf 2020 wordt er geen zelfrapportage meer uitgevraagd. Opmerking LSR: Uitgevoerd.	Deze aanbeveling werd *niet* overgenomen. Er worden nog steeds meerdere zelfrapportagevragenlijsten ingezet in het assessment (zie 3.4.1 Zelfrapportagevragenlijsten).
Het assessment anders in te richten: • Het is het overwegen waard om het assessment in te richten als 'assessment center', bestaande uit opdrachten en rollenspellen die in inhoud en in vorm overeenkomen met belangrijke taken die in (het begin van) de rio-opleiding vervuld moeten worden. Dit zal waarschijnlijk een goede criteriumvaliditeit en positieve reacties van kandidaten opleveren, maar zal ook tot hogere kosten leiden • Het assessmentinterview te laten vervallen.	**De aanbeveling wordt gedeeltelijk overgenomen:** het assessment wordt vervangen door een assessmentcenter, bestaande uit rollenspellen die gericht zijn op het meten van voor het rechterschap relevante vaardigheden. Een gesprek met een psycholoog blijft onderdeel van de procedure. Opmerking LSR: Uitgevoerd.	Deze aanbeveling is niet uitgevoerd. Een assessment center houdt in dat er meerdere oefeningen gebruikt worden om meerdere dimensies te meten welke door meerdere onafhankelijke beoordelaars worden gescoord (Rupp et al., 2015). Op dit moment is het assessment slechts voor een klein deel ingericht als een assessment center, maar niet volgens bovenstaande standaard.
Eindgesprek		
Het eindgesprek verder te structuren, aangezien dit de betrouwbaarheid en de voorspellende waarde waarschijnlijk zal verhogen. De besluitvorming volgens een statistische procedure te laten verlopen, waarbij vooraf wordt gespecificeerd hoe zwaar ieder thema en iedere score gewogen moet worden. Dat geldt dan zowel voor de scores of beoordelingen uit het assessment als uit het eindgesprek. Dit zal de voorspellende waarde waarschijnlijk verhogen.	**De aanbeveling wordt gedeeltelijk overgenomen, zoals ook de lokale selectie wordt gestandaardiseerd (zie aldaar).** Ook hier geldt dat de klinische blik niet verloren mag gaan. Daarom zal er geen statistische procedure worden ingevoerd. Opmerking LSR: Uitgevoerd zoals aangegeven, de eindgesprekken zijn niet gestandaardiseerd en er is niet afgesproken hoe zwaar elke thema weegt tov anderen (alles even zwaar).	Deze aanbeveling is gedeeltelijk uitgevoerd. Ook hier geldt zoals bij de lokale selectie: De klinische blik (holistische combinatie) verminderd de kwaliteit van selectiebeslissingen en maakt het moeilijker om selectiebeslissingen te evalueren en te verbeteren omdat de beslissing niet meer transparant is.

Aanbevelingen uit de laatste evaluatie (Niessen & Meijer, 2019)	Besluit & Opmerkingen LSR 2022	Implementatie zoals beoordeeld door onderzoekers
Een eventuele beoordeling van het assessmentrapport en het voeren van de eindgesprekken niet door dezelfde personen te laten doen, zodat de informatie afkomstig uit beide onderdelen onafhankelijk van elkaar kan worden meegenomen.	**Deze aanbeveling wordt niet overgenomen.** Landelijke Selectiecommissie voert eindgesprek conform huidige gebruiken. Opmerking LSR: Advies niet overgenomen.	Deze aanbeveling werd niet overgenomen. Hierdoor wordt standaardisatiecomponent 4 uit Tabel 1 geschonden.
Cultuurwaardevrij selecteren		
Aan de cultuurwaardevrijheid van de gehanteerde tests en assessments blijvend aandacht te besteden. Momenteel lijken er weinig kandidaten met een migratieachtergrond te solliciteren voor een rio-plaats. Als het aandeel kandidaten met een migratieachtergrond toeneemt zullen er wellicht meer knelpunten wat betreft cultuurwaardevrij selecteren aan het licht komen.	**De aanbeveling wordt geheel overgenomen.** Opmerking LSR: Nog steeds actueel, er wordt getracht dit te verbeteren.	
Acceptatie van de procedure door de kandidaten		
Informatie te verstrekken aan kandidaten over de reden om de procedure op een bepaalde manier in te richten. Dit kan de acceptatie vergroten.	**De aanbeveling wordt geheel overgenomen.** Opmerking LSR: Uitgevoerd, website up to date + social media inzetten + vraagbaak pool.	Wij hebben niet actief geprobeerd de uitvoering van deze aanbeveling te evalueren omdat dit onderdeel geen hoofddoel van dit onderzoek vormde.
Waar mogelijk selectie-instrumenten te gebruiken die inhoudelijk goed aansluiten bij de rio-opleiding en de functie van rechter, zoals een ‘assessment center’. Deze instrumenten zorgen waarschijnlijk voor een grotere tevredenheid van kandidaten (en hebben een goede criterium-validiteit), maar zijn kostbaarder dan het huidige assessment.	**De aanbeveling wordt geheel overgenomen.** Zie assessment. Opmerking LSR: Uitgevoerd.	Deze aanbeveling werd niet overgenomen, zie onder assessment.

Aanbevelingen uit de laatste evaluatie (Niessen & Meijer, 2019)	Besluit & Opmerkingen LSR 2022	Implementatie zoals beoordeeld door onderzoekers
Beoordeling in de rio-opleiding		
Indien mogelijk beoordelingen niet per taakniveau te laten uitvoeren, maar een algemene beoordeling te hanteren en het gewenste prestatie-niveau aan te passen aan de mate van vordering in de opleiding. Voor evaluatieonderzoek is het noodzakelijk dat beoordelingen uit de opleiding- samen geanalyseerd kunnen worden. Ook voor het in kaart brengen van de ontwikkeling van r(h)io's is het inzichtelijker als verschillende beoordelingen met elkaar vergeleken kunnen worden. De beoordelingslijst zo in te richten dat iedere vraag onder één competentie geschaard kan worden. De huidige structuur heeft waar-schijnlijk tot gevolg dat de competentiescores niet empirisch van elkaar onderscheiden kunnen worden. Een weging toe te passen, zodat compe-tenties die op basis van een groter aantal vragen worden beoordeeld niet langer (onbedoeld) meer gewicht krijgen in de totaalscore.	**Over deze aanbeveling heeft geen besluitvorming plaatsgevonden.** Het DB LBR heeft aangegeven om te gaan onderzoeken of het mogelijk is om een wegingsfactor 'taakniveau' toe te voegen aan de scorelijst, voor die criteria waar het taakniveau er ook toe doet.	De beoordeling in de rio-opleiding is nog steeds een aandachtspunt (zie Hoofdstuk 5 en Bijlage 3).
Beschikbaarheid gegevens		
In de nabije toekomst een beter, completer en handzamer systeem voor het bewaren en opslaan van gegevens te realiseren. Dit maakt het mogelijk om in de toekomst evaluatieonderzoek uit te voeren op basis van meer en meer complete gegevens, wat tot nauwkeurigere conclusies zal leiden.	Er is een bestuurlijke wens voor een dergelijk systeem. De financiële middelen ontbreken echter voor 2019.	Inmiddels werkt het systeem voor het bewaren van gegevens beter. Blijvende aandachtspunten zijn ervoor te zorgen dat de data zoals beoordelingen uit gesprekken daadwerkelijk worden opgeslagen.
Vervolgonderzoek		
Gezien de beperkingen van de gegevens die in het huidige onderzoek gebruikt konden worden voor statistische analyses, bevelen we aan om dit onderzoek, na aanpassingen van de selectieprocedure en het realiseren van een verbeterd systeem voor het bijhouden van gegevens, nog eens te herhalen.	**De aanbeveling wordt overgenomen.**	Deze aanbeveling werd overgenomen.

3 De kwaliteit van de selectie-instrumenten, het gebruik en ervaringen van betrokkenen

In dit hoofdstuk bespreken wij de kwaliteit van de individuele selectie-instrumenten en hoe deze instrumenten gebruikt en ervaren worden door betrokkenen. Daarbij onderscheiden we *gebruikers* (LSR-selecteurs, selecteurs betrokken bij de lokale selectie, LTP-adviseurs, medewerkers van het bureau LSR) en *oud-kandidaten* (geselecteerde rio's die de selectieprocedure ervaren hebben), omdat deze groepen in hun percepties van de selectieprocedure kunnen verschillen. We bespreken de kwaliteit van de selectie-instrumenten en het gebruik ervan gezamenlijk in één hoofdstuk omdat niet alleen de kwaliteit van een instrument maar ook de manier waarop het ingezet wordt bepalend is voor de kwaliteit van selectiebeslissingen.

Methode

Om in kaart te brengen hoe gebruikers de selectie-instrumenten gebruiken en de selectieprocedure ervaren zijn documenten van de LSR geanalyseerd en gesprekken gevoerd met 19 personen, waarvan twee van het Bureau LSR, drie van het Presidium, drie interne LSR-selecteurs, drie externe LSR-selecteurs, vier personen die betrokken zijn bij de lokale selectie, en een LTP-adviseur. Om in kaart te brengen hoe oud-kandidaten de selectieprocedure hebben ervaren is een online-vragenlijst afgenomen onder geselecteerde oud-kandidaten (zie Bijlage 2 voor de vragenlijst en informatie over de steekproef). Per selectie-instrument zijn er vragen gesteld over vier relevante procedurele rechtvaardigheidscriteria; de effectiviteit en eerlijkheid van het instrument, de relatie met de baan, de ogenschijnlijke validiteit en de kans om vaardigheden en kwaliteiten te laten zien (Bauer et al., 2001; Steiner & Gilliland, 1996). Respondenten beoordeelden deze vragen op een vijfpuntsschaal (zie Bijlage 2), maar konden hun antwoorden op deze vragen indien gewenst tekstueel toelichten. De antwoorden op deze open vragen zijn geanalyseerd aan de hand van thematische analyse (Braun & Clarke, 2006). Via de online vragenlijst werden oud-kandidaten ook gevraagd de selectie-procedure als geheel te beoordelen. Over het algemeen waren de respondenten positief over de selectieprocedure en kwamen de resultaten sterk overeen met de resultaten uit de laatste evaluatie (Niessen & Meijer, 2019). Wij bespreken daarom in dit hoofdstuk vooral de specifieke ervaringen per selectie-instrument.

De reacties van afgewezen oud-kandidaten zijn vaak gemiddeld negatiever dan de reacties van geselecteerde oud-kandidaten (Schmitt et al., 2004). Dit was ook het geval in de laatste evaluatie van de selectieprocedure voor rio's (Niessen & Meijer, 2019). Helaas konden aan dit onderzoek alleen personen deelnemen die ook daadwerkelijk geselecteerd werden; contactgegevens van afgewezen kandidaten waren niet voorhanden. De reacties van een aantal kandidaten die meer dan een keer gesolliciteerd hebben en dus wel minimaal een negatief besluit hebben gehad konden wel meegenomen worden. Daarom moeten de volgende bevindingen met dit punt in het achterhoofd geïnterpreteerd worden.

Uitleg van vaktermen en interpretatie van statistieken

Voor ieder selectie-instrument volgt eerst een beschrijving, gevolgd door een beoordeling van de betrouwbaarheid, constructvaliditeit en criteriumvaliditeit. *Betrouwbaarheid* is de mate waarin een testscore of een beoordeling vrij van meetfout is. Hier gaat het dus om de nauwkeurigheid van een meting. Betrouwbaarheid wordt meestal uitgedrukt volgens een coëfficiënt tussen 0 en 1, waarbij volgens de COTAN (Commissie Testaangelegenheden Nederland) >= .80 als voldoende en >= .90 als

goed wordt beschouwd voor individuele selectiebeslissingen (Evers et al., 2010). Bij instrumenten zoals selectiegesprekken is de interbeoordelaarsbetrouwbaarheid belangrijk. Dit is de mate waarin een beoordeling onafhankelijk is van de persoon die de beoordeling uitvoert. *Constructvaliditeit* is de mate waarin een instrument erin slaagt datgene te meten wat men beoogt te meten (bijvoorbeeld, meet een persoonlijkheidsvragenlijst inderdaad persoonlijkheid, en niet onbedoeld ook cognitieve capaciteiten). *Criteriumvaliditeit* is de mate waarin een test voorspellende waarde heeft voor relevante criteria zoals prestaties in banen of opleidingen (bijvoorbeeld, voorspelt een analytische test prestatie in de rio-opleiding). Om de criteriumvaliditeit te bepalen is op basis van data over 148 rio's het verband tussen scores en beoordelingen uit de selectieprocedure en beoordelingen uit de rio-opleiding berekend. De prestatie in de rio-opleiding was de som van de beoordelingen op alle 69 items die gebruikt werden om rio's te beoordelen op relevante competenties, zoals analytisch vermogen, overredingskracht, en samenwerken. Meer informatie over de prestatiemaat is te vinden in Bijlage 3. Er werd ook overwogen om verbanden tussen de selectie-instrumenten en een dichotoom uitkomstmaat 'succes' (d.w.z., heeft iemand de opleiding succesvol afgerond of niet) te onderzoeken. Dit werd uiteindelijk niet gedaan omdat er maar zeven rio's waren die aan het eind van de opleiding afgevallen zijn en het aantal afvallers dus te klein was om deze uitkomstmaat in analyses te gebruiken. Criteriumvaliditeit wordt uitgedrukt volgens een correlatiecoëfficiënt tussen -1 en 1 waarbij 0 betekent dat er geen verband bestaat. Een negatieve coëfficiënt betekent een negatief verband en een positieve coëfficiënt betekent een positief verband. Wij interpreteren correlaties van r = |.10|, r = |.30|, en r = |.50| als klein, middelgroot, en groot (Cohen, 1988). Rondom deze coëfficiënten zijn ook 95% betrouwbaarheidsintervallen weergegeven, welke de onnauwkeurigheid van de schatting van de correlatie duidelijk maken. Omdat een correlatiecoëfficiënt tussen een score op een selectie-instrument en prestatie in de opleiding alleen voor rio's berekend kan worden die ook geselecteerd zijn hebben wij de gerapporteerde correlatiecoëfficiënten gecorrigeerd voor *restrictie in range*. Een toelichting waarom dit nodig is en hoe dit gedaan werd is te vinden in Bijlage 3. Bijlage 3 bevat ook de ongecorrigeerde correlatiecoëfficiënten. De gecorrigeerde correlaties zijn niet voor onbetrouwbaarheid van de prestatiebeoordelingen gecorrigeerd, zoals wel wenselijk is, omdat die betrouwbaarheid op basis van de beschikbare gegevens niet geschat kon worden. Dit heeft een drukkend effect op de correlaties. Wel komen de schattingen van de correlaties goed overeen met de schattingen uit de literatuur (Sackett et al., 2022). Verder is er, net zoals in het vorige rapport (Niessen & Meijer, 2019), geen analyse uitgevoerd om de constructvaliditeit, ofwel de onderscheidbaarheid van de verschillende competenties te onderzoeken. Dit zal per definitie niet het geval zijn, omdat meerdere vragen uit de beoordelingslijst bij meerdere competentiescores worden meegeteld. Wij merken op dat de huidige manier waarop prestatie in de opleiding beoordeeld wordt vrij complex is en beoordelingen op verschillende taakniveaus een vergelijking tussen rio's moeilijk maakt. Wij raden aan om te evalueren of en hoe prestatie in de opleiding (het criterium) eenvoudiger kan worden gemaakt.

Ook wordt beoordeeld in hoeverre het gebruik van een instrument invloed heeft op de diversiteit van geselecteerde kandidaten. Daarvoor is op basis van data uit het rio-volgsysteem gekeken in hoeverre er verschillen in scores tussen mannen en vrouwen bestaan (verschillen op basis van etnisch-culturele achtergrond konden niet onderzocht worden omdat data hierover ontbrak). Deze analyses zijn gebaseerd op data van 182 mannen en 250 vrouwen. Voor andere kandidaten was geen informatie over hun geslacht aanwezig. Verschillen in scores tussen mannen en vrouwen zijn uitgedrukt via de coëfficiënt *Cohen's d*, waarbij een d van 0.2, 0.5, en 0.8 als klein, middelgroot, en groot effect geïnterpreteerd werd (Cohen, 1988). Een positieve d waarde betekent in dit rapport dat mannen hoger scoorden dan vrouwen. Een negatieve d waarde betekent dat vrouwen hoger scoorden dan

mannen. Ook is het verband tussen leeftijd en scores op een selectie-instrument bekeken en uitgedrukt via correlatiecoëfficiënten (voor de interpretatie zie de uitleg van 'Criteriumvaliditeit' hierboven). Voor deze analyses zijn gegevens over kandidaten gebruikt waarvan informatie over prestaties tijdens de selectieprocedure en beoordelingen van prestaties tijdens de opleiding beschikbaar waren. De gegevens zijn aangeleverd door de LSR, Studiecentrum Rechtspleging, en LTP.

Het gebruik van een selectie-instrument dat verschillen in scores tussen groepen laat zien betekent echter niet per se dat het een *onterecht* nadeel voor de groep met lagere scores oplevert. Daarvoor moet ook onderzocht worden of de prestaties in de opleiding op basis van het selectie-instrument systematisch onderschat worden voor de lager scorende groep ('predictieve bias'). In dit onderzoek kon predictieve bias niet worden onderzocht omdat de steekproef te klein was. Om de kwaliteit van de selectie-instrumenten te beoordelen zijn de beoordelingen van de COTAN en de handleidingen van LTP geraadpleegd, en is de data uit het rio-volgsysteem geanalyseerd.

3.1 Briefselectie

De briefselectie vindt in twee stappen plaats. Eerst controleert de LSR de formele eisen. Daarna wordt iedere kandidaat door twee personen beoordeeld; iemand van het selecterende lokale gerecht en een lid-secretaris vanuit de LSR. Daarbij heeft het lid-secretaris een adviserende rol. Naar aanleiding van de laatste evaluatie van de selectieprocedure is een scorelijst ten behoeve van de briefselectie opgesteld waarin factoren zoals opleiding, werkervaring, nevenactiviteiten, motivatie, en taalvaardigheden op een vijfpuntsschaal worden beoordeeld. Daarnaast bestaat er ook de mogelijkheid om bijzonderheden zoals diversiteit, reistijd/bereidheid om te verhuizen, en het afronden van een dissertatie te beoordelen. Wij merken op dat (relevante) werkervaring, gemeten in tijd, geen voorspellende waarde heeft voor prestaties in banen en opleidingen, onder andere omdat werkervaring niet betekent dat iemand in eerdere banen *goed* functioneerde (Van Iddekinge et al., 2019). Bovendien bestaan er middelgrote verschillen in werkervaring tussen kandidaten met en zonder etnisch-culturele achtergrond (Sackett et al., 2022). Ook kunnen factoren zoals motivatie en taalvaardigheden beter door gestandaardiseerde gesprekken en tests gemeten worden. Deze factoren kunnen niet goed door motivatiebrieven worden gemeten en indrukken o.b.v. motivatiebrieven hebben geen voorspellende waarde voor prestaties in opleidingen (Ferguson et al., 2000; Niessen & Neumann, 2022).

Uiteindelijk wordt holistisch besloten of de kandidaat wel of niet doorgaat naar de volgende fase van het selectieproces. Er bestaan dus geen vooraf vastgelegde grensscores die een kandidaat op de factoren moet behalen en geen vooraf opgestelde beslisregel waarmee de scores op de verschillende factoren tot een oordeel gecombineerd worden. Verder zijn er geen ankers die per factor aangeven wanneer een kandidaat een bepaalde score zou moeten krijgen (bijvoorbeeld, wanneer is de werkervaring minder goed of goed). Daarmee is de briefselectie niet voldoende gestandaardiseerd.
De betrouwbaarheid, constructvaliditeit, criteriumvaliditeit, en de invloed van de briefselectie op diversiteit kon niet geëvalueerd worden, omdat er geen scores van de brieven beschikbaar waren. Wetenschappelijk onderzoek naar predictieve bias op basis van cv's en motivatiebrieven is schaars. De literatuur laat wel zien dat het standaardiseren van een briefselectie helpt om te voorkomen dat kandidaten met een etnisch-culturele achtergrond lagere beoordelingen ontvangen als zij even geschikt zijn voor een functie als kandidaten zonder etnisch-culturele achtergrond (Odijk et al., 2024). Daarom raden wij ook op basis van de doelstelling om cultuurwaardevrij te selecteren aan om de briefselectie te standaardiseren.

3.1.1 Ervaringen van gebruikers

In het algemeen werd de briefselectie door gebruikers als positief en nuttig ervaren, vooral omdat men vond dat de motivatiebrief en cv veel nuttige informatie oplevert. Er werd vermeld dat de scorelijsten ter beoordeling altijd meegestuurd worden naar de gerechten, maar dat deze nauwelijks ingevuld en teruggestuurd worden. Deze opvatting wordt gedeeld door andere betrokkenen. De scorelijst wordt eerder als een soort leidraad gebruikt om naar de factoren te kijken die in de scorelijst zijn opgenomen. Betrokkenen geven aan de factoren nauwelijks te scoren en dat gerechten vaak al een interne selectie hebben gedaan voordat zij de kandidaten met een lid-secretaris van de LSR bespreken. De scorelijst wordt dus niet op de juiste manier gebruikt omdat de factoren zelden gescoord worden en uiteindelijk niet op basis van een beslisregel gecombineerd worden. Meerdere betrokkenen gaven aan naar de woonplaats van een kandidaat te kijken, omdat de aanname bestaat daarmee te kunnen beoordelen of een kandidaat langdurig aan een gerecht verbonden zal blijven. Ook kijkt men of kandidaten ook bij een ander gerecht gesolliciteerd hebben, en zo ja, waar. Dit omdat de aanname bestaat dat gerechten in populaire regio's in Nederland aantrekkelijker zijn dan andere, en dat goede kandidaten die bij een gerecht in een populaire regio solliciteren daar aangenomen zullen worden. Een betrokkene gaf aan dat hier veel aannames gemaakt worden en daardoor soms potentieel geschikte kandidaten afgewezen worden. Naast de factoren die in de scorelijst zijn opgenomen kijkt men ook naar de algemene indruk. Verder gaf een betrokkene aan dat er veel factoren zijn die wel van belang zijn maar niet in de scorelijst zijn opgenomen. Wij concluderen op basis van het design van de scorelijst en de manier waarop gebruikers de scorelijst gebruiken dat de in Tabel 1 aanbevolen standaardisatiecomponenten onvoldoende toegepast worden.

3.1.2 Ervaringen van oud-kandidaten

De briefselectie werd in het algemeen gematigd positief beoordeeld en minder positief dan de meeste andere selectie-instrumenten. Figuren met de algemene beoordelingen en de beoordelingen op de vier rechtvaardigheidsprincipes van oud-kandidaten zijn voor alle selectie-instrumenten weergegeven in Bijlage 4. Vooral vond men de briefselectie niet geschikt om de eigen capaciteiten te kunnen tonen en om kandidaten goed van elkaar te kunnen onderscheiden. In totaal hebben 103 respondenten een toelichting gegeven. Het meest genoemde aspect was het gebrek aan transparantie: het was voor kandidaten erg onduidelijk waarop zij in deze stap beoordeeld werden (n = 22). Verder vond men de briefselectie geen effectieve methode (n = 14) en mistte men uitleg, vooral in het geval van een negatiefbesluit (n = 14). Ook rapporteerden enkele kandidaten dat zij bij het ene gerecht wel door de briefselectie kwamen, maar bij andere gerechten, soms gebaseerd op dezelfde brief, niet (n = 10). Dit lijkt regelmatig voor te komen, aangezien deze ervaring alleen door de 61 kandidaten die bij meer dan één gerecht solliciteerden gemeld kon worden. De reden hiervoor blijft onduidelijk. Een mogelijke verklaring is dat gerechten in populariteit verschillen, maar ook dat verschillende gerechten tot verschillende beoordelingen komen op basis van dezelfde brief. Verder merkten verschillende respondenten op dat zij de briefselectie wel als noodzakelijk beschouwen om een eerste selectie te maken (n = 8) en dat het onlinesysteem voor de briefselectie gebruiksonvriendelijk was (n = 8).

