



Voordelen Verify:

- ✓ **Betere testen: hogere betrouwbaarheid en validiteit**
- ✓ **Functiegericht testen i.p.v. opleidingsgericht**
- ✓ **Grote representatieve referentiegroepen**
- ✓ **Getoetst op 'fairness'**
- ✓ **Wereldwijd beschikbare testversies en referentiegroepen**
- ✓ **Efficiënte afname met hoge face validiteit**

Toelichting op Verify

WA 14 09 10

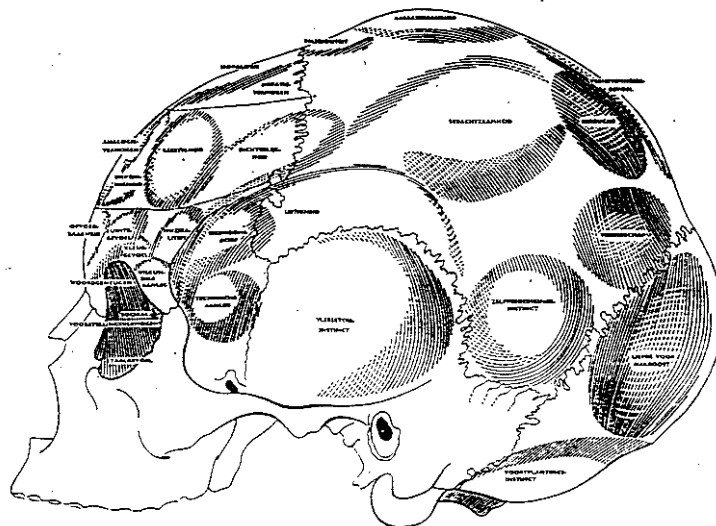
Doel van dit document

Recentelijk zijn de Verify testen geïntroduceerd als vervanger van de Occupational Testen. In dit document wordt nader ingegaan op de overeenkomsten tussen de testseries en de merites van Verify. Voor een verslag van het onderzoek naar Verify test versies wordt verwezen naar de 'Technical Manual-appendix/userguide' van Verify, die te downloaden is van de SHL site (www.SHL.com). U kunt ook terecht met vragen bij uw accountmanager.

Hieronder besteden we aandacht aan het doel van de Verify testreeks, vanuit een historisch kader. De belangrijke peiler van de Verify testserie, de latente trek als schaal model, volgt daarna. Ook gaan we in op het gebruik van Verify, waarbij het 'Solutions' aspect aan de orde komt. Verify is gebaseerd op het gebruik van de item response theorie (IRT). IRT staat toe de belangrijke informatie te vergaren die nodig is om Verify testen samen te stellen en zinvol te gebruiken. Daar waar nodig zullen wij dit toelichten in dit document.

Doel van Verify – historisch kader

In het begin van de vorige eeuw is de psychologie als wetenschap ontstaan. Het meten van individuele verschillen is een belangrijke peiler in deze wetenschap. Bij de start van de differentiële diagnostiek is afgesproken dat alle individuele kenmerken een Normaal Verdeling volgen – conform de leer van Galton. De meting van een individu krijgt pas betekenis bij een vergelijking met de resultaten van andere personen. Deze afspraak is essentieel voor de meting van psychologische kenmerken, zoals ook capaciteiten of intelligentie metingen. Rond 1900 waren de meetinstrumenten echter nog niet heel gesofisticeerd. Paul Broca (1824-1880) en Sir Francis Galton (1822-1911) waren één van de eerste wetenschappers die zich bezig hielden met het meten van intelligentie. Dat deden ze door de schedelomvang van mensen te meten. Hoe groter de schedel, hoe slimmer de mens, was hun uitgangspunt. Tegenwoordig is hun methodiek achterhaald en wordt deze niet meer als test ingezet en vormt een onderdeel van de geschiedenis van de psychologie.