3.1.3 Aanbevelingen

Het gebruik van een scorelijst is een verbetering ten opzichte van de vroegere selectieprocedure, waarbij nog geen scorelijst voor de briefselectie gebruikt werd (Niessen & Meijer, 2019). Wij bevelen aan ervoor te zorgen dat de scorelijst in eerste instantie op de juiste manier gebruikt en ingevuld wordt. Verder bevelen wij aan de scorelijst aan de hand van de standaardisatiecomponenten in

Tabel 1 verder te standaardiseren. Dit houdt in per factor te concretiseren wanneer welke numerieke beoordeling op de vijfpuntsschaal wordt gegeven. Als er factoren in de scorelijst missen, zoals door sommige gebruikers opgemerkt werd, kunnen deze factoren als zij relevant voor de opleiding zijn in de scorelijst opgenomen worden. Verder is het belangrijk om een beslisregel te gebruiken om tot een beoordeling te komen. Op dit moment gebeurt dit nog holistisch. Zodra een beslisregel is opgesteld kan het lid-secretaris van de LSR o.a. ingezet worden om te controleren of de beslisregel ook consistent wordt toegepast. Op dit moment lijkt het lid-secretaris vooral als een sparringpartner ingezet te worden en herinnert hij of zij selecteurs bij een lokaal gerecht eraan om op holistische wijze op bepaalde factoren te letten.

Aanbeveling 1 – De scorelijst voor de briefselectie verder standaardiseren en voor consistent gebruik zorgen

- Beschrijvende ankers per factor aan de vijfpuntsschaal toe voegen om te definiëren wanneer een factor als goed of minder goed beoordeeld wordt (bijvoorbeeld: wanneer zijn studieresultaten matig of sterk? Zie ook standaardisatiecomponent 3 uit Tabel 1).
- Per gerecht een beslisregel opstellen om tot een oordeel te komen over wanneer een kandidaat wel of niet doorgaat.
- Ervoor zorgen dat de scorelijst ook daadwerkelijk bij gerechten voor iedere kandidaat consistent gebruikt wordt. Op dit moment wordt de scorelijst slechts als ‘leidraad’ gebruikt (lage vorm van standaardisatie).

3.2 Analytische tests

De analytische tests vinden na de briefselectie plaats en zijn voor iedereen verplicht.
Er worden vier analytische tests afgenomen.

1. De test voor verbale relaties en analogieën (TVRA) voor verbaal redeneervermogen (ontwikkeld door LTP)
2. De Matrix test voor abstract redeneervermogen (ontwikkel door LTP)
3. De Watson-Glaser test voor kritisch redeneervermogen
4. De Numeriek Inzicht Test (NIT, ontwikkeld door LTP)

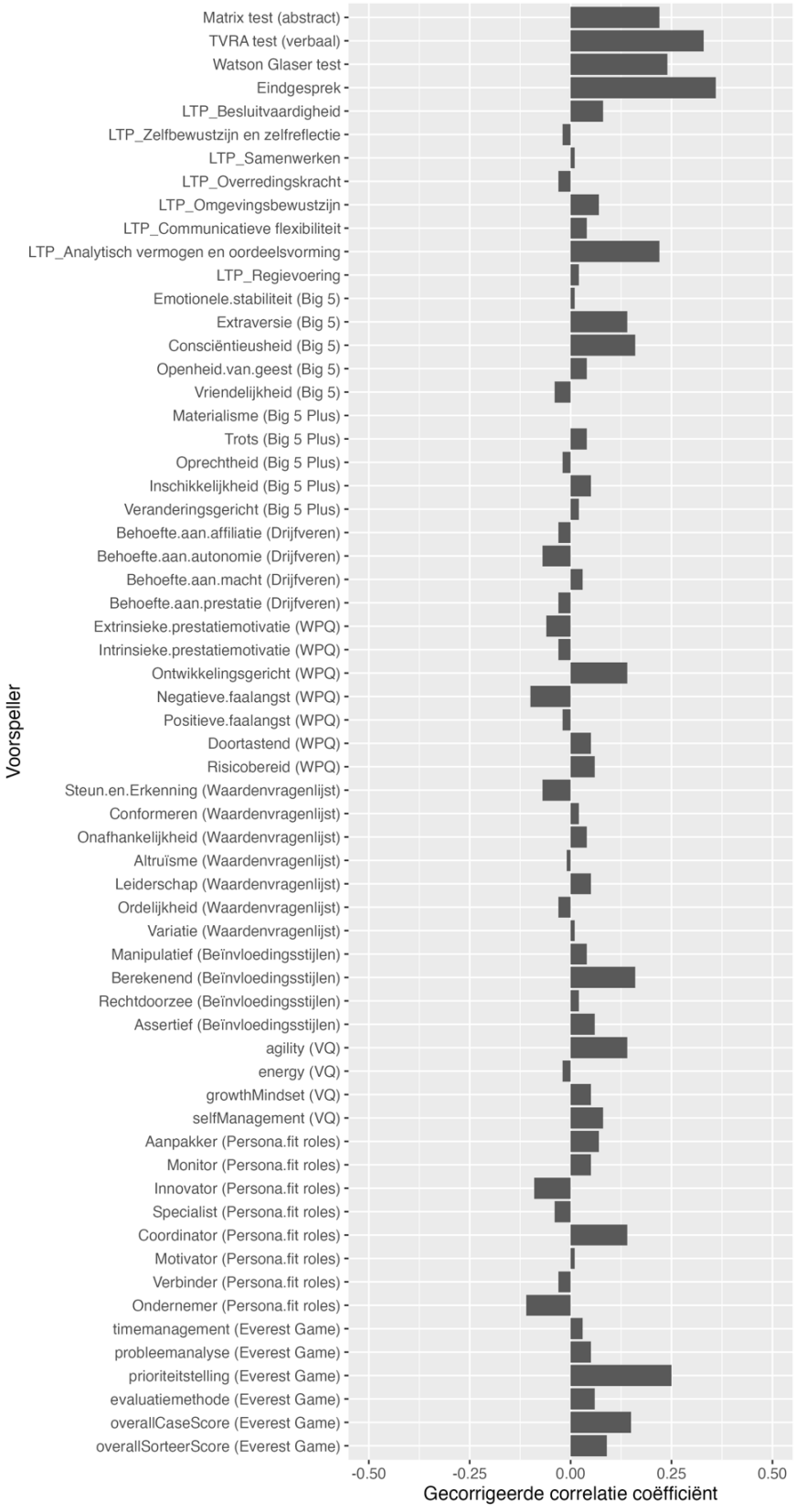
Van deze vier tests worden alleen de TVRA en de Watson-Glaser in deze fase van de selectie-procedure gebruikt om te beslissen wie doorgaat naar de lokale selectie. Wel worden alle vier tests in deze fase afgenomen. De Matrix test wordt echter pas in het assessment meegewogen (zie 3.4.5 Competentie-oordelen). De NIT werd pas na de start van dit evaluatieonderzoek in januari 2024 aan de selectieprocedure toegevoegd, als aanvulling op de al bestaande analytische tests. De NIT wordt op dit moment helemaal niet meegewogen omdat de LSR eerst wil onderzoeken of deze test voorspellende waarde heeft voor prestatie in de rio-opleiding. Dit wordt momenteel ook zo naar kandidaten gecommuniceerd.

De ruwe scores op de TVRA worden omgezet naar een zevenpuntsschaal met een gemiddelde van vier en een spreiding van een. Hiervoor wordt gebruik gemaakt van een WO-norm groep bestaande uit 450 kandidaten die bij LTP de TVRA in een selectiecontext hebben ingevuld. De scores op de Watson-Glaser zijn weergegeven op een schaal met een gemiddelde van vijf en een spreiding van twee (ook ‘stanines’ genoemd). Om de analytische tests te halen en door te gaan naar de lokale selectie dient een kandidaat een score van vier of hoger op de TVRA te halen en een score van

vijf of hoger op de Watson-Glaser. Ook kan een score van vier op de Watson-Glaser gecompenseerd worden met een score van vijf of hoger op de TVRA. In alle andere gevallen valt een kandidaat af.

De TVRA werd in 2021 door de Commissie Testaangelegenheden Nederland (COTAN) beoordeeld. De betrouwbaarheid van de TVRA werd als ‘voldoende’ beoordeeld. De constructvaliditeit werd als ‘onvoldoende’ beoordeeld omdat de veronderstelde unidimensionaliteit (de scores op de vragen kunnen het beste worden samengevat d.m.v. één testscore) niet is bevestigd en het onderzoek naar verbanden met scores op andere tests van onvoldoende kwaliteit was. Ook de criteriumvaliditeit werd als ‘onvoldoende’ beoordeeld door onvoldoende onderzoek. Analyses op basis van data uit het rio-volgsysteem lieten echter zien dat er een positieve en middelgrote samenhang was tussen scores op de TVRA test en de prestatie in de opleiding (r = .33). De verbanden tussen onderdelen uit de selectieprocedure en prestatie in de opleiding zijn weergegeven in Figuur 1.

Figuur 1 *Gecorrigeerde correlaties tussen onderdelen uit de selectieprocedure en de prestatie in de opleiding*



Noot. N = 81 – 148.

De Matrix test werd in 2020 net zoals de TVRA met 'voldoende' op betrouwbaarheid en 'onvoldoende' op constructvaliditeit en criteriumvaliditeit door de COTAN beoordeeld, gebaseerd op vergelijkbare argumenten. Analyses op basis van data uit het rio-volgsysteem lieten zien dat er een positieve en middelgrote samenhang was tussen scores op de Matrix test en de prestatie in de opleiding (r = .22). De Nederlandstalige Watson-Glaser test is niet door de COTAN beoordeeld. Onderzoek uit de Verenigde Staten laat zien dat de Watson-Glaser voldoende betrouwbaar is, de constructvaliditeit in orde is, en er middelgrote correlaties met werkprestaties gevonden werden (Watson & Glaser, 2010). Deze informatie is alleen niet zomaar te generaliseren naar de Nederlandse situatie en de Nederlandstalige versie van de test. In lijn met onderzoek uit de Verenigde Staten werd er op basis van data uit het rio-volgsysteem een positieve en middelgrote samenhang tussen scores op de Watson-Glaser test en prestatie in de opleiding gevonden (r = .24). De Matrix test en de Watson-Glaser test hadden echter geen toegevoegde waarde boven op de TVRA test. De voorspellende waarde wordt dus niet beter door naast de TVRA test ook nog de Matrix test en de Watson-Glaser test te gebruiken, en kan zelfs kleiner zijn dan de voorspellende waarde van de TVRA test op zichzelf als de drie tests suboptimaal worden gewogen. De resultaten over de criteriumvaliditeit van de analytische tests komen overeen met bevindingen uit de literatuur (Sackett et al., 2022, 2024) en met bevindingen uit de laatste evaluatie van de selectieprocedure (Niessen & Meijer, 2019). Samenvattend valt te zeggen dat vooral de TVRA test een redelijke voorspeller is voor prestatie in de opleiding, en dat de Watson-Glaser test en de Matrix test geen voorspellende waarde toevoegen.

De Numeriek Inzicht Test (NIT) dient in eerste instantie begripsvaardigheid (inductief redeneren) en in mindere mate ook leesvaardigheid en rekenvaardigheid te meten. De kandidaat dient aan de hand van tabellen en grafieken opgaven te beantwoorden. Omdat de NIT pas in januari 2024 aan de selectieprocedure toegevoegd werd om de voorspellende waarde te onderzoeken waren er nog geen scores van deze test beschikbaar in de data die door LTP werd aangeleverd waardoor een beoordeling van de kwaliteit van de NIT op basis van deze data niet mogelijk was. De NIT is nog niet beoordeeld door de COTAN. Verder ontbrak belangrijke informatie in de handleiding van LTP om de kwaliteit van de NIT te kunnen beoordelen. De gerapporteerde informatie impliceert echter dat de NIT onvoldoende betrouwbaar is. Er blijkt onvoldoende onderzoek verricht naar de kwaliteit van de NIT en de basis voor de in de handleiding gerapporteerde resultaten bleef soms onduidelijk of twijfelachtig. Zo vormde een afstudeerwerk uitgevoerd in 2005 de basis voor een deel van de analyses. Wij merken op dat het advies uit de laatste evaluatie (Niessen & Meijer, 2019), alleen tests te gebruiken die in voldoende mate zijn onderzocht op psychometrische kwaliteit (betrouwbaarheid en validiteit), dus niet consistent toegepast wordt. Omdat de NIT nog niet gebruikt wordt om selectiebeslissingen te nemen wordt deze test in de volgende analyses buiten beschouwing gelaten.

Analytische tests en diversiteit

Analyses op basis van de data uit het rio-volgsysteem lieten zien dat mannen en vrouwen zeer vergelijkbaar scoren op alle analytische tests. Verschillen op basis van geslacht waren dus verwaarloosbaar klein (d tussen 0.04 en 0.09). Zoals verwacht was er wel een middelgrote, negatieve samenhang tussen leeftijd en scores op de TVRA test (r = -.28), de Matrix test (r = -.38), en de Watson-Glaser test (r = -.24). Oudere kandidaten behaalden dus gemiddeld minder goede test scores dan jongere kandidaten. Zorgen omtrent mogelijke predictieve bias o.b.v. etnisch-culturele achtergrond zijn in het algemeen het grootst rondom tests voor cognitieve capaciteiten, omdat er op deze tests vaak grote en de grootste verschillen in scores worden gezien (Sackett et al., 2022). De algemene literatuur laat zien dat er weinig evidentie voor predictieve bias bij tests bestaat die

cognitieve capaciteiten dienen te meten (Berry, 2015; Dahlke & Sackett, 2022). Als het wordt gevonden, is het vaak in het voordeel van etnisch-culturele minderheidskandidaten; de prestaties van deze kandidaten worden enigszins overschat op basis van de testscores, niet onderschat. Een kanttekening is dat dit soort onderzoek vaak uitgaat van het gebruik van een enkele selectie-instrument en het gebruik van beslisregels. Dit is vaak niet representatief voor de praktijk. Verder is het meeste onderzoek gebaseerd op gegevens uit de V.S. Die resultaten zijn niet direct te generaliseren naar Nederland, maar het onderzoek uitgevoerd in Nederland duidt op vergelijkbare bevindingen (te Nijenhuis & van der Flier, 1999).

3.2.1 Ervaringen van gebruikers

In het algemeen waren betrokkenen van mening dat de analytische tests goede voorspellers zijn voor prestaties in de rio-opleiding. Tegelijkertijd bestonden er vermoedens dat kandidaten met een niet-Westerse achtergrond en oudere kandidaten gemiddeld lager scoren en de analytische tests daardoor een probleem voor de diversiteit zouden vormen. Sommige betrokkenen vonden het twijfelachtig of de analytische tests wel als een harde eis ingezet moeten worden en of de grensscores niet te streng zijn. Uit de gesprekken werd ook duidelijk dat voor de Matrix test en de op advies van LTP recent toegevoegde numerieke vaardighedentest geen harde grensscores gehanteerd worden, maar dat de scores uit de Matrix test wel in het assessment – holistisch – meegewogen worden. Verder hebben meerdere betrokkenen de wens genoemd de analytische tests voor de briefselectie te plaatsen, om zo te vermijden dat kandidaten die de briefselectie halen maar vervolgens op de analytische tests afvallen gezichtsverlies ervaren. Deze wijziging heeft al na de start van dit evaluatieonderzoek plaatsgevonden. Een betrokkene persoon nam aan dat de analytische tests makkelijk te oefenen zijn en vond de tests daarom ongeschikt. Als alternatief werd het maken van een juridische casus voorgesteld. Onderzoek liet inderdaad zien dat kandidaten iets hogere scores behalen als zij na de mogelijkheid om te oefenen een analytische test voor de tweede keer invulden, maar het effect was klein (Hausknecht et al., 2007; Lievens et al., 2012).

3.2.2 Ervaringen van oud-kandidaten

De analytische tests werden in het algemeen en op de vier procedurele rechtvaardigheidsprincipes positief beoordeeld. De tests werden ook gematigd positief beoordeeld ten opzichte van de andere instrumenten. In totaal hebben 113 respondenten een toelichting gegeven. Een veelvoorkomende opmerking was de indruk dat de tests makkelijk te oefenen zijn (n = 17). In het algemeen vond men de tests een effectieve methode (n = 14), maar werd, net zoals door gebruikers, een aangepaste volgorde van de tests en briefselectie gewenst (n = 16). Verder vonden sommige respondenten dat de grensscores van de tests geen harde eis zouden moeten zijn (n = 11) en dat de tests wellicht discriminerend zijn voor ouderen, kandidaten met een beperking zoals kleurenblindheid en kandidaten met een etnisch-culturele achtergrond (n = 10). Terwijl de Watson-Glaser test als nuttig werd ervaren (n = 10) twijfelden respondenten over het nut van de Matrix test (n = 12) en werd het als negatief ervaren dat het niet transparant was hoe de Matrix test, waarvoor geen harde grensscore wordt gehanteerd, werd gewogen (n = 5).

3.2.3 Aanbevelingen

Wij bevelen aan alleen de TVRA test, of tests die vergelijkbare constructen meten en van aantoonbaar goede psychometrische kwaliteit zijn, te handhaven omdat deze test betrouwbaar is en prestatie in de opleiding voorspelt. Wij raden aan de Watson-Glaser test en de Matrix test niet meer af te nemen omdat deze geen toegevoegde waarde boven op de TVRA test toonden. Wij merken op

dat in de vorige evaluatie aanbevolen werd de Watson-Glaser test *in plaats van* de Matrix test te gebruiken. Gezien de Matrix test nog steeds gebruikt wordt maar geen toegevoegde waarde boven op de TVRA test heeft, blijft de aanbeveling de Matrix test niet meer te gebruiken bestaan. De analyses uit deze evaluatie lieten echter ook zien dat de Watson-Glaser test, welke op zichzelf wel voorspellende waarde toonde (r = .24), ook geen toegevoegde waarde boven op de TVRA test toonde. De samenhang tussen scores op de Watson-Glaser test en scores op de TVRA test was zeer groot (r = .56). Onder deze omstandigheden zal het gebruik van zowel de Watson-Glaser test en de TVRA test *bij een optimale weging* in bijna dezelfde voorspellende waarde resulteren (r = .34) als de voorspellende waarde die enkel door gebruik van de TVRA test wordt bereikt (r =. 33) en zal bij een suboptimale weging zoals gelijke gewichten zelfs in minimaal kleinere voorspellende waarde resulteren (r = .32). De selectieprocedure kan dus door het weglaten van de Watson-Glaser test ingekort worden en wellicht attractiever voor kandidaten worden zonder voorspellende waarde te verliezen. Het handhaven van de Watson-Glaser test zal echter ook nauwelijks schadelijk zijn als de Watson-Glaser test niet meer gewicht krijgt dan de TVRA test.

Verder hingen de scores op de Matrix test het sterkste samen met leeftijd. Het is belangrijk te benadrukken dat het geen oplossing is om de Matrix test dan maar holistisch in het assessment mee te laten wegen. Hierdoor wordt het gebruik van de Matrix test slechts minder transparant. Ten slotte is er onvoldoende onderzoek gedaan naar de kwaliteit van de NIT test maar lijkt deze onvoldoende betrouwbaar te zijn. Daarom raden wij af de NIT in te zetten, in ieder geval totdat er evidentie kan worden aangeleverd waaruit blijkt dat de kwaliteit van de test op belangrijk aspecten in orde is.
Wij concluderen dat de analytische tests grotendeels aan de standaardisatiecomponenten in Tabel 1 voldoen, omdat iedere kandidaat dezelfde tests moet maken en de evaluatie en combinatie van de antwoorden gestandaardiseerd is. Alleen op die plaatsen waar geen harde grensscores bestaan en test scores holistisch in het assessment worden meegenomen ontbreekt standaardisatie. Dat raden we dus af.

Aanbeveling 2 – Alleen de TVRA test handhaven

- De TVRA test handhaven omdat deze prestatie in de opleiding voorspelt
- De Watson-Glaser test en de Matrix test niet meer gebruiken omdat deze geen voorspellende waarde toevoegen als de TVRA test al gebruikt wordt
- De NIT niet te gaan gebruiken omdat deze onvoldoende betrouwbaar lijkt te zijn

3.3 Lokale gesprekken

Kandidaten die de analytische tests gehaald hebben worden voor de lokale gesprekken uitgenodigd. Zij voeren gesprekken met de sollicitatie-adviescommissie en het bestuur van het wervende gerecht. Deze selecteurs zijn niet op systematische wijze geselecteerd. Een gerecht mag de commissie zelf samenstellen. Het bestuur ontvangt advies van de sollicitatie-adviescommissie om uiteindelijk een selectiebeslissing te nemen. Voorheen heeft de lokale selectie pas na het doorlopen van de landelijke selectieprocedure plaatsgevonden en was het doel de ‘fit’ met het gerecht te beoordelen, met de aanname dat de geschiktheid voor de rio-opleiding al door de LSR is vastgesteld. Inmiddels vindt de lokale selectie plaats voordat het assessment en de eindgesprekken worden afgenomen. Daardoor krijgt de lokale selectie meer invloed omdat deze gesprekken eerder in de selectieprocedure worden afgenomen. Het is daarom des te belangrijker dat de lokale gesprekken gestandaardiseerd zijn. Zoals hieronder beschreven zijn de lokale gesprekken echter vaak niet gestandaardiseerd en

worden zelden formeel gescoord. Daarom bevatte de data die door de LSR werd aangeleverd ook geen scores van lokale gesprekken, waardoor het niet mogelijk was de psychometrische kwaliteit van de lokale gesprekken en het effect op diversiteit te onderzoeken. Gezien het gebrek aan standaardisatie zal de betrouwbaarheid en validiteit van de lokale gesprekken echter waarschijnlijk laag zijn (Huffcutt et al., 2013, 2014; Sackett et al., 2022).

Referenties

Kandidaten dienen referenties van twee referenten aan te leveren, bij voorkeur uit de recente werksfeer. Een kandidaat kan tot twee dagen voor de lokale gesprekken de referenties aanleveren, omdat het de bedoeling is dat de referenties tijdens de lokale selectie worden bekeken. Het is echter ook mogelijk om de referenties al in de eerste stap van de selectieprocedure, tijdens het invullen van het sollicitatieformulier, aan te leveren. In sommige gevallen worden de referenties dan al als onderdeel van de briefselectie meegewogen. De referent wordt gevraagd om een aantal gestandaardiseerde vragen over de kandidaat te beantwoorden. Zo dienen de referenten te beoordelen in hoeverre de kandidaat beschikt over aspecten zoals analytisch vermogen, creativiteit, sociale vaardigheden en juridische kennis en ervaring, en in hoeverre de kandidaat integer is. Deze aspecten worden noch door de referenten noch door de LSR numeriek beoordeeld. Daarom kon de voorspellende waarde van de referenties niet onderzocht worden. In het algemeen is er weinig onderzoek verricht naar de voorspellende waarde van referenties, maar tonen bestaande studies aan dat de voorspellende waarde klein is (Hunter & Hunter, 1984; Mosel & Goheen, 1958; Reilly & Chao, 1982). Het bestaande wetenschappelijke onderzoek zal waarschijnlijk de voorspellende waarde van referenties en hoe deze in de praktijk gebruikt worden enigszins overschatten. Dit omdat de meeste referenties in een ‘vrijer format’ worden opgevraagd, waarbij de referent een tekst schrijft, maar er geen gestandaardiseerde, numerieke beoordelingen worden gegeven en niet over meerdere referenten heen wordt gemiddeld, wat in het bestaand wetenschappelijk onderzoek wel vaak het geval is. De voorspellende waarde van de referenties zoals die door LSR gebruikt worden zal dus waarschijnlijk klein zijn.