'Frenologie' of schedelmeting begin 1900

In de wetenschap bedoelt men met intelligentie/capaciteiten meestal cognitieve intelligentie. Een omschrijving is: "Het geheel van cognitieve of verstandelijke vermogens dat nodig is om

kennis te verwerven, en daar op een goede wijze gebruik van te maken, ten einde problemen op te lossen die een vast omschreven doel en structuur hebben"¹. In gewone mensentaal kan worden gezegd dat intelligentie gaat om hoe slim of pienter iemand is. De eerste objectieve testen voor de meting van capaciteiten bij volwassenen zijn rond 1915 ontwikkeld. Nadien is een groot scala aan testen gepubliceerd. Een test kan men zien als een gestandaardiseerd gedragsvoorbeeld dat weergegeven wordt met een numerieke schaal of categoriesysteem². De aanleg of het potentieel voor een specifieke taak wordt gemeten. Individuele prestaties worden afgezet tegen een voor het meetdoel relevante vergelijkings- of normgroep.

Er is veel onderzoek verricht naar de waarde van capaciteitentesten. Op basis hiervan wordt wel gesteld dat deze tests de beste voorspellers van succes in arbeidsorganisaties zijn, en gemiddelde validiteiten behalen in de orde van .50³. Bij deze algemene uitspraak kunnen wat kanttekeningen worden gemaakt⁴. Zo zijn capaciteitentesten niet de enige relevante voorspellers. Persoonlijkheid, motivatie, kennis en ervaring zijn vaak even zo goed van belang. Wanneer gekeken wordt naar specifieke competenties hebben capaciteitentesten soms een hogere dan wel een lagere validiteit. Wel wijst onderzoek uit dat zij voor de meeste posities, en dus ook de hogere posities, in arbeidsorganisaties relevant zijn om mee te nemen in het voorspellen van de mate van succes in het werk.

Met als doel een zo optimaal mogelijke capaciteitentest te maken die aansluit bij de huidige maatschappij en wereld van werken is de constructie van Verify ter hand genomen. De psychometrische eigenschappen, betrouwbaarheid en validiteit, staan voorop maar evenzo goed de praktische bruikbaarheid. Zo ook de toepasbaarheid online, in een omgeving door de deelnemer zelf gekozen – zonder dat fraude het resultaat beïnvloedt. Online toepasbaarheid is niet alleen makkelijk voor organisaties. Het leidt ook tot een grotere acceptabiliteit, aanvaardbaarheid, bij deelnemers.

Samenstelling van Verify

Is intelligentie één ding of bestaat het uit verschillende facetten? Hierover wordt al jarenlang discussie gevoerd. Charles Spearman nam de positie in dat er slechts één algemene intelligentiefactor is waarop alle cognitieve vaardigheden zijn terug te voeren. Anderen, zoals Vernon en Guilford, zijn het hier niet mee eens. Zij beweren dat er verschillende intelligentiefactoren zijn die (redelijk) los van elkaar staan. Er zijn door hen intelligentie modellen samengesteld die bestaan uit meer factoren, variërend van 4 tot 150 factoren.

Ook hier bepaalt het doel welk model het meest zinvol is in praktijk. Welke prestaties wil men voorspellen, en welke factor(en) blijkt uit onderzoek het meest valide hiervoor te zijn? Capaciteitentesten kunnen validiteiten behalen in de orde van .50, maar dit hangt uiteraard af van de betreffende criterium-predictor combinatie. Met één heel specifieke intelligentiefactor voorspelt men in de beoogde doelfunctie veelal ook een specifiek deel van prestaties. Er is een beperkter meetdomein dan in geval van een bredere intelligentiefactor die op zijn beurt voor een breder gebied wellicht een wat lagere validiteit heeft. Bij de voorspellingstaak is het ook relevant om na te gaan welke alternatieven er zijn om bepaalde prestatie onderdelen valide te meten. Oftewel: hoe breder het te voorspellen gedragsdomein des te lastiger is dit te dekken met één specifieke meting (die betrouwbaar, valide en praktisch is). Er is aangetoond dat een gedifferentieerde benadering, met name als deze is verbonden aan een goede analyse van de competenties die voor een functie nodig zijn, relevanter en sterker is dan de algemene capaciteiten (G-) factor. Ook geldt: hoe specifieker de intelligentiefactor, des te minder deze voorspelt van een breed functiedomein. De SHL- te

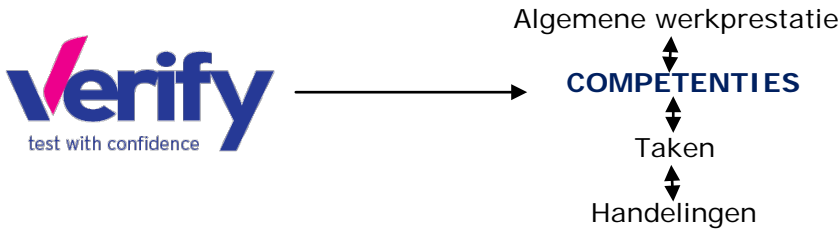
¹ Resing, W., & Drenth, P. (2007). *Intelligentie: weten en meten*. Amsterdam: Uitgeverij Nieuwezijds

² Cronbach, L.J. (1984). *Essentials of psychological testing (4th edn)*, New York: Harper Row

³ Schmidt, F., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262-274.

⁴ Lievens, F (2004). De voorspellende validiteit van selectietechnieken: bedenkingen vanuit de ivoren toren. HRM-net en www.allesoverassessmen.nl.

beoordelen, op een bepaald moeilijkheidsniveau. De Verify testen kunnen worden geplaatst op het niveau van voorspelling van bredere competenties, zoals problemen oplossen, oordelen, etc. Uit onderzoek is gebleken dat de Verify testen een hoge voorspellende waarde hebben voor dergelijke competentiegebieden.