3.3.1 Ervaringen van gebruikers

Uit de gesprekken bleek dat er veel variatie bestaat in hoe gerechten de lokale selectie inrichten. Zo varieert het aantal gesprekken en het aantal beoordelaars dat een gesprek bijwoont. Meerdere betrokkenen gaven aan dat de meeste gesprekken niet gestandaardiseerd zijn. Er worden vooraf slechts thema’s bepaald, maar het zijn vaak niet dezelfde vragen in dezelfde volgorde en er zijn meestal geen formele beoordelingsformulieren waarop numeriek gescoord wordt. Eén betrokkene gaf aan dezelfde vragen aan verschillende kandidaten te stellen. Dat is wenselijk (zie Tabel 1). Er bestonden echter geen voorbeelden van minder goede en goede antwoorden. Verder bestond er variatie in wat men beoogde te beoordelen in de lokale gesprekken. Sommige gesprekken richtten zich meer op de ‘fit’ met het gerecht, wat ook de bedoeling van de lokale selectie is volgens de Raad voor de rechtspraak. Andere gesprekken waren echter ook gericht op competenties, waarvan sommige ook in de eindgesprekken beoordeeld worden. Verder blijft het onduidelijk hoe ‘fit’ met het gerecht en andere aspecten die getoetst worden, zoals diversiteit, gedefinieerd zijn. Zo bleek uit een beoordelingsformulier wat bij een gerecht gebruikt werd dat onder diversiteit onder andere aan kandidaten gevraagd wordt of zij bekend zijn met de ‘Insights’ of soortgelijke ‘kleurentests’ (tests die beogen persoonlijkheidstypes in plaats van dimensies te meten) en zo ja, welk type zij zijn. Wij merken op dat onderzoek aantoont dat dit soort kleurentests geen wetenschappelijke theoretische onderbouwing hebben en onvoldoende betrouwbaar en valide zijn (Ashton & Lee, 2009) en raden daarom af dit soort tests in te zetten en kandidaten naar hun resultaten op

kleurentests te vragen. Kortom, er bestaat onduidelijkheid over en variatie in wat 'fit' en diversiteit überhaupt betekenen. Het is daarom ook raadzaam hierop in te gaan in de trainingen die voor selecteurs aangeboden worden.

Ook noemden betrokkenen verschillen in de populariteit van de gerechten, waardoor er bij sommige gerechten veel kandidaten solliciteren terwijl er nauwelijks kandidaten bij andere gerechten solliciteren. Dit werd ook als reden aangegeven waarom sommige lokale gesprekken niet gestandaardiseerd zijn; men is meer gericht op de werving van kandidaten dan de selectie. In de meeste gevallen worden er al oriënterende gesprekken voorafgaand aan de briefselectie gevoerd. Wij erkennen dat dergelijke gespreken verschillende doelen, zoals werving en selectie, kunnen hebben (Wingate & Bourdage, 2024). Dit hoeft echter niet te betekenen dat deze doelen per se in conflict staan. Zo kan een gesprek bijvoorbeeld met het selecterende gedeelte beginnen waarin op een gestandaardiseerde manier zoals in Tabel 1 aanbevolen de geschiktheid voor de functie wordt bepaald. Pas daarna zou het gesprek gericht kunnen zijn op de werving en het uitleggen van de baan en andere aspecten die belangrijk zijn. Wel moet er door gebruik van beslisregels gewaarborgd worden dat het wervende gedeelte de eerdere beoordeling niet meer kan beïnvloeden. Ten slotte hadden betrokkenen vanuit de LSR de indruk dat het draagvlak voor het landelijke traject iets is afgenomen sinds de lokale selectie voor de eindgesprekken plaats vindt. De indruk is dat het nu moeilijker is om aan gerechten uit te leggen waarom een kandidaat die lokaal geschikt is bevonden landelijk geen groen licht krijgt.

3.3.2 Ervaringen van oud-kandidaten

Op basis van de kwantitatieve beoordelingen werden de lokale gesprekken in het algemeen en ook op de vier rechtvaardigheidsprincipes positief beoordeeld. Ook ten opzichte van de andere instrumenten waren de lokale gesprekken een van de meest positief beoordeelde selectiemethoden. Op basis van de toelichtingen van 87 respondenten was de beoordeling echter minder positief. Veelvoorkomende opmerkingen waren dat de lokale gesprekken geen relatie met de baan hadden, niet gestandaardiseerd waren, en vooral als doel hadden om te bepalen of er een 'klik' tussen de partijen bestond (n = 27). Een aantal respondenten merkten op dat de lokale selectie een positieve ervaring was (n = 13), vooral omdat de lokale gesprekken kennismaking met het gerecht en toekomstige collega's mogelijk maakt. Anderen gaven aan dat de lokale selectie een negatieve ervaring was (n = 7). Zo werd bijvoorbeeld de hoeveelheid gesprekpartners in de commissie en de niet gestandaardiseerde form van het gesprek als intimiderend en negatief ervaren. Verder was er een wens om de lokale selectie later plaats vinden te laten (n = 6). Opvallend is ook dat een aantal respondenten die bij meerdere gerechten gesolliciteerd hadden opmerkten dat de lokale selectie bij verschillende gerechten verschillend verliep qua aantal gesprekken, aantal gesprekpartners, standaardisatie, en uitkomst en er dus 'ruis' in de lokale selectie van rio's bestaat (n = 10). Ook vond men dat de planning van de lokale gesprekken suboptimaal verliep en veel flexibiliteit van de kandidaten vereiste (n = 4).

3.3.3 Aanbevelingen

Zoals ook bij de vorige evaluatie vast gesteld werd (Niessen & Meijer, 2019) blijken de lokale gesprekken nog steeds niet gestandaardiseerd te zijn en blijkt er veel variatie te bestaan in hoe verschillende gerechten de lokale selectie uitvoeren. Daarom bevelen wij aan de lokale gesprekken aan de hand van de standaardisatiecomponenten uit Tabel 1 te standaardiseren. Als het doel is de 'fit' met het gerecht te beoordelen dienen de gesprekken *per gerecht* gestandaardiseerd te zijn, en niet over gerechten heen, omdat bij het beoordelen van 'fit' met het gerecht juist de assumptie wordt

gemaakt dat een kandidaat wel bij het ene maar niet bij het andere gerecht zou passen. Het blijft echter de vraag in hoeverre deze assumptie überhaupt gerechtvaardigd is. Wij bevelen aan concreet te bepalen wat onder 'fit' met het gerecht wordt verstaan. Als, zoals nu het geval is, fit in een gesprek wordt gemeten dienen er dus vooraf voor elke vraag die aan elke kandidaat wordt gesteld goede en minder goede antwoorden geformuleerd te worden. Verder blijft het onduidelijk wat de reden is dat er twee gesprekken worden gevoerd, één met de sollicitatie adviescommissie en een met een of meerdere leden vanuit het bestuur van het gerecht. Zo werd bijvoorbeeld uit een beoordelingsformulier duidelijk dat beide gesprekken dezelfde competenties dienen te beoordelen. Wij raden daarom aan een gezamenlijk gesprek te voeren waarbij iedere selecteur de kandidaat onafhankelijk beoordeeld. Dan kan een beslisregel worden opgesteld die bepaald wanneer een kandidaat wel of niet doorgaat. Ook merken wij op dat momenteel de selecteurs bij de lokale selectie vooraf aan de gesprekken het sollicitatiedossier ontvangen, bestaande uit het sollicitatieformulier, bijlagen, de uitslagen uit de analytische tests en de referenties. Hierdoor wordt dus de standaardisatiecomponent dat aanvullende informatie gecontroleerd moet worden, geschonden (zie standaardisatiecomponent 4 in Tabel 1).

Het is ook de vraag of de scheiding tussen lokale en landelijke selectie te verantwoorden is. Uit de gesprekken met betrokkenen bleek dat er veel aandacht is voor de kwaliteit van de eindgesprekken. Deze vormen echter de laatste stap in de huidige selectieprocedure. Kandidaten zijn dus al geselecteerd op basis van de brieven, analytische tests, en lokale gesprekken, waardoor de invloed van de eindgesprekken minder wordt. Op dit moment zijn de lokale gesprekken niet gestandaardiseerd. Er zijn verschillende mogelijkheden om dit te verbeteren. Een eerste oplossing is de lokale gesprekken met hulp van LSR te standaardiseren en ervoor te zorgen dat de gesprekken ook daadwerkelijk op de bedoelde manier worden uitgevoerd. Een tweede optie is om de lokale gesprekken en eindgesprekken samen te voegen. Er kan dan een gestandaardiseerd gesprek gevoerd worden waarin zowel de geschiktheid voor de functie als ook de fit met het gerecht wordt beoordeeld. Dit zal de doorlooptijd reduceren, maar is praktisch gezien lastig als zowel selecteurs van de LSR als afgevaardigden van de lokale gerechten als beoordelaars daarbij gehandhaafd zouden worden. Het is echter de vraag of de standaardisatie van de gesprekken gewaarborgd kan worden als de lokale gerechten zonder hulp van de LSR de gesprekken voeren. Een derde optie is om de lokale gesprekken uit te besteden aan de LSR. De lokale gerechten zouden dan (een deel van) de vragen kunnen opstellen en de gedragsankers kiezen, op een manier die passend is voor het doel van het lokale gesprek.

Referenties

De referenties zullen in de huidige vorm waarschijnlijk hooguit een lage voorspellende waarde hebben en ze blijken aspecten te meten die al op een betere manier door andere selectie-instrumenten worden gemeten (zoals analytische vaardigheden door de analytische tests). Daarom kan overwogen worden het opvragen van de referenties te laten vervallen, wat wellicht de selectieprocedure ook aantrekkelijker kan maken. Mocht het opvragen van referenties gehandhaafd blijven, dan raden wij aan de referenten te vragen belangrijke aspecten zoals integriteit numeriek te scoren. Ook hier raden wij aan de standaardisatiecomponenten uit Tabel 1 toe te passen en geankerde beoordelingsschalen te gebruiken om te expliciteren wat onder minder integer en integer gedrag wordt verstaan. Verder zou een beslisregel opgesteld moeten worden om te bepalen hoe informatie uit de referenties gebruikt en gecombineerd wordt met informatie uit de brieven. Op dit moment worden referenties nog holistisch geïnterpreteerd. Wij raden af om aspecten via referenties te beoordelen die niet relevant voor de baan lijken, zoals creativiteit, of veel beter met andere al gebruikte instrumenten

kunnen worden gemeten, zoals analytische vaardigheden met de analytische tests. Mocht het gebruik van referenties gehandhaafd blijven raden wij aan de referenties voor iedere kandidaat in dezelfde fase van de selectieprocedure te gebruiken (briefselectie of lokale gesprekken) en kandidaten niet te laten kiezen wanneer de referenties ingediend worden.

Aanbeveling 3 – Lokale gesprekken standaardiseren

- De lokale gesprekken standaardiseren aan de hand van de componenten die in Tabel 1 genoemd zijn.
- Per gerecht bepalen wat 'fit' met het gerecht precies betekent.
- Beslisregels opstellen om te bepalen wanneer een kandidaat doorgaat.
- Duidelijker naar kandidaten communiceren wat zij van de lokale selectie kunnen verwachten.
- De referenties hebben waarschijnlijk hooguit een kleine voorspellende waarde. Daarom kan overwogen worden de referenties weg te laten. Mochten de referenties gehandhaafd blijven, raden wij aan de evaluatie door referenten te standaardiseren en beslisregels op te stellen om te bepalen hoe informatie uit de referenties gebruikt wordt. Wij raden dan ook aan de referenties voor iedere kandidaat in dezelfde fase te gebruiken (briefselectie *of* lokale gesprekken).

3.4 Assessment

Het assessment bestaat uit veel verschillende onderdelen. Kandidaten vullen een aantal zelf-rapportagevragenlijsten in, worden geïnterviewd door een LTP-adviseur en doen een rollenspel. Ook wordt een Dilemmatest afgenomen en moeten kandidaten een simulatie (een spel) voltooien. Het assessment resulteert in een assessmentrapport geschreven door een LTP-adviseur. Naast een kwalitatief beschrijvend gedeelte bevat het assessmentrapport een dichotoom advies (positief of negatief) en oordelen op verschillende competenties uit het rechtersprofiel die zijn uitgedrukt op een vijfpuntsschaal. Wij bespreken eerst de kwaliteit van de selectie-instrumenten die deel uit-maken van het assessment. Daarna bespreken wij de voorspellende waarde van de competentie-oordelen die door LTP-adviseurs worden gegeven.

3.4.1 Zelfrapportagevragenlijsten

Big Five Plus

De Big Five Plus is een zelfrapportagevragenlijst ontwikkeld door LTP die niet door de COTAN is beoordeeld. De handleiding geeft aan dat de theoretische onderbouwing van dit instrument gebaseerd is op het gangbare Big Five persoonlijkheidsmodel. Onderzoek heeft aangetoond dat er vijf algemene factoren bestaan waarmee iemands persoonlijkheid kan worden beschreven. Deze 'Big Five' zijn Extraversie, Vriendelijkheid, Consciëntieusheid, Openheid van Geest, en Neuroticisme (ook Emotionele stabiliteit genoemd). In de vragenlijst van LTP bestaan er naast de Big Five nog vijf losse 'plusschalen': Veranderingsgericht, Inschikkelijkheid, Materialisme, Trots, en Oprechtheid. De theoretische basis van deze plusschalen is niet duidelijk. Er wordt aangegeven dat de plus-schalen Materialisme, Trots, en Oprechtheid de factor 'integriteit' beogen te meten. Integriteit is een zesde factor die onderdeel is van het gangbaar HEXACO-persoonlijkheidsmodel. Door LTP aangeleverde analyses lieten zien dat sommige schalen die apart worden gescoord soms veel samenhang toonden. De plusschalen konden soms onvoldoende worden onderscheiden van de Big Five schalen. De constructvaliditeit blijkt dus enigszins twijfelachtig te zijn.

De betrouwbaarheid werd in de handleiding voor de Big Five factoren Extraversie, Vriendelijkheid, Consciëntieusheid, Openheid van Geest, Emotionele Stabiliteit, en iedere plusschaal weergegeven. De Big Five factoren waren voldoende betrouwbaar (α = .83-.94) terwijl de plusschalen voor individuele selectiebeslissingen grotendeels onvoldoende betrouwbaar waren (α = .71-.85). Er wordt vermeld dat een grote steekproef in een onderzoekscontext (low-stakes) is verzameld en gebruikt om de betrouwbaarheid en validiteit te bepalen. Dit is problematisch omdat resultaten over de betrouwbaarheid en validiteit uit een onderzoekscontext niet zomaar naar een selectiecontext generaliseerd kunnen worden. In een selectiecontext zijn kandidaten immers gemotiveerd om de baan te krijgen (Morgeson et al., 2007), wat invloed heeft op de antwoorden die ze geven (sociale wenselijkheid). Zoals in de handleiding vermeld waren de gemiddelde scores op de NEO-PI-R (een ander persoonlijkheidsvragenlijst die gebruikt werd om de Big Five Plus vragenlijst te valideren) dan ook veel hoger in de selectiecontext. Voor de Big Five Plus vragenlijst – het selectie-instrument wat bij de selectieprocedure van rio's gebruikt wordt - is dit niet onderzocht. Opvallend is dat er aparte normen voor vrouwen en mannen gehanteerd worden, met als reden dat LTP verschillen in criteriumvaliditeit van de schalen voor mannen en vrouwen verwachtte. Het is onduidelijk waarom dit verwacht wordt. Het gevolg is dat de normscores geïnterpreteerd moeten worden als; hoe hoog scoort deze persoon op deze schaal, vergeleken met andere vrouwen/andere mannen. Er is in die zin dus geen sprake van gelijke behandeling, zonder duidelijke onderbouwing. Uit analyses op basis van data uit het rio-volgsysteem bleek dat verschillen tussen mannen en vrouwen verwaarloosbaar klein waren. Alleen op Extraversie scoorden mannen gemiddeld lager dan vrouwen (*d* = -0.51). Ook waren verbanden tussen de scores op de Big Five Plus schalen en leeftijd verwaarloosbaar klein. Zoals Figuur 1 laat zien toonde van de Big Five alleen Consciëntieusheid een kleine positieve correlatie met de prestatie in de opleiding (*r* = .16). De voorspellende waarde van Consciëntieusheid is op het randje om praktisch relevant te zijn. De voorspellende waarde van Emotionele Stabiliteit, Vriendelijkheid, Extraversie, en Openheid van Geest was dus verwaarloosbaar klein. Ook dit resultaat komt redelijk overeen met de bevindingen uit de literatuur (Sackett et al., 2022). De correlaties tussen de plusschalen en de prestatie in de opleiding waren ook verwaarloosbaar klein. Onder de aanname dat Consciëntieusheid en de TVRA test optimaal compensatorisch zouden worden gewogen (66% TVRA test, 34% Consciëntieusheid) zal de voorspellende waarde van beide selectie-instrumenten samengenomen *r* = .37 zijn, wat dus iets beter is dan het gebruik van enkel de TVRA test (*r* = .33). Als de TVRA test en Consciëntieusheid even zwaar zouden worden draagt het toevoegen van Consciëntieusheid nauwelijks bij aan de voorspellende waarde (*r* = .35). Voor de diversiteit zal het toevoegen van Consciëntieusheid een klein beetje helpen, omdat de verschillen in scores tussen kandidaten met en zonder een etnisch-culture achtergrond in het algemeen op dit selectie-instrument verwaarloosbaar klein zijn, en kleiner dan voor cognitieve tests (Sackett et al., 2022; Sackett & Ellingson, 1997). Verder kan ook overwogen worden om een vragenlijst te gebruiken waarin de items gecontextualiseerd zijn (d.w.z., waarin de items in de context van werk geformuleerd zijn), omdat gecontextualiseerde vragenlijsten hogere voorspellende waarde hebben dan niet gecontextualiseerde vragenlijsten (Shaffer & Postlethwaite, 2012). De Big Five Plus is niet gecontextualiseerd. Wij concluderen dat alleen Consciëntieusheid van de Big Five Plus voldoende voorspellende waarde heeft voor prestatie in de opleiding is en de rest van de Big Five Plus schalen dus weggelaten kan worden. Wij merken echter ook opnieuw op dat zelfrapportagevragenlijsten gemakkelijk sociaal-wenselijk in te vullen zijn en dit de voorspellende waarde negatief kan beïnvloeden (Anglim et al., 2018; Niessen & Meijer, 2019).

Work Personality Questionnaire

De Work Personality Questionnaire (WPQ) is een zelfrapportagevragenlijst ontwikkeld door LTP en beoogt smalle constructen te meten. De WPQ bestaat uit zeven schalen: Extrinsieke prestatiemotivatie, Intrinsieke prestatiemotivatie, Ontwikkelgericht, Negatieve faalangst, Positieve faalangst, Doortastend, en Risico nemen. De theoretische onderbouwing van de WPQ ontbreekt. Het blijft onduidelijk waarom deze constructen überhaupt belangrijk zouden zijn om werkprestaties te voorspellen. De WPQ is in 2020 door de COTAN beoordeeld. De betrouwbaarheid en constructvaliditeit werden als 'voldoende' beoordeeld. Verschillen tussen mannen en vrouwen en verbanden tussen de scores en leeftijd waren klein. De criteriumvaliditeit werd als 'onvoldoende' beoordeeld door de COTAN vanwege gebrek aan onderzoek waarin een prestatiemaat als uitkomst gehanteerd werd. De correlaties tussen de scores van de WPQ schalen en de prestatie in de opleiding waren verwaarloosbaar klein (*r*s <= |.14|). De WPQ heeft dus nauwelijks voorspellende waarde en kan daarom in zijn geheel weggelaten worden.

Drijfverenvragenlijst

De drijfverenvragenlijst is een zelfrapportagevragenlijst ontwikkeld door LTP. Dit instrument is niet beoordeeld door de COTAN. Het is voornamelijk gebaseerd op de Zelfdeterminatietheorie, welke veronderstelt dat mensen proberen hun drie basisbehoeftes aan autonomie, competentie, en affiliatie te bevredigen. In de handleiding wordt vermeld dat de behoefte aan competentie geen deel uitmaakt van dit instrument. Daar wordt geen reden voor genoemd. Naast behoefte aan autonomie en affiliatie bestaat dit instrument nog uit twee extra schalen: 'behoefte aan macht' en 'prestatiemotivatie'. De theoretische onderbouwing waarom deze aspecten werkprestatie zouden voorspellen ontbreekt grotendeels. De veronderstelde factorstructuur werd in onderzoek uitgevoerd door LTP in onvoldoende mate gevonden. De constructvaliditeit blijkt dus onvoldoende te zijn. De betrouwbaarheid van de schalen was voldoende (α = .80-.86) behalve van de schaal 'behoefte aan autonomie' (α = .76). Verschillen tussen mannen en vrouwen en verbanden tussen de scores en leeftijd waren klein. De verbanden tussen de scores op de schalen van de Drijfverenvragenlijst en prestatie in de opleiding waren verwaarloosbaar klein (*r*s <= |.07|). Dit instrument levert dus geen bijdrage aan het voorspellen van prestaties in de rio-opleiding en kan daarom weggelaten worden.

Waardenvragenlijst

De Waardenvragenlijst (WV) is een zelfrapportagevragenlijst ontwikkeld door LTP. Dit instrument is niet beoordeeld door de COTAN. De WV beoogt iemands persoonlijke en interpersoonlijke waarden in kaart te brengen en bestaat uit zeven constructen: Steun en Erkenning, Conformiteit, Onafhankelijkheid, Altruïsme, Leiderschap, Ordelijkheid en Variëteit. De theoretische onderbouwing is onduidelijk. De betrouwbaarheid van de schalen was voldoende (α = .80-.90). De constructvaliditeit werd onderzocht door de relaties tussen de WV en de Big Five Plus, de Werkstijlenvragenlijst, de Beïnvloedingstijlenvragenlijst, en de Drijfverenvragenlijst te berekenen. De constructvaliditeit kon echter niet beoordeeld worden omdat er geen duidelijke verwachtingen geformuleerd werden, waardoor het onduidelijk bleef welke relaties evidentie voor de constructvaliditeit vormen. Verschillen tussen mannen en vrouwen en verbanden tussen de scores op de schalen en leeftijd waren zeer klein. De correlaties tussen de scores op de schalen van de Waardenvragenlijst en prestatie in de opleiding waren verwaarloosbaar klein (*r*s <= |.07|). Ook dit instrument heeft dus geen voorspellende waarde en kan daarom weggelaten worden.