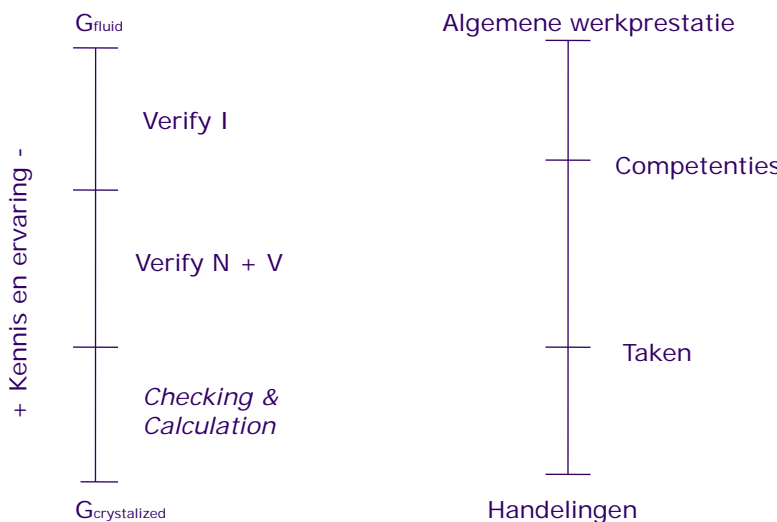


Meetdomein van Verify

Bewezen validiteit en betrouwbaarheid

Door hun meetbereik zijn Verify testen breed inzetbaar voor een groot scala aan werkgebieden. Uniek is dat de waarde van Verify op zeer grote steekproeven getoetst is. Hierdoor kunnen wij stellen dat de betrouwbaarheid en validiteit van Verify goed zijn, vergelijkbaar aan de waarden die Schmidt en Hunter noemen in hun overzichtstudie! Er zijn verschillende studies naar de predictieve validiteit uitgevoerd (Verify user & technical guide). In deze studies zijn organisaties uit verschillende sectoren vertegenwoordigd, waaronder bank- en verzekeringswezen, onderwijs, professional services, productie. De criteria zijn competentiebeoordelingen door managers. Voor de verbale test geldt een gemiddelde validiteit van .50 (totale steekproef 548), en voor de numerieke test .39 (N=760). Beide testen correleren relatief laag en een gezamenlijk gebruik leidt tot validiteiten van .60. De betrouwbaarheden zijn vergelijkbaar aan die van andere capaciteitentesten (interne consistentie tegen .80 en test-hertest ca. .90). Aangezien de Verify testen een geringe afname duur kennen, is de rendabiliteit van deze testen zeer hoog.

Algemene Opbouw



Algemene opbouw Verify testserie

Verify Vergelijkingsgroepen: predictief in plaats van retrospectief

De vergelijkingsgroepen voor Verify zijn gebaseerd op grote groepen van mensen die in arbeidsorganisaties werken. Wij wijzen hierbij op het onderscheid 'criterion vs norm referenced testing'⁵. Bij criterion referenced testing gaat men na wat een persoon kan, hoe goed iemand presteert ten opzichte van een bepaald criterium of niveau. De vraag is "in welke mate beheerst iemand een bepaalde taak?" Bij norm referenced testing meet men de plaats van een persoon in een bepaalde groep. De prestaties van de groep zijn het vergelijkingsuitgangspunt. Testen voor schoolprestaties zijn vaak criterion referenced; capaciteitentesten zijn veelal norm referenced. De validiteit van de norm referenced testen hangt logischerwijs sterk af van de relevantie van de testen relevant voor de beoogde (werk)prestaties. Voor de Verify testen is deze validiteit daadwerkelijk aangetoond.

Bij capaciteitentesten hanteert men van oudsher vaak een groepsindeling in opleiding. Bijvoorbeeld een groep mensen met afgeronde WO-opleiding of een groep met HBO diploma. De laatste jaren staat een dergelijke indeling sterk onder druk. Men kan zich afvragen hoe indicatief de indeling naar algemene opleidingscertificaten indicatief is voor cognitieve vermogens. Binnen opleidingsclusters is er een grote variatie aan opleidingen en niveau. Op algemeen niveau blijken bijvoorbeeld WO en HBO groepen zich niet te onderscheiden qua prestaties op capaciteitentesten: de verschillen zijn veelal niet groter dan een halve standaarddeviatie. Dit resulteert in d-waarden rond .2, een 'small effect', er wordt nog geen 1% van scorevariantie verklaard door groepslidmaatschap⁶. Klaarblijkelijk zijn algemene prestatieverschillen tussen WO en HBO eerder toe te schrijven aan andere factoren, zoals motivatie en persoonlijkheid. Voorts is het grote scala nieuwe opleidingen, in en buiten Nederland en België, vaak niet goed te classificeren binnen algemene opleidingsclusters. Tot slot, het voorspellingsdoel: wil men weten hoe iemand het doet in vergelijking met een oorspronkelijke groep mensen (retrospectief, opleiding) of hoe iemand presteert in vergelijking tot de groep mensen waartoe hij zal behoren (groep werknemers in doelpositie). De eerste vergelijking is 'statisch' en kan ook leiden tot unfair testgebruik.

Uitgebreide longitudinale studies tonen aan dat werkprestaties worden bepaald door 'conceptual ability' in plaats van opleiding. Het gaat om de wijze waarop mensen kunnen omgaan met werkzaamheden van een bepaald complexiteitsniveau⁷. De 'stratified systems theory' vat dit samen: werk is gestructureerd naar niveaus van toenemende complexiteit van competenties⁸. Bij toename van complexiteit van competenties wordt meer beroep gedaan op cognitieve vermogens. Dit principe wordt ook gevolgd bij o.a. de indeling naar beroepen en functies bij O*Net. Het onderzoek met O*Net job analysis database⁹ vormt het uitgangspunt voor de indeling van Verify referentiegroepen, naast empirisch onderzoek bij zeer grote groepen. Dit resulteerde in een indeling van testen naar moeilijkheidsniveau, opdat de mensen het juiste niveau van vragen krijgen voorgelegd, en naar een indeling in referentiegroepen om een vergelijking te maken met de beoogde functiegroep.