Beïnvloedingsstijlen

De Beïnvloedingsstijlenvragenlijst is gebaseerd op het concept van Machiavellisme en ontwikkeld door LTP. Dit instrument is niet beoordeeld door de COTAN. Volgens LTP is het doel van dit instrument in kaart te brengen of een kandidaat het gedrag van anderen met succes kan beïnvloeden en 'politiek inzicht' heeft. De theoretische onderbouwing waarom dit instrument prestatie in de opleiding zou moeten voorspellen is onduidelijk. De resultaten van een factoranalyse leidden tot de vier schalen Manipuleren, Berekenend, Rechtdoorzee en Assertiviteit. Het blijft onduidelijk in hoeverre dit de verwachting was en wat de theoretische basis hiervoor is. De samenhang tussen scores op de schalen was klein tot middelgroot. Op de schaal 'Manipuleren' na bleek de betrouwbaarheid van de schalen voor selectiedoeleinden onvoldoende te zijn (αs .75-.77). De verbanden tussen de vier scores op de schalen en geslacht en leeftijd waren klein tot middelgroot. Analyses van LTP lieten zien dat verschillen tussen mannen en vrouwen verwaarloosbaar klein waren. Dit werd bevestigd door analyses op basis van de data uit het rio-volgsysteem. Alleen op de schaal Rechtdoorzee scoorden mannen aanzienlijk lager dan vrouwen (d = -0.62). Correlaties tussen de scores en leeftijd waren verwaarloosbaar klein. Van de Beïnvloedingsstijlen toonden alleen de scores op de schaal Berekenend een kleine positieve verband met de prestatie in de opleiding (r = .16). De scores op deze schaal hingen echter ook samen met scores op de schaal Consciëntieusheid uit de Big Five Plus (r = .31). Wij raden aan de Beïnvloedingsvragenlijst niet meer te gebruiken omdat de scores niet betrouwbaar genoeg zijn en de enige schaal met enkele voorspellende waarde overlapt met Consciëntieusheid, wat al met de Big Five Plus vragenlijst wordt gemeten.

Persoonlijke Veranderkracht (VQ)

Op basis van de persoonlijkheidsvragenlijsten heeft LTP scores samengesteld om 'Persoonlijke Veranderkracht' (VQ) uit te drukken. De theoretische onderbouwing van VQ ontbreekt. Ook bestaat er geen onderzoek naar de kwaliteit van de VQ en ontbreekt een handleiding. De VQ is niet door de COTAN beoordeeld. Het blijft onduidelijk hoe scores uit de persoonlijkheidsvragenlijsten samengevoegd worden. VQ bestaat uit de vier schalen 'agility', 'energy', 'growth mindset', en 'self management'. Deze schalen beogen zelfvertrouwen, drive, openheid voor ontwikkeling, en het sturen van je eigen ontwikkeling te meten. Analyses op basis van de data uit het rio-volgsysteem lieten zien dat verschillen tussen mannen en vrouwen op de scores van de vier schalen klein waren en er kleine verbanden met leeftijd waren. De scores op de VQ schalen toonden verwaarloosbaar kleine correlaties met de prestatie in de opleiding (rs <= |.14|). De VQ is overbodig omdat deze scores geen voorspellende waarde hebben en een samenstelling zijn op basis van de persoonlijkheidsvragenlijsten die ook al op een ander wijze in het assessment worden meegenomen. Er wordt daardoor twee keer dezelfde informatie gebruikt. De VQ kan daarom weggelaten worden.

Teamrollen

Verder heeft LTP scores op basis van de persoonlijkheidsvragenlijsten scores samengesteld die resulteren in de teamrollen 'Aanpakker', 'Expert', 'Monitor', 'Coordinator', 'Verbinder', 'Motivator', 'Innovator', en 'Ondernemer'. Net zoals voor VQ blijft het onduidelijk welke scores op welke manier samengevoegd worden. Onderzoek naar de kwaliteit van de teamrollen en een handleiding ontbreken. Analyses op basis van de data uit het rio-volgsysteem lieten zien dat verschillen tussen mannen en vrouwen per team rol klein waren. Alleen op de rol Motivator scoorden mannen gemiddeld lager dan vrouwen (d = -0.48). De verbanden tussen de teamrollen en leeftijd waren verwaarloosbaar klein. De correlaties tussen scores op de Teamrollen en prestatie in de opleiding waren verwaarloosbaar klein (rs <= |.14|). De Teamrollen hebben dus geen voorspellende waarde en zijn overbodig omdat

zij op basis van andere persoonlijkheidsvragenlijsten tot stand komen en er dus dezelfde informatie twee keer wordt gebruikt. Wij raden aan de Teamrollen niet meer te gebruiken.

3.4.2 Everest Games

De 'Everest Games' (EG) is een digitaal spel/simulatie-opdracht en is niet beoordeeld door de COTAN. Kandidaten moeten in dit spel de meest geschikte locatie voor de toekomstige Everest games kiezen, op basis van veel relevante en minder relevante informatie. Het doel is om inzicht te krijgen in het vermogen van kandidaten om met veel informatie om te gaan, door een realistische werkomgeving te simuleren waarin een kandidaat door collega's wordt onderbroken. De theoretische basis voor de EG blijft onduidelijk. In de handleiding wordt vermeld dat dit spel het werk- en denkniveau op een andere wijze in kaart brengt dan traditionele capaciteitentests, dat de acceptatie van dit instrument onder kandidaten groot is en men het leuk vindt om te doen, omdat het als herkenbaar voor de eigen werksituatie wordt gezien.

Op basis van de handleiding werd niet duidelijk hoe de EG precies wordt gescoord. De dataset bevat de variabelen 'timemanagement', 'probleemanalyse', 'prioriteitstelling', 'evaluatiemethode', 'overallCaseScore', en 'overallSorteerScore'. Uit de handleiding wordt niet voldoende duidelijk wat deze variabelen precies beogen te meten. Er wordt vermeld dat de 'overall scores' samengesteld zijn uit enkele schalen, maar de details ontbreken. Er werd geen informatie over de betrouwbaarheid vermeld. Informatie over de validiteit van de EG is amper aanwezig. Er wordt vermeld dat het gemiddelde van de twee 'overall scores' sterk samenhangt met scores op een algemene cognitieve capaciteitentest (r = .60). Er wordt niet vermeld welke capaciteitentest hiervoor gebruikt werd en wat de steekproef was. Op basis hiervan lijkt er wel behoorlijke overlap te bestaan tussen de EG en de analytische tests. In de handleiding wordt dan ook vermeld dat dit instrument *in plaats van* een live-simulatie of intelligentiebatterij ingezet dient te worden. In de huidige selectieprocedure voor de rio-opleiding is de EG daarmee overbodig.

Ten slotte wordt vermeld dat scores op de EG negatief samenhangen met leeftijd. Hoe sterk dit verband was wordt niet vermeld. Ook bevatte de handleiding geen informatie over onderzoek naar predictieve bias. Analyses op basis van de data uit het rio-volgsysteem lieten zien dat er een middelgroot negatief verband bestond tussen de overallCaseScore en leeftijd (r = -.29) en een groot negatief verband tussen de overallSorteerScore en leeftijd (r = -.46). Verschillen tussen mannen en vrouwen waren verwaarloosbaar klein. De schaal Prioriteitstelling liet een middelgrote positieve correlatie met de prestatie in de opleiding zien (r = .25). Verbanden tussen de rest van de scores en de prestatie in de opleiding waren grotendeels verwaarloosbaar klein. De EG blijkt dus sterk samen te hangen met scores op cognitieve capaciteitentests, laat sterkere negatieve verbanden zien met leeftijd dan de analytische tests en heeft tegelijkertijd minder voorspellende waarde dan de analytische tests. Daarom raden wij aan de EG niet meer te gebruiken.

3.4.3 Dilemmatest

De Dilemmatest is een situational judgment test ontwikkeld door LTP, waarin kandidaten verschillende ethische dilemma's voorgelegd worden en waarbij zij moeten aangeven hoe zij in de situatie zouden reageren. Dit instrument is gebaseerd op Kohlberg's model van morele ontwikkeling. De inzet van dit model in deze context krijgt kritiek, onder andere omdat onderzoek naar dit model grotendeels gebaseerd is op kinderen en jongeren (Snarey, 1985). De Dilemmatest beoogt te meten in hoeverre een persoon op zichzelf ('egoDriven') of op regels ('ruleDriven') is gericht en in hoeverre

een persoon nadenkt over de gedachten achter de wetten ('valueDriven'). De vierde schaal beoogt de mate van strengheid of ethiek van de keuzes weer te geven ('stringency'). Uit documenten van LTP blijkt dat de Dilemmatest voornamelijk als praatstuk dient voor de adviseur tijdens het gesprek. De scores worden niet expliciet gebruikt in de beoordeling. Wat dit precies betekent en hoe de Dilemmatest wel of niet meegewogen wordt blijft onduidelijk en is dus niet transparant. Het blijft ook onduidelijk in hoeverre hoge scores op deze schalen wel of niet wenselijk zijn. De Dilemmatest is niet door de COTAN beoordeeld. Onderzoek naar de kwaliteit van dit instrument ontbreekt. Er was te weinig data beschikbaar om de criteriumvaliditeit met enige zekerheid te bepalen (n = 58), vermoedelijk omdat de Dilemmatest niet bij iedereen werd afgenomen. De resultaten suggereerden echter dat de verbanden tussen de scores van de Dilemmatest en prestatie in de opleiding verwaarloosbaar klein zijn. Wij raden af de Dilemmatest te gebruiken omdat informatie over de kwaliteit van dit instrument volledig ontbreekt en het instrument niet expliciet wordt meegewogen.

3.4.4 Assessmentgesprek en Rollenspel

Er werden door LTP geen scores van het gesprek en het rollenspel aangeleverd, waardoor het niet mogelijk was om de betrouwbaarheid en validiteit van deze instrumenten te onderzoeken. Uit gesprekken met LTP bleek echter ook dat de gesprekken niet gescoord worden. De evaluatiestandaardisatie ontbreekt dus bij dit selectie-instrument. Behalve voor het beoordelen van integriteit bestaat er geen vooraf opgestelde lijst met vragen die aan iedere kandidaat gesteld worden. Alleen de competenties die beoordeeld moeten worden zijn gestandaardiseerd. De standaardisatie van vragen ontbreekt dus ook, waardoor het gevaar bestaat dat ruis en bias geïntroduceerd worden als verschillende adviseurs verschillende vragen aan kandidaten stellen. Omdat de antwoorden op de vragen niet numeriek gescoord worden komt een oordeel op basis van het gesprek dus ook per definitie holistisch tot stand. Kortom, de standaardisatie ontbreekt voor het assessmentgesprek.

Tijdens het rollenspel wordt de kandidaat uitgedaagd zijn of haar competenties in een realistische werkomgeving te tonen. De casussen variëren tussen kandidaten om te voorkomen dat de inhoud van het rollenspel vooraf bekend is. Wat kandidaten precies in de casussen moeten doen bleef onduidelijk. Volgens LTP worden getrainde acteurs ingezet die gebruik maken van gedetailleerde instructies om consistentie te waarborgen. Zowel de adviseur als ook acteur beoordelen de kandidaat op een vijfpuntsschaal op de dimensies 'Denken', 'Voelen', 'Kracht' en 'Effectiviteit'. Dit is een model wat LTP intern gebruikt. Volgens LTP vatten deze dimensies de competenties uit het rechtersprofiel samen. De competenties uit het rechtersprofiel worden aan deze competenties gekoppeld. Hoe die koppeling eruitziet en waarom is echter onduidelijk. De theoretische onderbouwing voor dit model en onderzoek naar de validiteit ervan ontbreekt. Er worden geen gestandaardiseerde beoordelingsformulieren met gedragsankers gebruikt. De evaluatiestandaardisatie is dus laag (zie Tabel 1). De scores die door de adviseur en acteur worden gegeven ontbraken in het bestand wat van LTP werd ontvangen waardoor het niet mogelijk was de kwaliteit van het rollenspel te onderzoeken. Er wordt ook expliciet aangegeven dat de inzichten uit het rollenspel gebruikt worden als input voor het assessmentgesprek. Hierdoor wordt echter de onafhankelijkheid van de beoordelingen uit verschillende instrumenten geschonden (zie standaardisatiecomponent 4 in Tabel 1). Kortom, de standaardisatie van het rollenspel blijkt op het gebruik van casussen en instructies na laag te zijn, en het blijft onduidelijk hoe de scores uit het rollenspel gebruikt worden.

3.4.5 Competentie-oordelen

Het dichotoom selectieadvies (positief of negatief) en de competentie-oordelen op een vijfpunts-schaal worden gegeven door LTP-adviseurs op basis van al het bovenstaande en komen holistisch tot stand. Het blijft dus onduidelijk hoe de instrumenten precies gebruikt worden om tot competentie-oordelen en het selectieadvies te komen. Een mechanische combinatie van selectie-instrumenten is per definitie niet mogelijk omdat sommige indrukken uit de selectie-instrumenten (bijvoorbeeld het assessmentgesprek) niet gekwantificeerd zijn. Er werd aangegeven dat de Matrix test in het assessment wordt meegewogen maar het is niet transparant hoe dat gebeurt. Voor kandidaten waarvan scores over de prestatie in de opleiding beschikbaar waren was het advies in 136 van 145 gevallen positief. Gezien het te kleine aantal negatieve adviezen is er geen correlatie tussen het advies en de prestatie in de opleiding berekend. Wel zijn correlaties tussen de competentie-oordelen en de prestatie in de opleiding berekend (zie Figuur 1). De correlaties waren verwaarloosbaar klein. Alleen de correlatie tussen oordelen op de competentie 'Analytisch vermogen en oordeelsvorming' en de prestatie in de opleiding was iets groter, maar nog steeds klein (r = .22). Dit komt waarschijnlijk omdat LTP-adviseurs voor deze beoordeling vooral op de scores op de analytische tests letten. Het is echter opvallend dat de voorspellende waarde van dit competentie-oordeel minder was dan de voorspellende waarde van de meest voorspellende analytische test op zichzelf genomen (de TVRA test, r = .33). In lijn met bevindingen uit de literatuur bleken de holistische competentie-oordelen ook geen toegevoegde waarde boven op de scores uit de analytische tests te hebben[1] (Hoffman et al., 2018; Kuncel et al., 2013; Neumann et al., 2024; Yu & Kuncel, 2020, 2022), zelf terwijl deze holistische oordelen op veel meer informatie zijn gebaseerd. Sterker nog, aangezien de holistische competentie-oordelen en de scores op de analytische tests niet op een optimale manier gecombineerd worden (zoals wel het geval is in de gerapporteerde analyse in voetnoot 1), is het waarschijnlijk dat de holistische competentie-oordelen de valide informatie uit de analytische tests 'verwateren' en zo de validiteit van selectiebeslissingen zelfs verminderen (Dana et al., 2013; Kausel et al., 2016). Een robuuste bevinding bij individuele assessments is dat het holistische oordeel van een expert vaak *minder* valide is dan het consistente wegen van informatie volgens een beslisregel (Morris et al., 2015). Kortom, de competentie-oordelen op basis van het assessment hebben geen toevoegde waarde in de selectieprocedure, omdat deze prestatie in de opleiding nauwelijks voorspellen.

3.4.6 Ervaringen van gebruikers

Het assessment werd door betrokkenen met grote overeenstemming als positief en nuttig ervaren. Betrokkenen vonden het een goede aanvulling op de analytische tests en lokale gesprekken. Ook werd het gewaardeerd dat het assessment rapport inspiratie voor de eindgesprekken biedt[2]. Veel betrokkenen merkten echter op dat het vaak niet transparant is hoe het assessmentadvies in het rapport tot stand komt. De meerderheid van de betrokkenen ervoeren de niet-transparante aard van het advies negatief. Een minderheid gaf aan dat zij dit als minder erg ervoeren omdat zij vertrouwen op het oordeel van een deskundige adviseur en ervan uitgaan dat 'best practices' gevolgd worden. Verder werd duidelijk dat selecteurs die de eindgesprekken voeren ook regelmatig van het advies

---

1 Een vergelijking tussen een model met alleen de analytische tests als voorspellers en een model met de holistische competentie-oordelen erbij leverde geen evidentie op dat de competentie-oordelen significant meer variantie in de prestatie in de opleiding boven op de analytische tests verklaarden ($F(8, 133) = 0.59$, p = .78, ΔR^2 = .03, $\Delta R^2_{adj.}$ = -0.02).

2 Wij merken echter op dat hierdoor de standaardisatie van de eindgesprekken wordt geschonden omdat de eindgesprekken niet meer onafhankelijk van de informatie verzameld via andere selectie-instrumenten beoordeeld worden (zie component 4 in Tabel 1).

uit het assessment rapport afwijken. Dit gebeurt voornamelijk bij selecteurs die benadrukten dat zij het assessment als een momentopname ervaren of zwakke of sterke punten uit het rapport juist andersom in hun eigen gesprekken hadden ervaren. Dit is problematisch omdat het advies dan verschillend wordt gewogen voor verschillende kandidaten. Betrokkenen gaven ook aan dat een LTP-adviseur altijd beschikbaar is om de resultaten uit het assessmentrapport toe te lichten.

3.4.7 Ervaringen van oud-kandidaten

Zelfrapportagevragenlijsten

De zelfrapportagevragenlijsten werden gematigd beoordeeld en gemiddeld ten opzichte van andere selectie-instrumenten, zowel in het algemeen als ook voor iedere van de vier rechtvaardigheids-principes. Onder de 74 respondenten die een toelichting hebben gegeven was met afstand de meest voorkomende reactie dat de persoonlijkheidslijsten makkelijk sociaal wenselijk in te vullen zijn (*n* = 21). Andere respondenten hebben de vragenlijsten als een effectief instrument ervaren (*n* = 10) of vonden dat de vragenlijsten alleen ter voorbereiding op het gesprek zouden moeten dienen, waar de psycholoog vragen gebaseerd op de vragenlijsten gaat stellen (*n* = 5). Sommigen gaven ook aan dat dit op deze manier gebeurt. Wij merken opnieuw aan dat daardoor standaardisatie-component 4 uit Tabel 1 zou worden geschonden.

Dilemmatest

De Dilemmatest werd op een vergelijkbare manier beoordeeld als de zelfrappportagevragenlijsten. Ook de toelichtingen die door 78 respondenten werden gegeven waren vergelijkbaar. Respondenten vonden dat het erg makkelijk is om de Dilemmatest sociaal wenselijk in te vullen (*n* = 29). Anderen vonden de Dilemmatest juist geschikt voor het selecteren van rio's en benadrukten de relatie met de baan (*n* = 7). Sommige respondenten vonden de Dilemmatest echter niet effectief (*n* = 4), misten een optie voor toelichting (*n* = 5) of vonden dat het niet transparant was hoe de Dilemmatest werd gewogen (*n* = 3).

Everest Games

De Everest Games werden in het algemeen als gematigd positief beoordeeld, maar in vergelijking met de ander selectie-instrumenten negatiever beoordeeld. Vooral de relatie met de baan en de mate waarin het een logische selectiemethode leek werd negatief geëvalueerd. Dit kwam overeen met de toelichtingen die door 105 respondenten werden gegeven. Veel respondenten (*n* = 30) gaven aan dat de verwachtingen en de relevantie van de game voor het functioneren als rechter onduide-lijk waren. Men vroeg zich ook af wat de game zou meten. Ook merkten veel respondenten op dat zij geen feedback over de uitkomst kregen en dat het niet transparant was hoe de uitkomst van de game mee werden gewogen in het advies (*n* = 30). Sommige respondenten vonden de game geschikt en hebben het als een leuke activiteit ervaren (*n* = 13). Anderen vonden het juist niet leuk of effectief (*n* = 12). Ook uitten sommige respondenten bedenkingen over verschillen in scores gebaseerd op leeftijd en ervaringen met games, maar ook beperkingen zoals kleurenblindheid (*n* = 8).

Assessmentgesprek

Het assessmentgesprek werd in het algemeen als gematigd tot positief ervaren. Ten opzichte van de andere selectie-instrumenten werd het als positief ervaren. Dit geldt ook voor de vier rechtvaardig-heidsprincipes. In totaal hebben 74 respondenten hun antwoorden toegelicht. Daarvan gaven een paar respondenten aan dat de uitkomst van het gesprek van de 'klik' met de psycholoog afhangt (*n* = 13) en dat er ongepaste of invasieve vragen gesteld werden (*n* = 11). Anderen merkten op dat het een

prettige ervaring was (*n* = 10) en zij het gesprek nuttig of effectief vonden (*n* = 5), maar ook dat de uitkomst van de specifieke psycholoog afhangt (*n* = 6). Dit was de reden waarom sommige respondenten meer dan één beoordelaar wensten (*n* = 4). Andere opmerkingen gingen erover dat de psycholoog vroegtijdig conclusies trok, niet goed voorbereid was of geen goed beeld van de functie van rechter had.

Rollenspel

Het rollenspel werd net als de game als gematigd positief, maar in relatie tot de ander selectie-instrumenten als negatief beoordeeld. Ook de beoordelingen op de vier rechtvaardigheidsprincipes waren negatief in vergelijking met de andere instrumenten. In totaal hebben 99 respondenten een toelichting gegeven. Een groot aantal respondenten (*n* = 30) merkte op dat de het rollenspel en de gebruikte casus geen relatie met de functie als rechter had. Veel respondenten (*n* = 29) vonden het rollenspel vooral onprettig omdat het online plaatsvond, wat grotendeels door de COVID pandemie zo gedaan is. Verder werd het rollenspel als kunstmatig ervaren (*n* = 16) en vond men het een momentopname (*n* = 10). Andere respondenten merkten op dat de tijd te kort en de verwachtingen onduidelijk waren. Ook waren sommige respondenten het oneens met de door de psycholoog veronderstelde juiste antwoorden en vond men de conclusies die getrokken werden op basis van het rollenspel te stellig. Sommige respondenten vonden het oneerlijk dat er verschillende casussen gebruikt werden voor verschillende rio's, waar sommige casussen meer gerelateerd waren aan de functie van rechter dan anderen.

3.4.8 Aanbevelingen

We raden aan om het assessment te laten vervallen en alleen de schaal Consciëntieusheid uit de Big Five Plus vragenlijst te gebruiken. De Big Five Plus helemaal weglaten zal echter ook een optie zijn. De voorspellende waarde zal daardoor alleen minimaal kleiner worden. Omdat de Work Personality Questionnaire, en de Drijfveren- en Waardenvragenlijst geen voorspellende waarde toonden raden wij aan deze niet meer in de selectieprocedure op te nemen. Ook raden wij aan de Beïnvloedingsstijlenvragenlijst en de *plusschalen* van de Big Five Plus niet meer mee op te nemen omdat deze onvoldoende betrouwbaar waren. Verder raden wij aan de Dilemmatest niet meer mee op te nemen omdat deze op dit moment alleen maar als *praatstuk* wordt gebruikt in het assessment-gesprek. Dit betekent dus dat scores op de Dilemmatest holistisch gebruikt worden en de onafhankelijkheid tussen selectie-instrumenten wordt geschonden (zie Tabel 1). Zelfs als de Dilemmatest mechanisch, volgens een beslisregel zou gebruikt worden raden wij af dit instrument te gebruiken omdat onderzoek naar de kwaliteit van dit instrument volledig ontbreekt. De dilemma's die in de huidige Dilemmatest worden gebruikt zijn niet gericht op de taken die rio's moeten uitvoeren in hun baan (standaardisatiecomponent 1 uit Tabel 1 ontbreekt). Een alternatief zou kunnen zijn een situational judgment test (SJT) te gebruiken die aan de psychometrische eisen voldoet en ook daadwerkelijk de gewenste aspecten zoals integriteit in de rechtspraak meet. Lievens en Patterson (2011) lieten zien dat een SJT voorspellende waarde toonde in een 'high-stakes' context waar de test door sollicitanten werd ingevuld. Wij merken echter op dat het ontwikkelen van een goede SJT veel tijd en geld zou kosten. Daarom kan alternatief overwogen worden integriteit in de eindgesprekken te meten, door scenario's uit de rechtspraak te schetsen en kandidaten te vragen hoe zij in het verleden hebben gehandeld of in de toekomst zouden handelen (Hollwitz & Pawlowski, 1997).
De antwoorden kunnen dan met vooraf opgestelde goede en minder goede antwoorden vergeleken en gescoord worden. Wij raden af zelf-rapportagevragenlijsten te gebruiken om integriteit te meten, aangezien de voorspellende waarde in een high-stakes context zeer klein is (Van Iddekinge et al., 2012).