Verify testversies: het juiste moeilijkheidsniveau van vragen

Verify tests zijn ingedeeld naar vier niveaus van complexiteit. Deze indeling is gebaseerd (zie hierboven en schema) op theoretische uitgangspunten en empirisch onderzoek.

⁵ Anastasi, A. (1988). *Psychological Testing*. New York, New York: MacMillan Publishing Company.

⁶ Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey: Earlbaum.

⁷ Stamp, G. (1988). *Longitudinal research into methods of assessing managerial potential*. (Technical report 819) US Army, Research Institute for the Behavioural and Social Sciences.

⁸ Jaques, E. (1996) (Rev. 2nd ed.) *The requisite organisation: a total system for effective managerial organizational and managerial leadership for the 21st century*. Arlington, VA: Cason Hall.

⁹ Jeanneret, P. R., and Strong, M. H. (2003). *Linking O*NET job analysis information to job requirement predictors: An O*NET application*. *Personnel Psychology*, 56, 465-492.

Verify indeling (Global indeling)	Indicatie Nederlandstalige versies	Opleidingsniveau vergelijkbaar met opleiding op niveau	Functie complexiteit voorbeelden
Academische starter (graduate)	Academische start rollen met breed profiel	Hoog WO	Trainee/start rollen voor management of professionele functies (geen of minder dan 3 jaar werkervaring) – opleiding is bepalend voor complexiteit.
Manager/professional (Manager/professional)	Hoger opgeleide rollen met management en/of professioneel profiel	Hoog/midden WO, HBO	Management en professionele rollen, senior en midden niveau v.a. 3 jaar werkervaring – naast opleiding is ervaring en werkniveau bepalend
Midden management/non academische professional (Supervisor)	Middelbaar opgeleide rollen met coördinerend management en/of professioneel profiel a Midden coördinerend management b Midden professionals, evt teamleider	Midden HBO, MBO	Rollen waarbij sprake is van junior of midden management en/of professionele ervaring al dan niet met coördinerende verantwoording
Uitvoerend (Operator)	Lager opgeleide rollen met uitvoerend profiel a uitvoerend mensgericht b uitvoerend administratief	Laag MBO, VMBO of opleiding tot 16 jaar	Uitvoerende rollen

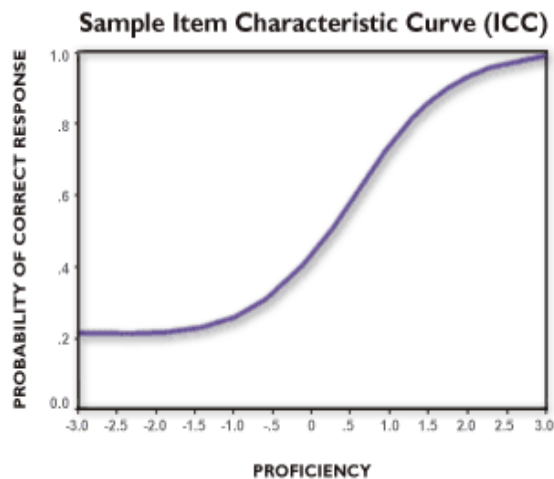
Indeling Verify testserie

Om te zorgen dat mensen niet een te makkelijke of te moeilijke testversie voorgelegd krijgen zijn Verify testen ingedeeld naar breed opleidingsniveau, naast werk(complexiteits)niveau. De laatste indeling correspondeert met de indeling van O*Net: lager, midden en hoger opleidingsniveau. Deze driedeling mag grof lijken. Deze volgt echter uit onderzoek op een bestand van tienduizenden items, afkomstig van testen voor lagere tot hogere opleidingsniveaus. Een verfijnde indeling leidt niet tot zinvol onderscheid, ook niet in Nederland of België. Hetzelfde geldt voor de indeling naar werkcomplexiteit. Dit impliceert overigens niet dat de gemiddelde niveaus op de gemeten capaciteiten niet verschillen voor groepen mensen. Studies geven aan dat indeling naar werksector dan het meest relevante onderscheid is. Samenvattend kennen de Verify tests verschillende versies – gebaseerd op werkcomplexiteit en de indeling naar opleidingsgraad; binnen deze versies worden referentie groepen onderscheiden naar werksector, naast een algemene referentiegroep. De indeling naar werksector is: algemeen, bank/verzekeringswezen en professionele dienstverlening, wetenschap en techniek, detailhandel/horeca en recreatie en publieke sector/overheid.

Berekening van scores Verify: Item Response Theorie als basis

Om de Verify testen samen te stellen is gebruik gemaakt van Item Response Theorie (IRT). Binnen deze theorie kan voor elk item de moeilijkheidsgraad en het onderscheidend vermogen (relatie met de te meten 'latente trek') worden geschat. Uitgangspunt van IRT is dat de items steekproef onafhankelijk zijn, en ongecorrleerd. Dit betekent dat de resultaten voor de calibratie ook gebruik worden voor het opstellen van het referentiemodel. Voorts draagt elk item in de test bij aan het eindresultaat op basis van de waarde die het item heeft in relatie tot de latente trek. Met andere woorden, elk item kan worden opgevat als een 'mini test', die bijdraagt aan het eindresultaat. Bij een test volgens de Klassieke Test Theorie (KTT) draagt een item wel of niet bij aan het totaal (men maakt het item goed of

fout). Bij IRT draagt elke item in een bepaalde mate (de moeilijkheidsgraad) bij aan het eindresultaat.



*Voorbeeld item informatie functie: Grafiek van latente trek in relatie tot kans op item antwoord.
Hoe hoger de positie op de trek, des te hoger de kans op 'correct antwoord'.
Intercept Y-as: moeilijkheidsgraad. Helling: discriminatie index.*

De steekproef onafhankelijkheid is essentieel. Bij KTT kan alleen een uitspraak worden gedaan over de desbetreffende verzameling items die onderzocht op de specifieke doelgroep. Verandert er iets aan de test (bijvoorbeeld deletie van een item) of aan de doelgroep dan zijn de uitspraken niet meer geldend/valide. Vergelijk een emmer water: bij bevriezing (andere omstandigheden) verandert de eigenschap van de inhoud wezenlijk. Bij IRT geldt dat voor elk item de validiteit bekend is, en dat deze geldt voor de grote groep van mensen waarvoor de test is samengesteld. Vergelijk een emmer zand: na bevriezing heeft het zand nog steeds dezelfde kenmerken als daarvoor. IRT geeft een meer 'vaste' basis voor het doen van voorspelling van kwaliteiten.

IRT toepassing

- De klassieke test theorie wordt toegepast op de test als geheel. De analyse en kenmerken hebben betrekking op de specifieke testversie en de betreffende groep mensen die deze heeft ingevuld.
- Item Response Theorie richt zich op de meting van het onderliggende kenmerk ('latente trek') in plaats van de prestatie op de test zelf. Het onderliggende kenmerk wordt in elk item gemeten en is steekproef onafhankelijk. Voor elk item zijn de meetkenmerken bekend, ongeacht de verzameling van items in een test en onafhankelijk van de groep mensen die het item heeft beantwoord.
- Bij IRT wordt voor elk item twee kenmerken bepaald. Moeilijkheidsgraad: hoeveel mensen 'bezitten' dit kenmerk. 'Discriminatie index: mate waarin het item onderscheid maakt tussen hoge en lage posities op de latente trek. Als iedereen een kenmerk wel/niet bezit heeft het weinig onderscheidend vermogen.
- Uitgangspunt voor de scoring van Verify is het patroon van alle antwoorden op de items. Zogenaamd goed of fout gemaakte items dragen elk bij tot het eindresultaat waarbij de relatie met de latente trek bepalend is (en dus niet alleen het model van goed of fout).