Ongeacht het selectie-instrument dat gebruikt wordt om integriteit te meten merken wij op dat het selecteren op integriteit in de context van de rio-opleiding heel moeilijk is. Dit omdat zonder selectie op intregiteit al zeer waarschijnlijk de grote meerderheid van rio's integer handelt (er is dus sprake van een zeer hoge 'base rate'). Gezien dat integriteit moeilijk te voorspellen is zou een selectie op integriteit tot veel vals-negatieve beslissingen leiden (d.w.z. er worden veel kandidaten ten onrechte als niet integer aangewezen), in ieder geval als er redelijk streng op wordt geselecteerd.

Wij raden ook aan de Everest Games niet meer te gebruiken. Er was weinig informatie beschikbaar om de kwaliteit van de Everest Games te beoordelen en dit instrument wordt momenteel holistisch gebruikt. Er werd door LTP vermeld dat de Everest Games *in plaats van* testen om cognitieve vaardigheden te meten gebruikt kunnen worden en niet als toevoeging. Scores uit de Everest Games hingen inderdaad sterk samen met scores op de analytische tests welke al cognitieve capaciteiten meten. Bovendien hingen scores uit de Everest Game sterk negatief samen met leeftijd en werd dit selectie-instrument ten opzichte van andere instrumenten als redelijk negatief door oud-kandidaten ervaren, zoals het ook al in de vorige evaluatie het geval was (Niessen & Meijer, 2019). Omdat de TVRA test voldoet aan de psychometrische eisen en onderzoek naar de kwaliteit van de Everest Games ontbreekt kan de TVRA test gehandhaafd worden en kan de Everest Games komen te vervallen.

Wij raden ook aan het assessmentgesprek met een LTP-adviseur te laten vervallen. Op dit moment is het assessmentgesprek op standaardisatie van competenties na niet gestandaardiseerd. Omdat de gesprekken niet gescoord worden is ook het gebruik van beslisregels (mechanische combinatie) per definitie onmogelijk. Bovendien blijft het onduidelijk wat in het assessmentgesprek wordt gemeten en worden er al gesprekken met de kandidaten gevoerd (lokale selectie en eindgesprekken). Het assessmentgesprek lijkt dus ook overbodig.

De potentiële bijdrage van het rollenspel blijft ook onduidelijk omdat de opzet onduidelijk is en de scores van adviseurs en acteurs niet beschikbaar waren. Verder bleek dat de evaluatie van prestatie in het rollenspel niet gestandaardiseerd is omdat er geen beoordelingsformulieren met gedragsankers worden gebruikt. Wel gebruikt een adviseur de indrukken uit het rollenspel om holistisch de competenties te scoren. Het is niet raadzaam om holistisch competentie-oordelen te vormen. De voorspellende waarde van de competentie-oordelen was dan ook verwaarloosbaar klein. Kortom, het rollenspel heeft zoals het op dit moment gebruikt wordt geen toegevoegde waarde. Tot dat er onderzoek is gedaan naar de voorspellende waarde van het rollenspel kan deze beter niet gebruikt worden. Er bestaan dus geen redenen om het rollenspel te handhaven. Mocht dit advies niet worden overgenomen, raden wij nadrukkelijk aan deze instrumenten te standaardiseren en te scoren zodat informatie via een beslisregel gecombineerd kan worden. Het moet dan echter duidelijk zijn wat deze instrumenten meten om te veel overlap met de lokale selectie en de eindgesprekken te voorkomen. Zoals ook in de vorige evaluatie opgemerkt werd (Niessen & Meijer, 2019) zal een rollenspel het beste een onderdeel vormen van een 'assessment center', en niet als een enkele methode worden gebruikt. Een assessment center houdt in dat er meerdere oefeningen zoals rollenspellen en presentaties gebruikt worden om meerdere aspecten te meten welke door meerdere onafhankelijke beoordelaars worden gescoord (Kuncel & Sackett, 2014). De oefeningen die in assessment centers worden ingezet sluiten dicht aan bij het werk wat kandidaten later moeten verrichten. In een assessment center worden dus ook de standaardisatiecomponenten uit Tabel 1 toegepast (Rupp et al., 2015). Als assessment centers op deze wijze ingericht zijn voorspellen ze werkprestaties redelijk goed (Sackett et al., 2022; Sackett, Shewach, et al., 2017). Het ontwikkelen

van een goede assessment center eist echter hulp van een assessmentpsycholoog en zal kostbaar zijn. Rollenspellen kunnen dus nuttig zijn als zij op de juiste manier gebruikt worden, idealiter in de vorm van een assessment center.

Ten slotte raden wij aan om instrumenten die tot stand komen op basis van een combinatie van persoonlijkheidsschalen niet meer te gebruiken (Persoonlijke Veranderkracht en Teamrollen). Hierdoor wordt dezelfde informatie twee keer gebruikt. Bovendien blijft het onduidelijk welke instrumenten onderdeel uitmaken van deze instrumenten.

Aanbeveling 4

- Alleen de Big Five Plus vragenlijst (of een vergelijkbaar instrument van goede kwaliteit) handhaven, en alleen de schaal Consciëntieusheid uit de Big Five Plus vragenlijst gebruiken.
- De Work Personality Questionnaire niet meer gebruiken omdat deze overlap met de Big Five Plus toont en geen voorspellende waarde heeft voor prestatie in de opleiding.
- De Drijfveren- en Waardenvragenlijst niet meer gebruiken omdat deze onvoldoende voorspellende waarde toonden.
- De Beïnvloedingsstijlenvragenlijst en de *plusschalen* van de Big Five Plus niet meer gebruiken omdat deze onvoldoende betrouwbaar waren.
- De Dilemmatest niet meer gebruiken omdat deze maar als *praatstuk* gebruikt wordt en onderzoek naar de kwaliteit van dit instrument ontbreekt.
- De Everest Game niet meer gebruiken omdat de scores sterk overlappen met scores op de analytische tests, de scores behoorlijk negatief samenhangen met leeftijd, en dit instrument minder positief door kandidaten wordt ervaren.
- Het assessmentgesprek niet meer gebruiken omdat de standaardisatie ontbreekt, het onduidelijk blijft wat er wordt gemeten, en er al meerdere gesprekken gevoerd worden (lokale selectie en eindgesprekken).
- In het algemeen en ook in de toekomst alleen tests gebruiken waarvan de psychometrische kwaliteit is aangetoond (bijv. o.b.v. een COTAN-oordeel).
- Beslisregels opstellen om selectie-instrumenten te combineren.
- De evaluatie van het rollenspel standaardiseren en het rollenspel alleen gebruiken als vervolgonderzoek aantoont dat het prestatie in de opleiding voorspelt. Alternatief een assessment center laten inrichten waarbij er naast het rollenspel meerdere praktijkopdrachten (simulaties) gebruikt worden en kandidaten door meerdere onafhankelijke beoordelaars beoordeeld worden. Wij raden aan de richtlijnen voor assessment centers te gebruiken (Rupp et al., 2015).

3.5 Eindgesprekken

In totaal worden er met elke kandidaat drie eindgesprekken gevoerd. De selecteurs die de eindgesprekken voeren zijn door de LSR geselecteerd op basis van online assessments en een gestandaardiseerd gesprek. De selecteurs volgen ook een training van LTP over het selecteren van kandidaten en het interpreteren van een assessmentrapport. Bij elk gesprek zijn er twee selecteurs aanwezig en wordt een kandidaat op drie competenties beoordeeld. In juni 2024 heeft de Rechtspraak nieuwe beoordelingsformulieren ingevoerd die de standaardisatie in positieve zin verhogen. Met deze beoordelingsformulieren beoordelen beide selecteurs een kandidaat steeds onafhankelijk op een vijfpuntsschaal, op de drie competenties die bij een gesprek horen. Een uitzondering vormt de

beoordeling van de competentie zelfreflectie en leervermogen. Hierover worden geen vragen gesteld. Beoordelaars worden alleen gevraagd om aan te geven of zij tijdens het gesprek de indruk hebben gekregen dat de kandidaat over voldoende zelfreflectie en leervermogen beschikt (nee of ja). Een betere manier zou zijn om concrete vragen over deze competentie te stellen en deze volgens de in Tabel 1 genoemde standaardisatiecomponenten te scoren. Nu kan het immers voorkomen dat er nee moet worden gekozen, enkel omdat de beoordelaar hier geen voldoende indruk van heeft kunnen vormen op basis van het gesprek.

In een apart bestand zijn gedragsankers weergegeven die bij de numerieke beoordelingen van een competentie horen. Ten slotte worden de beoordelingen per beoordelaar volgens een vooraf opgestelde beslisregel gecombineerd. Dit is zeer wenselijk en een positieve verandering sinds de vorige evaluatie (Niessen & Meijer, 2019). Echter wordt er dus niet per vraag gescoord, en worden er nog steeds verschillende vragen aan verschillende kandidaten gesteld. Op dit moment volgt *per interviewer* een negatief oordeel als 1) drie of meer competenties als 'matig' beoordeeld worden, 2) er één competentie als 'onvoldoende' wordt beoordeeld, of 3) het vermogen tot zelfreflectie en leervermogen van de kandidaat niet positief wordt beoordeeld. Als maar één of twee competenties als 'matig' (2) beoordeeld worden volgt een eindoordeel 'matig positief'. Als alle competenties 'voldoende' tot 'uitstekend' zijn volgt een eindoordeel 'positief'. Na afloop van de drie eindgesprekken vindt de beraadslaging plaats. Tijdens de beraadslaging gaan selecteurs hun oordelen met elkaar bespreken. Na het overleg vindt er een anonieme stemming plaats. Als vier of meer selecteurs een positief oordeel geven wordt de kandidaat aangenomen en anders afgewezen.

Betrouwbaarheid van de eindgesprekken

Om de betrouwbaarheid van de eindgesprekken te bepalen is de interbeoordelaarsbetrouwbaarheid per thema en eindbeoordeling van een gesprek berekend op basis van de beoordelingen die zijn gegeven door de selecteurs. Hierbij moet opgemerkt worden dat de gerapporteerde resultaten gebaseerd zijn op de manier hoe de eindgesprekken doorgevoerd werden voordat de nieuwe beoordelingsformulieren ingevoerd werden. Deze resultaten reflecteren dus nog niet de nieuwe veranderingen die hierboven zijn beschreven en pas in juni 2024 ingevoerd zijn. Daarom is het niet mogelijk een uitspraak over de betrouwbaarheid van de huidige eindgesprekken te doen. De data werden aangeleverd door de LSR. Er was data van 214 tot 252 kandidaten beschikbaar, afhankelijk van de competentie die beoordeeld werd. De interbeoordelaarsbetrouwbaarheid geeft aan in hoeverre de twee selecteurs in hun oordelen overeenkomen en is in Tabel 3 weergegeven. De betrouwbaarheid van de themascores en eindoordelen was onvoldoende. Dit betekent dus dat twee selecteurs die dezelfde kandidaat in hetzelfde gesprek hebben gezien weinig overeenstemming in hun oordelen tonen. We raden aan om dit soort onderzoek voor de huidige vorm van eindgesprekken te herhalen. Vervolgonderzoek zou nuttig zijn zodra genoeg (ongeveer 100, maar liefst meer) kandidaten volgens de nieuwe vorm van de eindgesprekken beoordeeld zijn. Het onderzoek kan herhaald worden, bijvoorbeeld zodra er weer 100 nieuwe kandidaten zijn beoordeeld, om een nauwkeurigere schatting van de betrouwbaarheid te krijgen. Het onderzoek zal zeker herhaald moeten worden als de vorm of standaardisatie van de gesprekken veranderd, ook is minder standaardisatie niet wenselijk. Wij merken ook op dat als de gesprekken op onze aanbevolen manier afgenomen en gescoord worden, onderzoek naar de betrouwbaarheid al plaats kan vinden voor de volgende reguliere evaluatie. Dit geldt ook voor andere selectie-instrumenten zoals de lokale gesprekken en de briefselectie. Voor een analyse naar de betrouwbaarheid zijn alleen maar onafhankelijke beoordelingen van de selecteurs uit de gesprekken, brieven, en referenties nodig, maar geen gegevens over de prestatie van rio's.

Als de data systematisch wordt opgeslagen zal een betrouwbaarheidsanalyse makkelijk uit te voeren zijn.

Tabel 3 *Interbeoordelaarsbetrouwbaarheid van de themascores en eindoordelen per eindgesprek*

Thema/eindoordeel	*ICC*
Gesprek 1	
Maatschappelijke nevenactiviteiten	.55 [.43, .65]
Samenwerken	.38 [.21, .52]
Evenwichtigheid	.51 [.36, .72]
Eindoordeel selecteurs gesprek 1	.72 [.63, .78]
Gesprek 2	
Visie op maatschappelijke ontwikkelingen	.56 [.43, .66]
Besluitvaardigheid	.43 [.27, .56]
Regievoering	.58 [.46, .68]
Eindoordeel selecteurs gesprek 2	.48 [.32, .60]
Gesprek 3	
Visie op rechterschap	.30 [.09, .46]
Inlevingsvermogen	.34 [.14, .49]
Flexibiliteit	.28 [.06, .44]
Eindoordeel selecteurs gesprek 3	.46 [.29, .59]

Notitie. ICC = intra-klasse correlatie. ICC (2, *k*) is berekend met de icc functie van het irr package in *R*. Interpretatie: < .80 = onvoldoende, < 0.9 = voldoende, >= .90 goed (Evers et al., 2010). 95% betrouwbaarheidsinterval tussen haakjes. Het betrouwbaarheidsinterval geeft de nauwkeurigheid van een bepaalde schatting weer (bijv. van een correlatie, een gemiddelde, of een betrouwbaarheidscoëfficiënt). *N* = 214 – 252.

Voorspellende waarde van de eindgesprekken

Uit de eindgesprekken volgde geen enkele totaalscore, zoals bijvoorbeeld een gemiddeld oordeel over de zes selecteurs heen. Deze situatie zorgt ervoor dat de validiteit van de eindgesprekken zoals die daadwerkelijk gebruikt werden niet te schatten valt. Desondanks hebben wij de validiteit van de eindgesprekken bepaald om een indruk te geven *wat de voorspellende waarde zou kunnen zijn* als de eindgesprekken op die manier ingericht worden. We hebben eerst per competentie totaalscores gevormd door steeds het gemiddelde van de twee selecteurs op iedere competentie te berekenen. In gevallen waar alleen maar een oordeel van een interviewer beschikbaar was werd dit

enkele oordeel gebruikt omdat er anders niet genoeg data beschikbaar was[3]. Vervolgens werd het gemiddelde over deze competentie scores berekend om een totaalscore 'Eindgesprek' te vormen (zie ook Figuur 1). Het verband tussen deze eindgesprekscore en prestatie in de opleiding was middelgroot tot groot (r = .38). Ook dit resultaat komt overeen met bevindingen uit de literatuur (Sackett et al., 2022). Echter is het hierbij heel belangrijk te benadrukken dat deze analyse niet representatief is voor hoe selectiebeslissingen op basis van de eindgesprekken op dit moment worden genomen. In de huidige selectieprocedure worden de competentie-oordelen van de selecteurs immers niet gemiddeld en wordt er vervolgens ook geen gemiddelde over competenties heen berekend. Deze resultaten geven dus een indicatie van hoe valide de oordelen uit de eindgesprekken *zouden kunnen zijn*, als ze op deze manier tot stand zouden komen. In de huidige selectieprocedure kunnen selecteurs hun oordelen achteraf in een groepsverband bespreken en bijstellen. Uit de literatuur blijkt dat dit ruis introduceert en tot slechtere beslissingen leidt dan het consistente gebruik van een beslisregel (Dilchert & Ones, 2009; Kuncel et al., 2013). Wij concluderen dat de voorspellende waarde van de eindgesprekken zoals ze tot juni 2024 uitgevoerd werden hooguit klein tot middelgroot is, maar hoog zou kunnen zijn als selecteurs hun oordelen zouden middelen en van de beraadslaging afzien. Omdat alleen kandidaten met een positief eindoordeel werden aangenomen was het niet mogelijk de validiteit van de uiteindelijke selectiebeslissingen te onderzoeken.

3.5.1 Ervaringen van gebruikers

De eindgesprekken worden door gebruikers als zeer positief en nuttig ervaren. Dit was ook te verwachten, gezien veel gebruikers als selecteurs in de eindgesprekken fungeren en op basis van deze gesprekken selectiebeslissingen nemen. Men had het gevoel dat de selectie zorgvuldig verloopt en het trainen van de selecteurs bijdraagt aan het selecteren van goede rio's. De meeste gebruikers gaven aan open te staan voor veranderingen die selectiebeslissingen zouden verbeteren en bias zouden kunnen voorkomen, en dat zij ook regelmatig deelnemen aan terugkomdagen en trainingen over selectie. Het voorstel, de beraadslaging te laten vervallen en in plaats daarvan een beslisregel te gebruiken om selectiebeslissingen te nemen werd echter door enige gebruikers als nadelig ervaren en vond men dat daardoor een waardevolle uitwisseling van informatie verloren zou gaan. Wij merken opnieuw op dat beraadslagingen tot minder goede beslissingen leiden dan het consistente combineren van informatie volgens een beslisregel (Dilchert & Ones, 2009).

3.5.2 Ervaringen van oud-kandidaten

De eindgesprekken werden positief ervaren. In totaal gaven er 96 respondenten een toelichting. Daarbij gaven respondenten aan dat zij drie gesprekken positief hebben ervaren omdat zij dachten dat een wat minder goed verlopend gesprek zo makkelijker gecompenseerd kon worden door een goed gesprek (n = 22). Terwijl sommige respondenten de eindgesprekken prettig vonden (n = 8) vonden anderen de gesprekken onprettig, willekeurig, of niet effectief (n = 15). Verder gaven respondenten aan dat gesprekken in deze fase (na het assessment en het lokale gesprek) niet meer veel zouden uit moeten maken (n = 5). Sommigen vonden het aantal gespreken te hoog (n = 4) of de online afname van de onlinegesprekken onprettig (n = 6). Een aantal respondenten vond het betrekken van externe leden niet passend omdat deze als onplezierig werden ervaren (n = 6). Verder is aan de respondenten

3 Wij merken op dat er vaak data ontbrak. Dit betekent in ieder geval dat de scores uit de eindgesprekken niet consequent werden opgeslagen. Het valt niet te beantwoorden of in gevallen waar data ontbrak in eerste instantie überhaupt competentie-oordelen werden gegeven door de selecteurs, wat bewijs voor een gebrek aan standaardisatie zou zijn (zie Tabel 1).

gevraagd in hoeverre zij open staan voor het gebruik van beslisregels om zo wel informatie binnen een gesprek, als informatie uit verschillende instrumenten te combineren, om een selectie beslissing te nemen (zie Bijlage 5 voor de vragen). In totaal had 58% van de respondenten een voorkeur antwoorden in gesprekken en informatie uit verschillende selectie-instrumenten via groepsoverleg te combineren. Terwijl er dus een lichte voorkeur voor holistische selectiebeslissingen bestond was een grote minderheid (42%) geneigd informatie volgens een beslisregel te combineren.

3.5.3 Aanbevelingen

Wij bevelen aan om, zoals het voor andere competenties ook wordt gedaan, vragen naar concreet gedrag te stellen om het vermogen tot zelfreflectie en leervermogen op een vijfpuntsschaal te beoordelen. Op dit moment wordt een interviewer alleen na afloop van het gesprek gevraagd aan te geven of die 'een indruk heeft kunnen vormen' in hoeverre de kandidaat zelfreflectie toont. Als er geen concrete vragen naar dit gedrag gesteld worden bestaat dus de kans dat sommige kandidaten per toeval hierover iets vertellen en andere kandidaten niet. Zonder concrete vragen te stellen blijft het dus onduidelijk of de kandidaat niks over zelfreflectie zei of op de kandidaat op deze competentie daadwerkelijk laag scoort.

Wij bevelen ook aan om aan alle kandidaten dezelfde vragen in dezelfde volgorde te stellen, de antwoorden *per vraag* en niet alleen per competentie te scoren, en concrete gedragsankers met minder goede en goede antwoorden per vraag op te stellen. Verder raden wij aan om te overwegen de eindgesprekken samen te voegen en een eindgesprek te voeren waar meerdere beoordelaars ingezet worden om dezelfde competenties onafhankelijk te beoordelen. Deze onafhankelijke competentie-oordelen kunnen dan over beoordelaars heen gemiddeld worden en vervolgens tot een beoordeling opgeteld worden (compensatorische combinatie). Alternatief kunnen er minimumeisen voor de gemiddelde competentie-oordelen gehanteerd worden (noncompenatorische combinatie). Een voorbeeld zou zijn dat een kandidaat afgewezen wordt als het gemiddelde competentie-oordeel lager is dan twee op een vijfpuntsschaal. In alle andere gevallen wordt de kandidaat aangenomen. Als besloten wordt drie eindgesprekken te handhaven raden wij ook aan om de onafhankelijke competentie-oordelen van de selecteurs per gesprek te middelen en een beslisregel zoals hierboven voorgesteld te gebruiken om tot een oordeel te komen. Ten slotte is het belangrijk om te voorkomen dat aanvullende informatie zoals brieven, assessment rapporten, en analytische test scores de evaluatie van de competenties in de eindgesprekken beïnvloeden (zie Tabel 1). Wij raden dus af om vooraf aan de eindgesprekken het dossier van een kandidaat te bekijken en op basis daarvan verschillende vragen aan verschillende kandidaten te stellen. Hierdoor wordt ruis geïntroduceerd en zullen verschillende selecteurs de informatie uit het dossier verschillend gaan wegen voor verschillende kandidaten. Uiteraard kan aanvullende informatie zoals analytische test scores met gekwantificeerde oordelen uit de eindgesprekken gecombineerd worden om een finale selectiebeslissing te nemen. Wij legen het combineren van verschillende instrumenten uit in het volgende hoofdstuk.

Aanbeveling 5 – Eindgesprekken standaardiseren en beslisregels toepassen

- Competentie-oordelen van selecteurs middelen en een beslisregel opstellen om tot een eindoordeel voor het eindgesprek/de eindgesprekken te komen. Dit kan gebeuren door de gemiddelde competentie-oordelen simpelweg op te tellen of minimaal eisen per competentie op te stellen.
- Antwoorden scoren in plaats van globale competenties. Gedragsankers met minder goede en goede antwoorden per vraag opstellen.
- De competentie zelfreflectie en leervermogen op basis van gedragsrelevante vragen beoordelen en zoals andere competenties op een vijfpuntsschaal scoren.
- Ervoor zorgen dat aanvullende informatie zoals brieven, assessment rapporten en test scores niet vooraf aan de eindgesprekken wordt bekeken.

4 Inrichting van de gehele selectieprocedure

In dit hoofdstuk bespreken wij hoe de verschillende selectie-instrumenten als geheel het beste gebruikt kunnen worden. Wij gaan daarbij uit van een scenario waarin de eerdere aanbevelingen worden overgenomen en besteden ook aandacht aan de toekomstige geschatte prestatie en de diversiteit van geselecteerde rio's.

Het validiteits-diversiteitsdilemma en nieuwe schattingen van de voorspellende waarde van selectie-instrumenten

Een bevinding uit de literatuur is dat sterk cognitief georiënteerde selectie-instrumenten zoals analytische tests vaak de grootste verschillen laten zien tussen kandidaten met en zonder een etnisch-culturele achtergrond en tussen jongere en oudere kandidaten, maar tegelijkertijd ook hoge voorspellende waarde voor toekomstig werkprestatie tonen (Dahlke & Sackett, 2017). Dit wordt ook het *validiteits-diversiteitsdilemma* genoemd (Ployhart & Holtz, 2008). Recent zijn hierover nieuwe bevindingen in de literatuur verschenen (Sackett et al., 2022, 2024) die het verband tussen scores op cognitieve capaciteitentests en werkprestatie iets lager schatten dan eerder onderzoek (Schmidt & Hunter, 1998). Dit komt mede door het toepassen van andere statistische technieken. Niet cognitieve capaciteiten, maar gestandaardiseerde gesprekken worden nu als de beste voorspeller van werkprestatie gezien en tonen redelijk kleine verschillen in scores tussen kandidaten met en zonder etnisch-culturele achtergrond (Sackett et al., 2022). Een grotere rol voor gestandaardiseerde gesprekken is daarmee een effectieve manier om zowel prestaties als diversiteit te verbeteren. Geheel vrij van verschillen zijn gestandaardiseerde gesprekken echter niet. Wij merken ook op dat de gesprekken waarop deze bevindingen gebaseerd zijn uiterst gestandaardiseerd en meer gestandaardiseerd zijn dan de gesprekken in de huidige selectieprocedure. Daarbij geldt dat het praktisch lastig is om bij een grote hoeveelheid kandidaten gestandaardiseerde gesprekken in te zetten. Het is daarom goed voorstelbaar dat er eerst andere instrumenten worden ingezet, zoals de analytische tests.

Beslisregels opstellen

Ten eerste moet bij het opstellen van een beslisregel besloten worden of de selectie-instrumenten noncompensatorisch of compensatorisch gebruikt worden (of een combinatie hiervan). In de huidige selectieprocedure is het volgens de LSR de bedoeling dat de selectie-instrumenten noncompensatorisch gebruikt worden. Het idee is bijvoorbeeld dat een kandidaat die de analytische tests behaald vervolgens voldoende moet worden beoordeeld tijdens de lokale selectie. Daarbij is het de bedoeling dat voor elke stap in de selectieprocedure slechts informatie verzameld tijdens die stap wordt gebruikt, zonder ook nog informatie uit eerdere selectie-instrumenten mee te wegen. Dat impliceert dat een kandidaat die voldoende in de eindgesprekken scoort automatisch zou moeten worden aangenomen, en dat er geen beraadslaging nodig zou zijn. In deze aanpak zouden dus van tevoren per selectie-instrument harde grensscores bepaald moeten worden (bijvoorbeeld, een kandidaat gaat door wanneer de gemiddelde eindbeoordeling op een vijfpuntsschaal in het lokale gesprek een 3 of hoger is, een kandidaat wordt aangenomen wanneer de eindbeoordeling in het eindgesprek een 3 of hoger is). Wij raden deze aanpak aan omdat deze makkelijk uit te voeren is en zo alle kandidaten op dezelfde manier worden beoordeeld.

In de huidige selectieprocedure wordt informatie echter zowel noncompensatorisch als ook compensatorisch gebruikt. Kandidaten moeten eerst grensscores op de analytische tests behalen en kunnen bij de briefselectie en de lokale gesprekken afvallen (ookal ontbreken voor de briefselectie en de lokale gesprekken nog expliciete grensscores). Uiteindelijk wordt informatie echter, zoals uit de gesprekken met selecteurs bleek, holistisch gecombineerd, waarbij selecteurs informatie uit de eindgesprekken combineren met eerdere informatie zoals scores op de analytische tests en indrukken uit het assessmentrapport. Informatie wordt dus ook in zekere mate compensatorisch gebruikt. Hoe dit precies gebeurt blijft echter eerder een 'black box'. Verschillende selecteurs gebruiken informatie verschillend, en wellicht noncompensatorisch, compensatorisch of een combinatie daarvan. Voor het geval dat de LSR toch een voorkeur heeft om in de laatste fase van de selectieprocedure informatie compensatorisch te gebruiken, zoals het nu het geval is, geven wij in het volgende een voorbeeld aan hoe dit het beste kan gedaan worden. Dit voorbeeld is ook van toepassing als door gebrek aan kandidaten wordt besloten geen getrapte selectieprocedure meer te handhaven (d.w.z., eerst op analytische vaardigheden selecteren en alleen de resterende kandidaten op basis van andere selectie-instrumenten te selecteren). In een krappe arbeidsmarkt zouden dus alle kandidaten alle selectie-instrumenten kunnen doorlopen.

Als men informatie na het behalen van de TVRA test en de lokale selectie compensatorisch zou willen gebruiken, zou een simpele beslisregel kunnen zijn: Geschiktheid = Score TVRA test*$w1$ + Score lokale gesprek*$w2$ + Score eindgesprek*$w3$. De score op het lokale gesprek zou de (gemiddelde) beoordeling uit het lokale gesprek zijn. De score op het eindgesprek zou de gemiddelde beoordeling over meerdere beoordelaars en competenties zijn. De w's bepalen hoe zwaar de scores worden gewogen. Dat kan ook gelijk zijn, maar dat hoeft niet. Omdat er al eerder is geselecteerd op basis van de TVRA test en het lokale gesprek, is het als dat zo blijft, aan te bevelen om het meeste gewicht aan het eindgesprek toe te kennen. Hierbij moet er wel op gelet worden dat de TVRA testscore en de scores op beide gesprekken op dezelfde schaal (bijvoorbeeld een vijfpuntsschaal) worden geplaatst (Meijer et al., 2020). De testaanbieder zou wellicht kunnen helpen om de scores op de TVRA test om te zetten naar de gewenste schaal. Een handleiding die binnenkort als addendum bij de Algemene Standaard Testgebruik van het Nederlands Instituut voor Psychologen (NIP) wordt gepubliceerd bevat informatie hoe dit het beste kan worden gedaan (Niessen et al., 2024). Als minder wenselijk maar gemakkelijker uitvoerbaar alternatief zouden de selecteurs de scores zelf kunnen omzetten volgens een simpele regel zoals weergegeven in het beoordelingsformulier in Figuur 2. Daarin worden scores op een cognitieve capaciteitentest omgezet door scores die in een bepaald bereik vallen dezelfde score te geven. Ten slotte moet nog bepaald worden hoe de totaalscore gebruikt wordt om selectiebeslissingen te nemen. Dit kan op een 'top-down' manier (d.w.z. de hoogst scorende kandidaten krijgen de beschikbare opleidingsplekken aangeboden). Alternatief kan een grensscore vastgelegd worden. Als kandidaten deze grensscore behalen worden zij aangenomen.

Figuur 2 *Voorbeeld van een Beoordelingsformulier uit Meijer et al. (2020)*

General information

Name applicant: Candidate 1

Cognitive ability test	1 ≤95, 1.5: 96-100, 2: 101-105, 2.5: 106-110, 3: 111-115, 3.5: 116-120, 4: 121-125, 4.5: 126-130, 5: > 130 **Score: 121**				
	☐ 1	☐ 2	☐ 3	x 4	☐ 5
Conscientiousness	**Score: 4.8**				
	☐ 1	☐ 2	☐ 3	☐ 4	x 5
Biodata scale	**Score: 3.1**				
	☐ 1	☐ 2	x 3	☐ 4	☐ 5
Structured interview rating	**Mean across raters: 3.2**				
	☐ 1	☐ 2	x 3	☐ 4	☐ 5
Final rating (sum)	**15**				

Remarks

Pareto-optimalisatie

Zoals eerder uitgelegd betekent het validiteits-diversiteitsdilemma dat selectie-instrumenten met de hoogste voorspellende waarde vaak ook de grootste verschillen in scores laten zien tussen bijvoorbeeld kandidaten met en zonder een etnisch-culturele achtergrond. Dit dilemma heeft als gevolg dat organisaties niet tegelijkertijd de voorspellende waarde van de selectieprocedure en het percentage diverse kandidaten die geselecteerd worden kunnen maximaliseren. Een organisatie kan dan een afweging maken over het relatieve belang van beide doelen. Zodra deze beleidskeuze is gemaakt moet de vraag beantwoord worden hoe de selectie-instrumenten het beste kunnen worden gewogen om zo goed mogelijk recht te doen aan het relatieve belang van beide doelstellingen. Hiervoor kan Pareto-optimalisatie gebruikt worden. Dit is een vrij nieuwe techniek waarbij selectie-instrumenten zodanig worden gewogen dat voor een bepaalde gewenste ratio in de kans op aangenomen te worden voor kandidaten met en zonder etnisch-culturele achtergrond, de hoogst mogelijke voorspellende waarde wordt behouden (Rupp et al., 2020), of andersom, waarbij voor een minimaal gewenste voorspellende waarde de maximale diversiteit wordt bereikt.

Het aanpassen van de weging van selectie-instrumenten om tot een gewenste balans van diversiteit en validiteit te komen vereist wel dat selectie-instrumenten gescoord en volgens een beslisregel gecombineerd worden. Het doel van de Raad voor de rechtspraak om zowel goede prestaties als diversiteit te waarborgen kan dus het beste behaald worden door het gebruik van beslisregels, omdat dit het mogelijk maakt bepaalde wegingen van selectie-instrumenten zo toe te passen dat de gewenste balans bereikt wordt. Een vrij toegankelijke tool om Pareto-optimalisatie toe te passen is via deze link te vinden: https://qchelseasong.shinyapps.io/ParetoR/. Het toepassen van Pareto-optimalisatie vereist echter hulp van onderzoekers of een assessmentpsycholoog die de nodige expertise bezit. Ook kunnen wij op dit moment geen concreter advies over de toepassing van Pareto-optimalisatie geven, omdat de gewenste diversiteit/voorspellende waarde en informatie zoals het aantal kandidaten met en zonder etnisch-culturele achtergrond niet bekend zijn.

Bij de eerdergenoemde beslisregel (Geschiktheid = Score TVRA test*$w1$ + Score lokaal gesprek*$w2$ + Score eindgesprek*$w3$) kan ook overwogen worden om in de eerste fase waarin op de TVRA test wordt geselecteerd de Consciëntieusheid schaal uit de Big Five Plus toe te voegen en de TVRA test en de Consciëntieusheid schaal compensatorisch te wegen. Om de criteriumvaliditeit te maximaliseren zou meer gewicht aan de analytische tests worden moeten gegeven. Meer gewicht geven aan Consciëntieusheid zou echter voordelig voor de diversiteit zijn, omdat verschillen in Consciëntieusheid scores tussen kandidaten met en zonder etnisch-culturele achtergrond vaak zeer klein zijn (Sackett et al., 2022). De optimale oplossing gegeven gewenste uitkomsten zou zoals eerder genoemd via Pareto-optimalisatie te vinden zijn. Het gebruik van technieken zoals Pareto-optimalisatie eist echter kennis over het aantal kandidaten met een etnisch-culturele achtergrond en informatie over verschillen in scores op de verschillende selectie-instrumenten. Die informatie is momenteel niet beschikbaar. Bij gebrek daaraan kunnen schattingen gebruikt worden, maar ideaal is dat niet. Als dit een belangrijke doelstelling is, raden we daarom aan om over een manier na te denken waarop die informatie in de toekomst toch beschikbaar kan komen voor onderzoek, uiteraard op vertrouwelijke wijze en rekening houdend met de gevoeligheid waarmee dat gepaard kan gaan.

Aanbeveling 6 – Etnisch-culturele achtergrond van kandidaten opslaan

- Op vertrouwelijke wijze informatie over etnisch-culturele achtergrond van kandidaten vragen, uitsluitend voor onderzoeksdoeleinden en zonder dat selecteurs of anderen daar toegang toe hebben, om analyses naar diversiteit mogelijk te maken.

Acceptatie van beslisregels verhogen

Als de LSR informatie compensatorisch op een optimale manier wil gebruiken raden wij aan Pareto-optimalisatie toe te passen. Hierdoor verliezen selecteurs echter zeggenschap over de beslisregel. Dit roept vaak weerstand op en accepteren selecteurs soms eerder een aanpak waarin zij zelf de beslisregel mogen opstellen, ten opzichte van een ‘optimale’ beslisregel (Niessen et al., 2024; Nolan & Highhouse, 2014). Mocht dit het geval zijn kunnen selecteurs bij het opstellen van de beslisregel betrokken worden en de gewichten meebepalen. Het is moeilijk om in deze fase concreter advies over een ‘optimale’ beslisregel te geven, omdat dit afhankelijk is van onder ander de selectie-instrumenten die gebruikt worden, de selectie-ratio’s, en de kwaliteit van de selectie-instrumenten, welke voor sommige instrumenten onbekend zijn. Dit is echter geen reden om geen beslisregel te gebruiken, omdat zelfs ‘suboptimale’ beslisregels in aanzienlijk betere beslissingen resulteren dan holistische combinatie (Dawes, 1979). Het gebruik van beslisregels zal de selectie-procedure bovendien transparanter maken, waarborgt gelijke behandeling, en maakt het de Rechtspraak en onderzoekers mogelijk om de selectieprocedure in de toekomst beter te evalueren. Wij bevelen daarom aan om de meerwaarde van het gebruik van beslisregels op te nemen in de trainingen die worden aangeboden voor selecteurs. Onderzoek liet zien dat regelmatige training het gebruik van beslisregels en de acceptatie ervan kan bevorderen (Neumann et al., 2022). Mocht het gebruik van beslisregels veel weerstand oproepen, dan is het in ieder geval raadzaam om selecteurs de uitkomst op basis van een beslisregel als uitgangspunt of advies te geven, omdat dit tot betere selectiebeslissingen leidt dan puur holistische combinatie zonder advies van een beslisregel (Neumann et al., 2024). Het leidt echter tot slechtere beslissingen dan het consistent gebruik van een beslisregel (Hoffman et al., 2018; Neumann et al., 2024), omdat gebruikers vaker ten onrechte dan terecht denken dat er een uitzondering gemaakt moet worden of een bijstelling op zijn plaats is. Daarom heeft deze aanpak niet de eerste voorkeur.

Aanbeveling 7

- In trainingen die aangeboden worden voor selecteurs van de LSR en de lokale gesprekken de evidentie behandelen dat beslisregels tot betere selectiebeslissingen leiden dan het menselijke oordeel. Daarbij is het raadzaam selecteurs te informeren dat het consistente gebruik van een beslisregel belangrijker is dan de exacte vorm van de beslisregel, tenminste als alleen maar selectie-instrumenten van goede kwaliteit worden gebruikt. Dit is omdat beslisregels schadelijke inconsistentie in menselijke oordelen reduceren. Dit soort informatie kan ervoor zorgen dat selecteurs een beslisregel consistenter toepassen (Neumann et al., 2022).
- Wij benadrukken dat het belangrijk is beslisregels *consistent* te gebruiken. Mocht er grote weerstand bestaan raden wij aan op zijn minst de voorspelling van een beslisregel als advies te beschouwen en uiterst voorzichtig te zijn deze te overrulen omdat dit de kwaliteit van selectiebeslissingen *vermindert* (Hoffman et al., 2018; Neumann et al., 2024).
- Beslisregels in alle stappen van de selectieprocedure gebruiken en niet alleen om de finale selectiebeslissing te nemen. Beslisregels moeten dus ook gebruikt worden om te bepalen wanneer een kandidaat de briefselectie en de lokale gesprekken haalt (en uiteraard ook om het advies van een adviseur te bepalen mocht het assessment toch gehandhaafd blijven).

5 Aanbevelingen

In Tabel 4 geven wij een overzicht van onze aanbevelingen. De aanbevelingen zijn per thema grof geordend van zeer belangrijk tot minder belangrijk. Wij onderscheiden ook tussen 'must have' aanbevelingen en 'nice to have' aanbevelingen. 'Nice to have' aanbevelingen zijn wenselijk om uit te voeren maar zullen een kleiner effect op de kwaliteit van selectiebeslissingen, diversiteit, en acceptatie door kandidaten hebben en hebben daarom niet de hoogste prioriteit. Ook beschrijven wij in Tabel 4 welke informatie van adviesbureaus opgevraagd zou moeten worden om in de toekomst de kwaliteit van assessments te bepalen.

Tabel 4 *Overzicht van aanbevelingen*

'Must have' aanbevelingen
Briefselectie
• De scorelijst voor de briefselectie verder standaardiseren en vooral voor consistent gebruik zorgen • Beschrijvende ankers toevoegen om te definiëren wanneer een factor als goed of minder goed beoordeeld wordt (bijvoorbeeld: wanneer zijn studieresultaten matig of sterk? Zie ook standaardisatiecomponent 3 uit Tabel 1) • Per gerecht een beslisregel opstellen om tot een oordeel te komen over wanneer een kandidaat wel of niet doorgaat • Toezicht vanuit LSR op juiste manier gebruik van de scorelijst. De huidige begeleiding door de LSR is minder effectief omdat selecteurs herinnerd worden holistisch op de factoren in de scorelijst te letten (bijvoorbeeld: "Heb je ook goed naar de studieresultaten gekeken"?). De LSR zou beter kunnen checken of 1) de factoren op de vijfpuntsschaal in de huidige scorelijst ook daadwerkelijk gescoord zijn en niet leeg blijven, 2) een beslisregel daadwerkelijk wordt gebruikt om tot een oordeel te komen en er dus niet van de beslisregel wordt afgeweken
Analytische tests
• Alleen de TVRA test handhaven omdat deze prestatie in de opleiding redelijk goed voorspelt • De Watson-Glaser en de Matrix test niet meer gebruiken omdat deze geen toegevoegde voorspellende waarde hebben boven op de TVRA test • De NIT niet te gaan gebruiken omdat deze onvoldoende betrouwbaar lijkt te zijn
Lokale selectie
• De lokale gesprekken standaardiseren aan de hand van de componenten die in Tabel 1 genoemd zijn. Dit houdt in alleen vragen te stellen die relevant zijn voor de baan, dezelfde vragen aan iedere kandidaat te stellen, en beschrijvende ankers toevoegen om te definiëren wanneer een antwoord als goed of minder goed beoordeeld wordt • Per gerecht bepalen wat 'fit' met het gerecht precies betekend • De referenties hebben waarschijnlijk hooguit een kleine voorspellende waarde. Daarom kan overwogen worden de referenties weg te laten. Mochten de referenties gehandhaafd blijven, raden wij aan de evaluatie door referenten te standaardiseren en beslisregels op te stellen om te bepalen hoe informatie uit de referenties gebruikt wordt. Wij raden dan ook aan de referenties voor iedere kandidaat in dezelfde fase te gebruiken (briefselectie of lokale gesprekken) • Beslisregels opstellen om te bepalen wanneer een kandidaat doorgaat

Assessment

- Alleen de Big Five Plus vragenlijst (of een vergelijkbaar instrument van goede kwaliteit) handhaven, en alleen de schaal Consciëntieusheid gebruiken
- De Work Personality Questionnaire, Drijfveren- en Waardenvragenlijst niet meer gebruiken omdat deze geen voorspellende waarde heeft voor prestaties in de opleiding
- De Beïnvloedingsstijlenvragenlijst en de *plusschalen* van de Big Five Plus niet meer gebruiken omdat deze onvoldoende betrouwbaar waren
- De Dilemmatest niet meer gebruiken omdat deze maar als *praatstuk* gebruikt wordt en onderzoek naar de kwaliteit van dit instrument ontbreekt
- De Everest Game niet meer gebruiken omdat de scores sterk overlappen met scores op de analytische tests, de scores behoorlijk negatief samenhangen met leeftijd, en dit instrument minder positief door kandidaten wordt ervaren
- Het assessmentgesprek niet meer gebruiken omdat de standaardisatie ontbreekt, het onduidelijk blijft wat er wordt gemeten, en er al meerdere gesprekken gevoerd worden (lokale selectie en eindgesprekken)
- Alleen tests gebruiken waarvan de psychometrische kwaliteit is aangetoond (bijv. o.b.v. een COTAN-oordeel)
- Indien het assessment blijft bestaan, beslisregels opstellen om selectie-instrumenten te combineren en bepalen hoe het advies gevormd wordt. Bij handhaving het assessment idealiter als een assessment center laten inrichten waarbij er naast het rollenspel meerdere praktijk-opdrachten (simulaties) gebruikt worden en kandidaten door meerdere onafhankelijke beoordelaars beoordeeld worden

Eindgesprekken

- Competentie-oordelen van selecteurs middelen en een beslisregel opstellen om tot een eindoordeel voor het eindgesprek/de eindgesprekken te komen. Dit kan gebeuren door de gemiddelde competentie-oordelen simpelweg op te tellen of minimaal eisen per competentie op te stellen
- Antwoorden scoren in plaats van globale competenties. Gedragsankers met minder goede en goede antwoorden per vraag opstellen
- De competentie zelfreflectie en leervermogen op basis van gedragsrelevante vragen beoordelen en op een vijfpuntsschaal scoren
- Ervoor zorgen dat aanvullende informatie zoals brieven, assessment rapporten en test scores niet vooraf aan de eindgesprekken wordt bekeken

Beslisregels in alle stappen van de selectieprocedure toepassen en meer informatie over beslisregels in trainingen opnemen

- In alle trainingen die aangeboden worden de evidentie behandelen dat beslisregels tot betere selectiebeslissingen leiden dan het menselijke oordeel. Selecteurs informeren dat het consistente gebruik van een beslisregel belangrijker is dan de exacte vorm van de beslisregel, tenminste als alleen maar selectie-instrumenten van goede kwaliteit worden gebruikt. Dit is omdat beslisregels schadelijke inconsistentie in menselijke oordelen reduceren. Dit soort informatie kan ervoor zorgen dat selecteurs een beslisregel consistenter toepassen (Neumann et al., 2022).
- Beslisregels *consistent* gebruiken. Mocht er grote weerstand bestaan kan de voorspelling van een beslisregel op zijn minst als advies worden beschouwen. Wij raden echter af de voorspelling van de beslisregel te overrulen omdat dit de kwaliteit van selectiebeslissingen *vermindert*
- Beslisregels in alle stappen van de selectieprocedure gebruiken. Beslisregels moeten dus ook gebruikt worden om te bepalen wanneer een kandidaat de briefselectie en de lokale gesprekken haalt

‘Nice to have’ aanbevelingen

Informatie opslaan om diversiteit beter te kunnen onderzoeken

- Omdat diversiteit een belangrijk doel vormt van de selectieprocedure is het wenselijk verschillen in scores op de selectie-instrumenten tussen verschillende groepen van kandidaten te evalueren. Dit was echter niet mogelijk in deze evaluatie omdat de nodige informatie ontbrak. Wij bevelen aan op vertrouwelijke wijze informatie over etnisch-culturele achtergrond van kandidaten te vragen, uitsluitend voor onderzoeksdoeleinden en zonder dat selecteurs of anderen daar toegang toe hebben, om analyses naar diversiteit mogelijk te maken

Verwachtingen van kandidaten managen

- De verwachtingen naar kandidaten voor elke stap in de selectieprocedure en vooral de lokale selectie duidelijker communiceren

De manier hoe prestatie in de opleiding wordt gemeten verbeteren

- Evalueren hoe prestatie in de opleiding eenvoudiger kan worden gemeten om meer vergelijkbaarheid tussen rio’s te schaffen

Evaluatieonderzoek herhalen

- Dit evaluatieonderzoek na het uitvoeren van de aanpassingen herhalen. Hierbij is het belangrijkste dat er dan ook scores beschikbaar zijn van selectie-instrumenten die nu niet onderzocht konden worden omdat data ontbrak, zoals beoordelingen uit lokale gesprekken

Informatie om bij toekomstige aanbestedingen uit te vragen

- Recente handleidingen over de gebruikte selectie-instrumenten, waaruit duidelijk wordt of de selectie-instrumenten voldoen aan de nodige psychometrische kenmerken (bijvoorbeeld betrouwbaarheid, validiteit, normen, onderzoek naar verschillen tussen groepen en predictieve bias).
- Documentatie over ‘kwalitatieve’ selectie-instrumenten zoals gesprekken en simulatieopdrachten (rollenspellen of andere opdrachten die deel uitmaken van een assessment center). Informatie over de interbeoordelaarsbetrouwbaarheid van dit soort selectie-instrumenten. Afnameprotocollen waaruit duidelijk wordt dat de standaardisatie gewaarborgd wordt. Hiervoor kan Tabel 1 uit dit rapport gebruikt worden om de juiste standardisaatiecomponenten uit te vragen
- Documentatie over expliciete beslisregels die gehanteerd worden om de gebruikte selectie-instrumenten/competenties te combineren, of een plan van aanpak dat weergeeft hoe die beslisregels zullen worden opgesteld als de opdracht wordt gegund.

6 Conclusie en antwoorden op de onderzoeksvragen

Wij concluderen dat er door de Rechtspraak veel moeite wordt gedaan om goed presterende en diverse rio's te selecteren en dat het selectieproces verloopt zoals vastgelegd in het handboek LSR (onderzoeksvraag 1). Een positief aspect is dat de analytische tests in het algemeen prestatie in de opleiding redelijk goed voorspellen. Een aantal andere selectie-instrumenten zoals de meeste zelfrapportagevragenlijsten toonden echter nauwelijks voorspellende waarde (onderzoeksvraag 12, 19 en 20). De selectieprocedure kan dus nog aanzienlijk verbeterd en ingekort worden door selectie-instrumenten die niet bijdragen aan het voorspellen van prestaties in de rio-opleiding en grote verschillen in scores laten zien tussen kandidaten met en zonder een etnisch-culturele achtergrond weg te laten of deze te verbeteren of vervangen (onderzoeksvraag 2, 10, 14, 15). Dat houdt onder ander in gesprekken te standaardiseren. Vooral in de lokale gesprekken ontbreekt nog grotendeels standaardisatie en blijft het onduidelijk wat in deze gesprekken wordt gemeten, waardoor het ook voor kandidaten onduidelijk blijft wat zij van deze stap in de selectieprocedure moeten verwachten (onderzoeksvraag 3 en 4). Zoals al uit de laatste evaluatie bleek is het ook belangrijk meerdere selectie-instrumenten volgens een beslisregel te combineren. Het consistente gebruik van beslisregels zou de kwaliteit van selectiebeslissingen verbeteren (onderzoeksvraag 6 en 7) en maakt cultuurwaardevrij selecteren makkelijker te evalueren, als er in de toekomst meer informatie over dat aspect wordt verzameld. Selectie-instrumenten kunnen volgens een beslisregel dan ook op een manier gecombineerd worden waarbij rekening met de diversiteit wordt gehouden, bijvoorbeeld door gestandaardiseerde gesprekken meer gewicht te geven (onderzoeksvraag 16). Het blijft echter door gebrek aan gegevens onduidelijk hoeveel kandidaten met een etnisch-culturele minderheidsachtergrond überhaupt solliciteren, waardoor het moeilijk is concreter advies over een beslisregel te geven. Uit gesprekken met betrokkenen lijken deze aantallen klein te zijn. Het is daarom de vraag in hoeverre de *selectie* de diversiteit van rio's beïnvloedt. In eerste instantie moet de *werving* een voldoende aantal kandidaten met een etnisch-culturele achtergrond aantrekken. Verder zal de selectieprocedure voor kandidaten waarschijnlijk aantrekkelijker worden door instrumenten zoals de Everest Game, welke als redelijk negatief en niet gerelateerd aan de baan werden ervaren, niet meer te gebruiken (onderzoeksvraag 5). Het weglaten van de Everest Games zal de kwaliteit van de selectiebeslissingen niet negatief beïnvloeden omdat cognitieve capaciteiten al met de analytische tests worden gemeten, welke prestatie in de opleiding goed voorspellen.

Er ontbrak data in het bestand dat werd ontvangen van de LSR omdat niet alle selectie-instrumenten consistent gescoord worden (bijvoorbeeld de briefselectie en lokale gesprekken). Het is daarom zeer belangrijk alle instrumenten te scoren om vervolgonderzoek naar de kwaliteit van deze selectie-instrumenten mogelijk te maken. Selecteurs van de lokale selectie worden niet systematisch getraind maar hebben de mogelijkheid om trainingen te volgen. Daartegenover doorlopen de selecteurs van de LSR een selectieprocedure en volgen zij een training van LTP. De selectie en training lijkt dus zorgvuldiger te zijn voor landelijke selecteurs, vergeleken met lokale selecteurs (onderzoeksvraag 7 en 8). Wij benadrukken echter dat het belangrijkste is dat de selecteurs ook daadwerkelijk hulpmiddelen zoals beoordelingsformulieren en beslisregels krijgen en gaan gebruiken. Dit blijkt grotendeels nog niet het geval te zijn (onderzoeksvraag 17 en 21). Zo ontbreekt de standaardisatie grotendeels voor de lokale gesprekken, en blijft het onduidelijk hoe sommige instrumenten en het advies uit het assessment rapport door de selecteurs gebruikt worden. Wel bestaan er sinds juni 2024 nieuwe beoordelingsformulieren voor de eindgesprekken. Daarin worden competenties numeriek

beoordeeld (maar wordt er nog niet gescoord per vraag), en bestaan er beslisregels om per selecteur tot een persoonlijk oordeel te komen. Vervolgens vindt echter een beraadslaging plaats, wat wij afraden, omdat dit tot minder goede selectiebeslissingen leidt dan het gebruik van een beslisregel. Ook de afstemming tussen de lokale en landelijke selectie ontbreekt deels (onderzoeksvraag 9), omdat in de lokale selectie ook al gedeeltelijk competenties getoetst worden die eigenlijk pas in de eindgesprekken beoordeeld horen te worden.

Afsluitend valt te concluderen dat een paar redelijk simpele interventies, zoals het weglaten van onderdelen die niet bijdragen aan de doelen van de selectieprocedure, het standaardiseren van gesprekken en het toepassen van beslisregels, de selectieprocedure aanzienlijk zullen verbeteren. De implementatie van de laatste twee aspecten vergt enige voorbereiding, afstemming en waarschijnlijk advies en ondersteuning van experts op het gebied van personeelsselectie en assessment. Een wat grotere drempel is daarin waarschijnlijk de implementatie van gestandaardiseerde gesprekken bij de lokale gerechten. Als deze zaken eenmaal geïmplementeerd zijn, zijn ze echter gemakkelijk en efficiënt uit te voeren en te gebruiken en kost de procedure naar onze inschatting minder tijd en middelen dan nu het geval is, terwijl de effectiviteit toeneemt.

Dankwoord

We danken Jumana Tuama voor haar hulp bij het transcriberen van de gesprekken en het opstellen van de enquête voor de reacties van de kandidaten die deelnamen aan de selectieprocedure. Verder danken we alle personen die hebben meegewerkt aan dit onderzoek voor hun inzet.

Referenties

Anglim, J., Bozic, S., Little, J., & Lievens, F. (2018). Response distortion on personality tests in applicants: Comparing high-stakes to low-stakes medical settings. *Advances in Health Sciences Education*, *23*(2), 311–321. psyh. https://doi.org/10.1007/s10459-017-9796-8

Arkes, H. R., González-Vallejo, C., Bonham, A. J., Kung, Y.-H., & Bailey, N. (2010). Assessing the merits and faults of holistic and disaggregated judgments. *Journal of Behavioral Decision Making*, *23*(3), 250–270. https://doi.org/10.1002/bdm.655

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 417–439). Springer US. https://doi.org/10.1007/978-0-306-47630-3_19

Ashton, M. C., & Lee, K. (2009). An investigation of personality types within the HEXACO personality framework. *Journal of Individual Differences*, *30*(4), 181–187. https://doi.org/10.1027/1614-0001.30.4.181

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, *54*(2), 387–419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, *50*(3), 655–702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x

Chapman, D. S., & Zweig, D. I. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. *Personnel Psychology*, *58*(3), 673–702. https://doi.org/10.1111/j.1744-6570.2005.00516.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Dahlke, J. A., & Sackett, P. R. (2017). The relationship between cognitive-ability saturation and subgroup mean differences across predictors of job performance. *Journal of Applied Psychology*, *102*(10), 1403–1420. https://doi.org/10.1037/apl0000234.supp

Dahlke, J. A., & Sackett, P. R. (2022). On the assessment of predictive bias in selection systems with multiple predictors. *Journal of Applied Psychology*, *107*(11), 1995–2012. https://doi.org/10.1037/APL0000996

Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making*, *8*(5), 512–520.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment*, *17*(3), 254–270. https://doi.org/10.1111/j.1468-2389.2009.00468.x

Dipboye, R. L., Fontenelle, G. A., & Garner, K. (1984). Effects of previewing the application on interview process and outcomes. *Journal of Applied Psychology*, *69*(1), 118–128. https://doi.org/10.1037/0021-9010.69.1.118

Evers, A., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests. Utrecht: Nederlands Instituut voor Psychologen.* Nederlands Instituut voor Psychologen.

Ferguson, E., Sanders, A., O'Hehir, F., & James, D. (2000). Predictive validity of personal statements and the role of the five-factor model of personality in relation to medical training. *Journal of Occupational and Organizational Psychology*, *73*(3), 321–344. psyh. https://doi.org/10.1348/096317900167056

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The Clinical-Statistical Controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*(3), 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, *133*(2), 765–800. https://doi.org/10.1093/qje/qjx042

Hollwitz, J. C., & Pawlowski, D. R. (1997). The development of a structured ethical integrity Interview for pre-employment screening. *The Journal of Business Communication (1973)*, *34*(2), 203–219. https://doi.org/10.1177/002194369703400206

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*(2), 184–190. https://doi.org/10.1037/0021-9010.79.2.184

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment*, *21*(3), 264–276. https://doi.org/10.1111/ijsa.12036

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology. *International Journal of Selection and Assessment*, *22*(3), 297–309. https://doi.org/10.1111/ijsa.12078

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72–98. https://doi.org/10.1037/0033-2909.96.1.72

Kausel, E. E., Culbertson, S. S., & Madrid, H. P. (2016). Overconfidence in personnel selection: When and why unstructured interview information can hurt hiring decisions. *Organizational Behavior and Human Decision Processes*, *137*, 27–44. https://doi.org/10.1016/j.obhdp.2016.07.005

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, *98*(6), 1060–1072. https://doi.org/10.1037/a0034156

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology*, *99*(1), 38–47. https://doi.org/10.1037/a0034147

Lievens, F., Buyse, T., Sackett, P. R., & Connelly, B. S. (2012). The effects of coaching on situational judgment tests in high-stakes selection. *International Journal of Selection and Assessment*, *20*(3), 272–282. https://doi.org/10.1111/j.1468-2389.2012.00599.x

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology*, *96*(5), 927–940. https://doi.org/10.1037/a0023496

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*(4), 599–616. https://doi.org/10.1037/0021-9010.79.4.599

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000

Meijer, R. R., Neumann, M., Hemker, B. T., & Niessen, A. S. M. (2020). A tutorial on mechanical decision-making for personnel and educational selection. *Frontiers in Psychology*, *10*, Article 3002. https://doi.org/10.3389/fpsyg.2019.03002

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683–729. https://doi.org/10.1111/J.1744-6570.2007.00089.X

Morris, S. B., Daisley, R. L., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationship between individual assessments and job performance. *Journal of Applied Psychology*, *100*(1), 5–20. https://doi.org/10.1037/a0036938

Mosel, J. N., & Goheen, H. W. (1958). The Validity of the Employment Recommendation Questionnaire in Personnel Selection: I. Skilled Traders. *Personnel Psychology*, *11*(4), 481–490. https://doi.org/10.1111/j.1744-6570.1958.tb00034.x

Murphy, K. R. (2019). Understanding how and why adding valid predictors can decrease the validity of selection composites: A generalization of Sackett, Dahlke, Shewach, and Kuncel (2017). *International Journal of Selection and Assessment*, *27*(3), 249–255. https://doi.org/10.1111/ijsa.12253

Neumann, M., Hengeveld, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2022). Education increases decision-rule use: An investigation of education and incentives to improve decision making. *Journal of Experimental Psychology: Applied*, *28*(1), 166–178. https://doi.org/10.1037/xap0000372

Neumann, M., Niessen, A. S. M., Linde, M., Tendeiro, J. N., & Meijer, R. R. (2024). "Adding an egg" in algorithmic decision making: Improving stakeholder and user perceptions, and predictive validity by enhancing autonomy. *European Journal of Work and Organizational Psychology*, *33*(3), 245–262. https://doi.org/10.1080/1359432X.2023.2260540

Niessen, A. S. M., Hurks, P. M., Neumann, M., & Meijer, R. R. (2024). *Handleiding voor het opstellen en gebruiken van beslisregels*.

Niessen, A. S. M., & Meijer, R. R. (2019). Evaluatie van het selectieproces voor de rio-opleiding. *Research Memoranda*, *2*. https://www.rechtspraak.nl/SiteCollectionDocuments/evaluatie-van-het-selectieproces-voor-de-rio-opleiding.pdf

Niessen, A. S. M., Meijer, R. R., & Neumann, M. (2019). Mis(ver)standen in de selectiepraktijk: Een goed verhaal maakt nog geen goede beslissing. *De Psycholoog*, *54*(11), 46–55.

Niessen, A. S. M., & Neumann, M. (2022). Using personal statements in college admissions: An investigation of gender bias and the effects of increased structure. *International Journal of Testing*, *22*(1), 5–20. https://doi.org/10.1080/15305058.2021.2019749

Nolan, K. P., & Highhouse, S. (2014). Need for autonomy and resistance to standardized employee selection practices. *Human Performance*, *27*(4), 328–346. https://doi.org/10.1080/08959285.2014.929691

Odijk, J. M., Hiemstra, A. M. F., & Born, M. Ph. (2024). *Kansengelijkheid in selectie en assessment: Een labstudie naar de effectiviteit van structurering om kansengelijkheid voor etnisch-cultureel diverse sollicitanten te bevorderen*.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, *35*(1), 1–62. https://doi.org/10.1111/j.1744-6570.1982.tb02184.x

Rupp, D. E., Hoffman, B. J., Bischof, D., Byham, W., Collins, L., Gibbons, A., Hirose, S., Kleinmann, M., Kudisch, J. D., Lanik, M., Jackson, D. J. R., Kim, M., Lievens, F., Meiring, D., Melchers, K. G., Pendit, V. G., Putka, D. J., Povah, N., Reynolds, D., … Thornton, G. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

Rupp, D. E., Song, Q. C., & Strah, N. (2020). Addressing the so-called validity–diversity trade-off: Exploring the practicalities and legal defensibility of pareto-optimization for reducing adverse impact within personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *13*(2), 246–271. https://doi.org/10.1017/iop.2020.19

Sackett, P. R., Dahlke, J. A., Shewach, O. R., & Kuncel, N. R. (2017). Effects of predictor weighting methods on incremental validity. *Journal of Applied Psychology*, *102*(10), 1421–1434. https://doi.org/10.1037/apl0000235

Sackett, P. R., Demeke, S., Bazian, I. M., Griebie, A. M., Priest, R., & Kuncel, N. R. (2024). A contemporary look at the relationship between general cognitive ability and job performance. *Journal of Applied Psychology*, *109*(5), 687–713. https://doi.org/10.1037/apl0001159.supp

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*(3), 707–721. https://doi.org/10.1111/j.1744-6570.1997.tb00711.x

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*(10), 1435–1447. https://doi.org/10.1037/apl0000236

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*(11), 2040–2068. https://doi.org/10.1037/APL0000994

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*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., & Ramsay, L. J. (2004). The impact of justice and self-serving bias explanations of the perceived fairness of different types of selection tests. *International Journal of Selection and Assessment*, *12*(1–2), 160–171. https://doi.org/10.1111/j.0965-075X.2004.00271.x

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, *65*(3), 445–494. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Snarey, J. R. (1985). Cross-cultural universality of social-moral development: A critical review of Kohlbergian research. *Psychological Bulletin*, *97*(2), 202–232. https://doi.org/10.1037/0033-2909.97.2.202

Steiner, D. D., & Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology*, *81*(2), 134–141. https://doi.org/10.1037/0021-9010.81.2.134

te Nijenhuis, J., & van der Flier, H. (1999). Bias research in The Netherlands: Review and implications. *European Journal of Psychological Assessment*, *15*(2), 165–175. https://doi.org/10.1027//1015-5759.15.2.165

van Andel, C. E. E., Born, M. P., Themmen, A. P. N., & Stegers-Jager, K. M. (2019). Broadly sampled assessment reduces ethnicity-related differences in clinical grades. *Medical Education*, *53*(3), 264–275. https://doi.org/10.1111/medu.13790

Van Iddekinge, C. H., Arnold, J. D., Frieder, R. E., & Roth, P. L. (2019). A meta-analysis of the criterion-related validity of prehire work experience. *Personnel Psychology*, *72*(4), 571–598. https://doi.org/10.1111/peps.12335

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology*, *97*(3), 499–530. https://doi.org/10.1037/a0021196

Watson, G., & Glaser, E. M. (2010). *Watson-Glaser II Critical Thinking Appraisal: Technical manual and user's guide.* NCS Pearson, Inc.

Wingate, T. G., & Bourdage, J. S. (2024). What are interviews for? A qualitative study of employment interview goals and design. *Human Resource Management*, *63*(4), 555–580. https://doi.org/10.1002/hrm.22215

Yu, M. C., & Kuncel, N. R. (2020). Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions*, *6*(2), 1–10. https://scholarworks.bgsu.edu/pad/vol6/iss2/2

Yu, M. C., & Kuncel, N. R. (2022). Testing the value of expert insight: Comparing local versus general expert judgment models. *International Journal of Selection and Assessment*, *30*(2), 202–215. https://doi.org/10.1111/IJSA.12356

Bijlagen

Bijlage 1 – Specifieke onderzoeksvragen voor de evaluatie

Tabel B1 *Specifieke onderzoeksvragen voor de evaluatie*

Het selectieproces
1. Verloopt het selectieproces zoals vastgelegd in het Handboek LSR
2. Zo ja, wat gaat goed en wat kan eventueel verbeterd worden?
3. Is voor alle betrokkenen (kandidaten en selecteurs) duidelijk wat de formele eisen zijn, wat de definitie van die eisen is en hoe die in de praktijk worden toegepast?
4. Vindt de selectie plaats op basis van het competentieprofiel van rechter of raadsheer?
5. Draagt de inrichting van de selectieprocedure ertoe bij dat potentieel geschikte kandidaten ook deelnemen aan het selectieproces?
6. Draagt de inrichting van de selectieprocedure (in te zetten instrumenten, de volgorde waarin de instrumenten worden ingezet en de weging van die instrumenten) ertoe bij dat geschikte kandidaten ook succesvol door de selectieprocedure komen?
7. Op welke wijze worden leden van de landelijke selectiecommissie (LSR) geselecteerd en opgeleid om de taak van selecteur uit te voeren?
8. Hoe zijn de lokale selectiecommissies (bij de gerechten) samengesteld en hoe worden de selecteurs geselecteerd en opgeleid om de taak van selecteur uit te voeren?
9. Is het lokale proces van selectie goed afgestemd op het landelijke proces van selectie?
10. Wat is de doorlooptijd van de selectieprocedure en zou (met behoud van kwaliteit) nog versnelling mogelijk zijn?
11. Wat is de uitval per fase en wat is de reden van uitval? Hoe had die uitval voorkomen kunnen worden?
De werking van de selectie-instrumenten
12. Zijn de selectie-instrumenten die worden ingezet betrouwbaar en valide?
13. In hoeverre dragen de selectie-instrumenten die worden ingezet bij aan de voorspellende waarde van de selectie?
14. Zijn alle selectie-instrumenten die worden ingezet waardevrij?
15. Zou de inzet van andere selectie-instrumenten kunnen bijdragen aan het vergroten van de diversiteit onder rio's?
16. Zou de onderlinge weging van de selectie-instrumenten kunnen bijdragen aan het afvallen van kandidaten met een niet-westerse achtergrond?
17. Hoe voeren de selecteurs bij LSR en de lokale selectiecommissie in de praktijk hun taak uit?
18. Kunnen we op een andere manier rekening houden met kandidaten met een beperking, zoals dyslexie?
19. Als een kandidaat de selectieprocedure succesvol doorloopt, wat betekent dit dan voor de kans dat de kandidaat de rio-opleiding succesvol zal afronden?
20. Wat is de voorspellende waarde van de selectie? Komt de voorspelling over een competentie bij de selectie overeen met de feitelijke prestaties gedurende de rio-opleiding?
21. (Aanvullend) In hoeverre worden beslisregels toegepast en geaccepteerd door gebruikers?

Bijlage 2 – Vragenlijst over ervaringen van oud-kandidaten en informatie over de steekproef

Tabel B2 *Vragenlijst over ervaringen van oud-kandidaten*

Vragen over de selectieprocedure als geheel.
Deelnemers gaven antwoorden op een 5-punt schaal (1 = zeer mee oneens, 5 = zeer mee eens).

1. In het algemeen was ik tevreden met de selectieprocedure.
2. Deelname aan de selectieprocedure was een positieve ervaring.
3. De beslissing die over mij genomen is naar aanleiding van de selectieprocedure was een eerlijke beslissing.
4. De personen die zijn aangenomen op basis van de selectieprocedure verdienen die beslissing.
5. Ik wist van tevoren hoe de selectieprocedure eruit zou zien.
6. De inhoud van de selectieprocedure was gepast.
7. De resultaten van de selectieprocedure boden aanknopingspunten voor mijn ontwikkeling als rechter in opleiding.

Vragen over de specifieke onderdelen van de selectieprocedure.
Deze vragen werden over elk specifiek onderdeel gesteld. In dit voorbeeld werden er vragen over de briefselectie gesteld.

1. Hoe beoordeelt u de effectiviteit van de briefselectie om gekwalificeerde rechters te identificeren? (1 = *zeer ineffectief*, 5 = *zeer effectief)*.
2. Als u niet zou worden aangenomen op basis van de briefselectie, wat zou u dan van de eerlijkheid van deze procedure vinden? (1 = *zeer oneerlijk*, 5 = *zeer eerlijk*).
3. De inhoud van de briefselectie was duidelijk gerelateerd aan de functie van rechter. (1 = *zeer mee oneens*, 5 = *zeer mee eens*).
4. De briefselectie laat belangrijke kwaliteiten van personen zien die hen van anderen onderscheiden. (1 = *zeer mee oneens*, 5 = *zeer mee eens*.
5. De briefselectie is een logische methode om te gebruiken voor het identificeren van gekwalificeerde rechters. (1 = *zeer mee oneens*, 5 = *zeer mee eens*).
6. Ik kon mijn vaardigheden en capaciteiten goed laten zien op de briefselectie.
7. De inhoud van de selectieprocedure was gepast. (1 = *zeer mee oneens*, 5 = *zeer mee eens*).
8. De resultaten van de selectieprocedure boden aanknopingspunten voor mijn ontwikkeling als rechter in opleiding. (1 = *zeer mee oneens*, 5 = *zeer mee eens*).

Informatie over de steekproef

De vragenlijst werd aan 391 personen verstuurd. In totaal makten er 205 respondenten de vragenlijst compleet af, waarvan 38% mannelijk, 61% vrouwelijk, en 1% anders waren. De meeste respondenten (43%) waren tussen 40 en 49 jaar oud en 98% van de respondenten gaven aan dat zij in Nederland geboren waren. Van de respondenten had 16% zich aangemeld voor een plaats voor rio's met twee tot zes jaar werkervaring, had 63% zich aangemeld voor een plaats voor rio's met minstens zes jaar werkervaring en had 21% zich aangemeld voor een plaats voor rio's met minstens 10 jaar werkervaring. Verder was de gemiddelde algemene werkervaring van de respondenten ten tijde van hun sollicitatie 16 jaar (*SD* = 8.1, variërend van 2.5 tot 40 jaar). 70% van de respondenten had één keer gesolliciteerd, terwijl 22% twee keer solliciteerde. Zesentachtig percent van de respondenten solliciteerde tussen 2020 en 2022. Respondenten hebben eerst vragen over de selectieprocedure als geheel beantwoord. Daarna werden vragen over de verschillende onderdelen van de procedure zoals de analytische test en het assessment gesteld.

Bijlage 3 – Operationalisatie van prestatie in de rio-opleiding en technische details

De prestaties tijdens de rio-opleiding worden door leden van een beoordelingscommissie op basis van het portfolio van een kandidaat beoordeeld. Daarvoor wordt gebruik gemaakt van een scoringslijst. De prestaties worden met 69 items beoordeeld. Deze items horen bij verschillende competenties die zijn weergegeven in Tabel B3.

Tabel B3 *Beoordeelde competenties in de rio-opleiding*

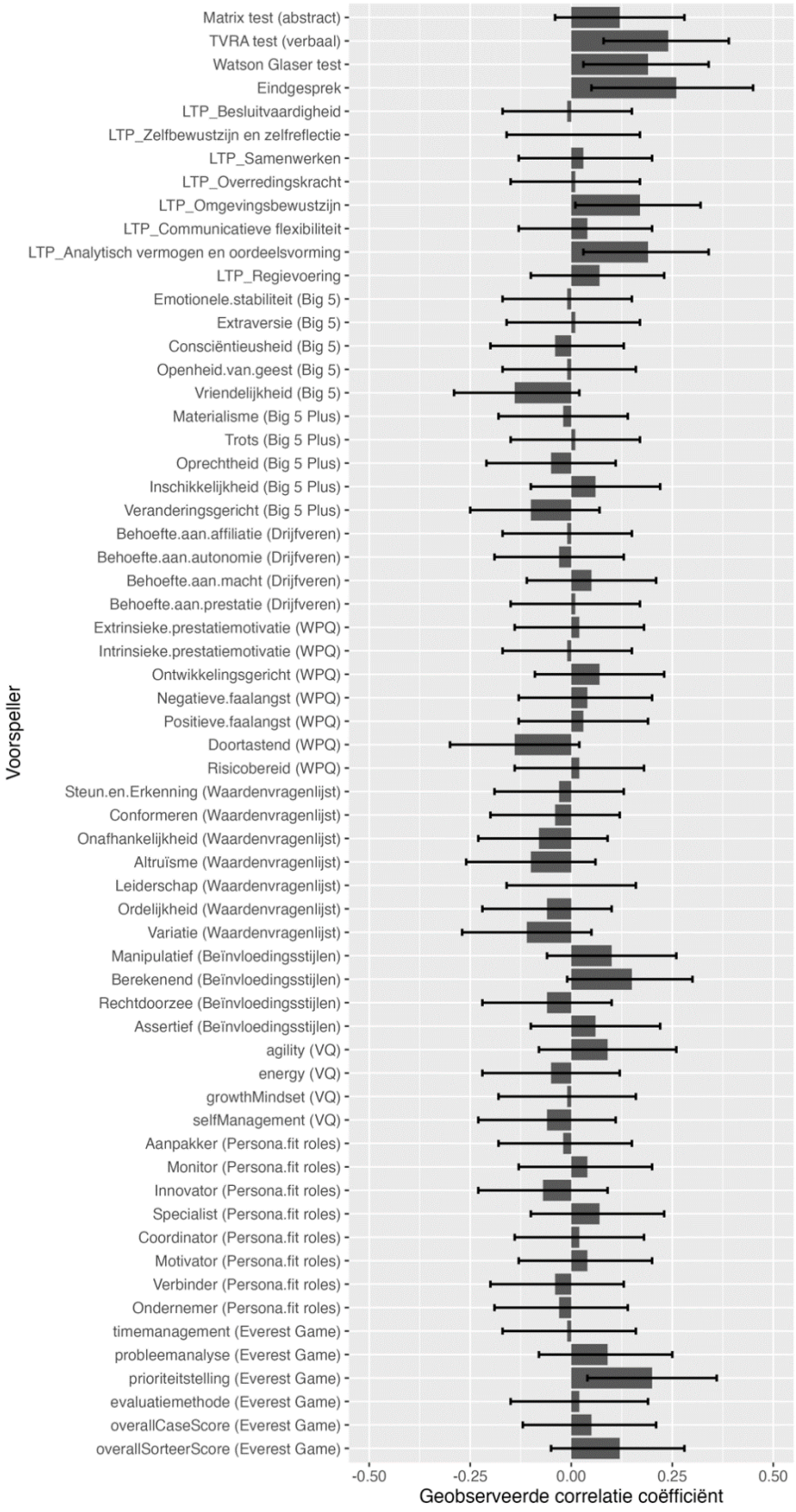
Competenties
Omgevingsbewustzijn
Analytisch vermogen en Oordeelsvorming
Luisteren
Overredingskracht
Regievoering
Samenwerken
Zelfvertrouwen en Authenticiteit
Flexibiliteit
Zelfbewustzijn
Leervermogen en Zelfreflectie
Besluitvaardigheid
Juridische kennis

De som van de scores op deze 69 items resulteert in een totaalscore, welke wordt vergeleken met een cesuurscore. Enkele items zijn kritisch, wat betekent dat onvoldoende scores op deze items niet gecompenseerd kunnen worden door hogere scores op andere items. Verder worden er verschillende taakniveaus gehanteerd. Het taakniveau geeft aan op welk niveau een rio globaal functioneert. Deze omstandigheden zorgen ervoor dat de totaalscores van rio's niet zomaar met elkaar vergeleken kunnen worden. Een analyse per taakniveau was niet mogelijk vanwege te kleine steekproefgroottes. Daarom is besloten alleen de groep rio's te analyseren die een eindbeoordeling hebben gekregen en ook aan alle kritische criteria voldeden. Deze groep rio's worden per definitie op het hoogste taakniveau (D) beoordeeld en hebben altijd dezelfde cesuurscore. Dit maakt het mogelijk de totaalscore tussen rio's te vergelijken. Daarom werd in deze analyses de som van de beoordelingen op alle 69 items als prestatiemaat gebruikt.

Om de criteriumvaliditeit te onderzoeken werd een steekproef van 81-148 rio's gebruikt, omdat voor deze groep zowel scores op de selectie-instrumenten als ook scores van de prestaties in de rio-opleiding beschikbaar waren. In deze steeproef (N = 148) waren 51% mannelijk en was de gemiddelde leeftijd 45 jaar (SD = 8). Een correlatie tussen de scores op de selectie-instrumenten en prestaties in de rio-opleiding kan alleen voor rio's berekend worden die ook geselecteerd zijn. Rio's die bij de selectie zijn afgevallen zijn immers nooit aan de opleiding begonnen en hebben daarom ook geen beoordeling gekregen. Dit zorgt voor *restrictie in range*, wat resulteert in een onderschatting van de correlatie tussen voorspeller en criterium. Om dit te voorkomen hebben wij de *geobserveerde correlaties* voor restrictie in range gecorrigeerd (*gecorrigeerde correlaties*). Hiervoor werd de multivariate correctie uit het *psychmeta* package in *R* en de functie *correct_matrix_mvrr* gebruikt, met als

input alle selectie-instrumenten die in Figuur 1 zijn weergegeven. De gecorrigeerde correlaties zijn in Figuur 1 in de tekst weergegeven. De geobserveerde correlaties zijn hieronder in Figuur B1 weergegeven.

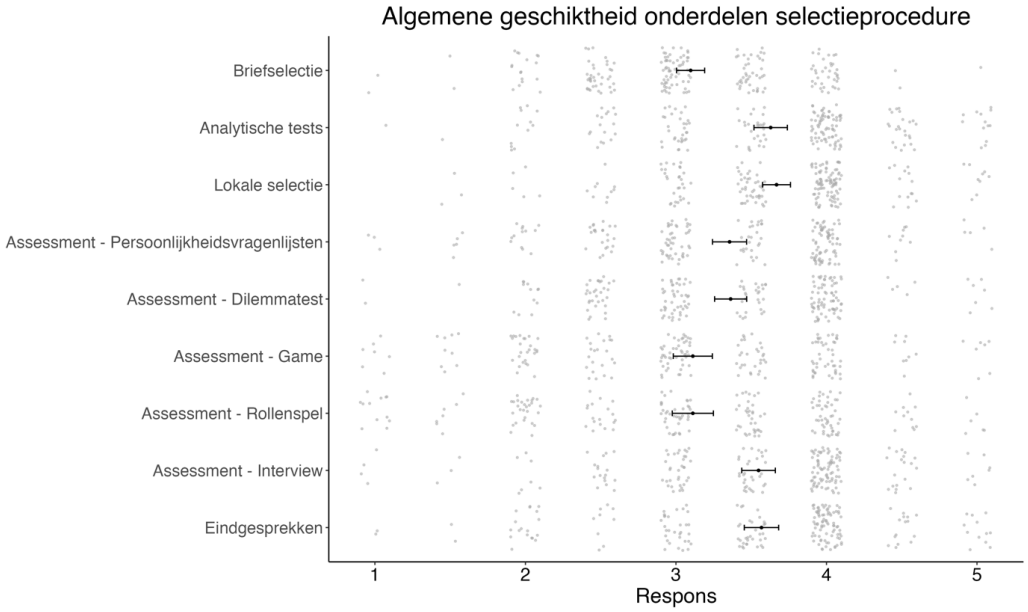
Figuur B1 *Geobserveerde correlaties tussen onderdelen uit de selectieprocedure en de totaalscore*



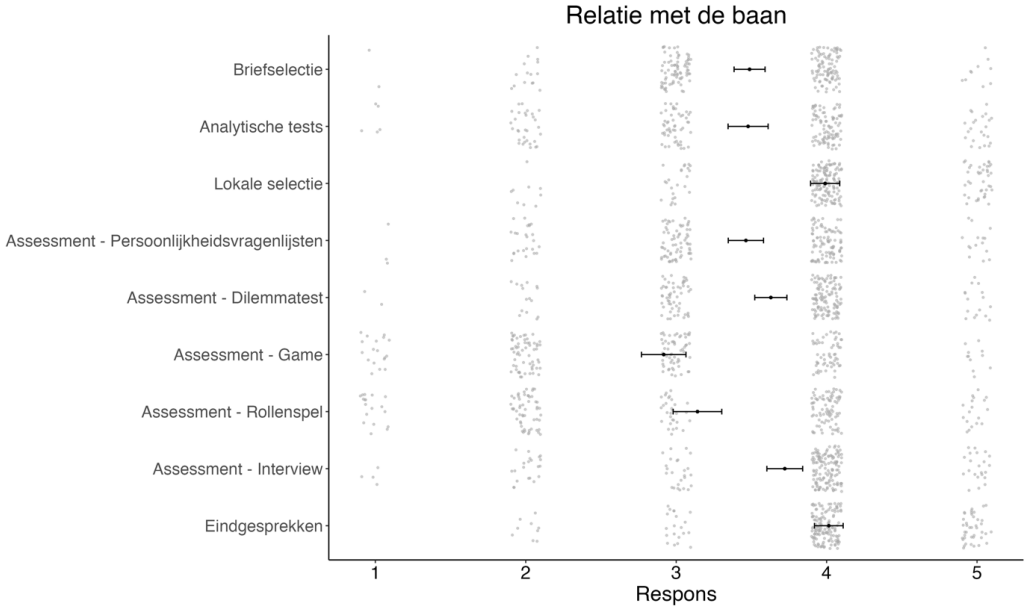
Noot. N = 81-148. De error bars geven het 95% betrouwbaarheidsinterval weer.

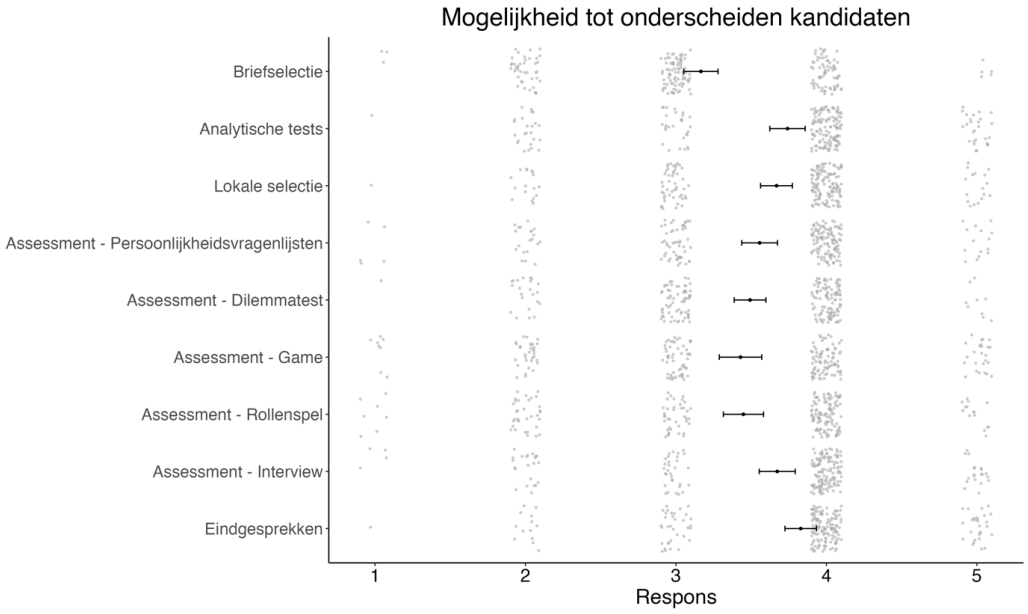
Bijlage 4 – Gemiddelde beoordelingen van de selectieprocedure dooroud-kandidaten

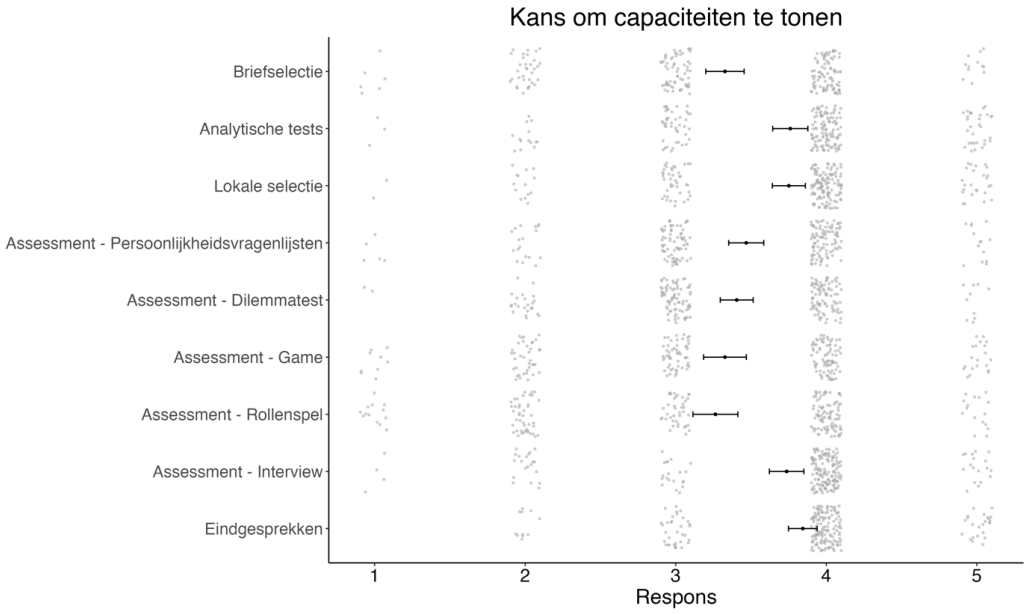
Figuur B2 *Gemiddelden en 95% betrouwbaarheidsintervallen voor algemene geschiktheid van de verschillende onderdelen van de selectieprocedure voor de rio-opleiding*

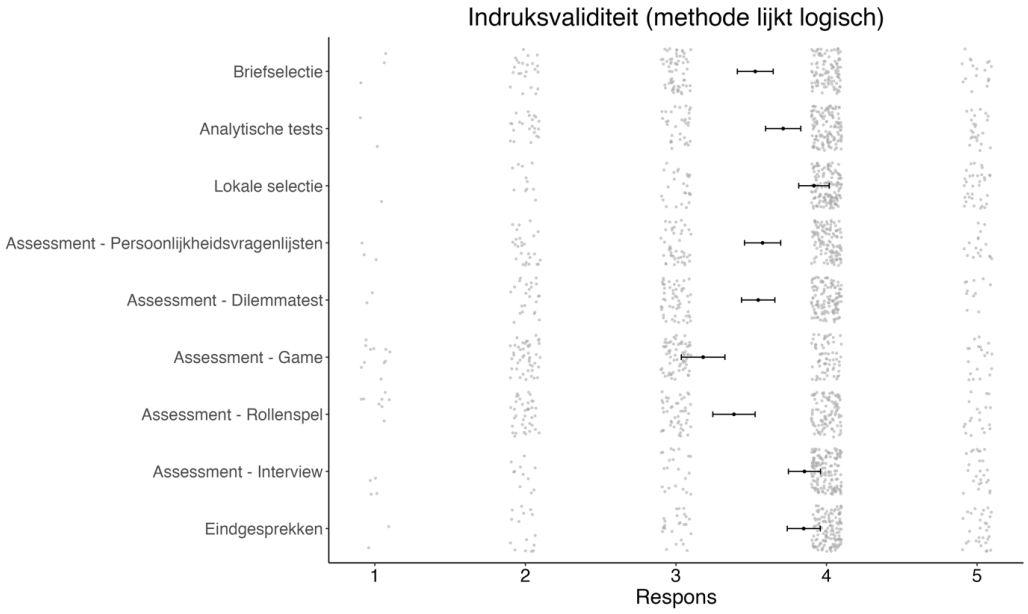


Figuur B3 *Gemiddelden en 95% betrouwbaarheidsintervallen voor de procedurele rechtvaardigheidscriteria van de verschillende onderdelen van de selectieprocedure voor de rio-opleiding*




Mogelijkheid tot onderscheiden kandidaten
Briefselectie
Analytische tests
Lokale selectie
Assessment - Persoonlijkheidsvragenlijsten
Assessment - Dilemmatest
Assessment - Game
Assessment - Rollenspel
Assessment - Interview
Eindgesprekken
1
2
3
4
5
Respons


Kans om capaciteiten te tonen
Briefselectie
Analytische tests
Lokale selectie
Assessment - Persoonlijkheidsvragenlijsten
Assessment - Dilemmatest
Assessment - Game
Assessment - Rollenspel
Assessment - Interview
Eindgesprekken
1
2
3
4
5
Respons


Indruksvaliditeit (methode lijkt logisch)
Briefselectie
Analytische tests
Lokale selectie
Assessment - Persoonlijkheidsvragenlijsten
Assessment - Dilemmatest
Assessment - Game
Assessment - Rollenspel
Assessment - Interview
Eindgesprekken
1
2
3
4
5
Respons

Bijlage 5 – Vragen aan oud-kandidaten over hun voorkeur voor beslisregels

De respondenten zijn gevraagd in hoeverre zij open staan voor het gebruik van beslisregels om zo wel informatie binnen een gesprek, als informatie uit verschillende instrumenten te combineren, om een selectie beslissing te nemen. Voor de eerste vraagstelling werden respondenten als volgt geïnformeerd:

Wetenschappelijk onderzoek heeft aangetoond dat er betere interviewbeoordelingen gemaakt worden wanneer selecteurs antwoorden van de kandidaten kwantitatief scoren en vervolgens middels een beslisregel combineren in plaats van middels groepsoverleg op basis van hun menselijke oordelen. Met een beslisregel wordt iedere kandidaat op dezelfde manier beoordeeld. Voor de eindgesprekken zou een simpele beslisregel kunnen zijn: Selecteurs beoordelen uw interview antwoorden onafhankelijk van elkaar en tellen de beoordelingen daarna bij elkaar op. De kandidaten met de hoogste scores worden, gebaseerd op de eindgesprekken, als meest geschikt geacht. In de huidige selectieprocedure wordt er geen beslisregel gebruikt om tot een interviewoordeel te komen. In plaats daarvan wordt een interviewbeoordeling middels groepsoverleg bepaald. Vergeleken met een interviewbeoordeling middels groepsoverleg, hoe zou u een interviewbeoordeling via een beslisregel vinden?

1. Gesprekpartners in de eindgesprekken integreren kwantitatieve scores op interviewantwoorden **via groepsoverleg** waarbij het mogelijk is dat verschillende kandidaten op verschillende manieren beoordeeld worden.
2. Gesprekpartners in de eindgesprekken integreren kwantitatieve scores op interviewantwoorden **via een beslisregel** (bijvoorbeeld optellen van scores) waarbij het NIET mogelijk is dat verschillende kandidaten op verschillende manieren beoordeeld worden.

Daarna werd respondenten het volgende vertelt en gevraagd:

Een beslisregel kan ook gebruikt worden om informatie uit meerdere instrumenten zoals het assessment en de eindgesprekken te combineren en zo uiteindelijk tot een selectiebeslissing te komen. In de huidige selectieprocedure wordt er geen beslisregel gebruikt om tot een selectiebeslissing te komen. In plaats daarvan wordt een selectiebeslissing middels groepsoverleg bepaald. Als u zou mogen kiezen, welke van de twee onderstaande manieren zou u voorkeur hebben?

1. Selecteurs integreren het eindoordeel uit een interview en het assessment advies **via groepsoverleg** waarbij het mogelijk is dat verschillende kandidaten op verschillende manieren beoordeeld worden.
2. Selecteurs integreren het eindoordeel uit een interview en het assessment advies **via een beslisregel** (bijvoorbeeld optellen van scores) waarbij het NIET mogelijk is dat verschillende kandidaten op verschillende manieren beoordeeld worden.