Verify referentiegroepen: zeer grote vergelijkings(norm)groepen

Voor de samenstelling van Verify is gebruik gemaakt van zeer grote referentiegroepen. Dit betekent dat de schatting van de parameters zeer accuraat is, en dat de testen een hoge betrouwbaarheid en validiteit kennen. Ook worden de resultaten van een deelnemer vergeleken met een referentiegroep, die vele malen groter is dan de doorsnee normgroep bij testen die het principe van de KTT volgen. De vergelijking met de referentiegroep is meer accuraat en betrouwbaar, en meer indicatief. Voor de Nederlandse en Belgische deelnemers in de referentiegroepen is er geen significante afwijking in antwoordpatronen met de deelnemers uit de 'global' referentiegroep (Benchmark factsheet Verify). Voor Nederlandstalige deelnemers kunnen de omvangrijke referentiegroepen gebruikt worden

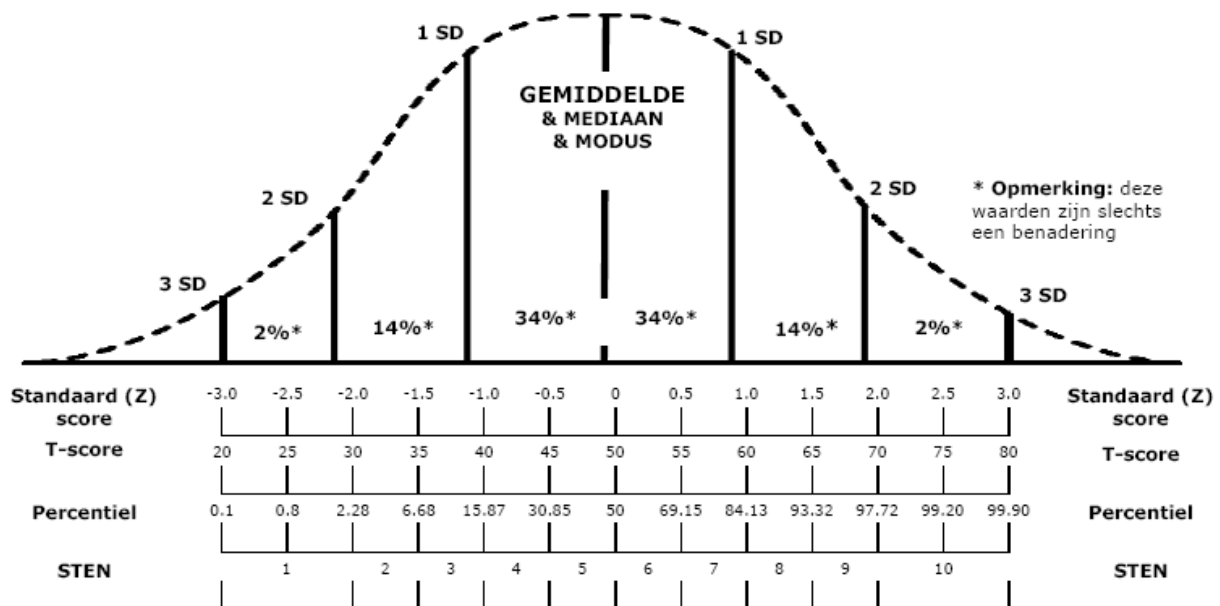
(steekproeven van duizenden mensen). Ook zijn geen relevante seksegerelateerde, leeftijdsgerelateerde of etnisch gerelateerde verschillen geconstateerd (d-maten tussen 0 en 0.2). Verify testen voldoen aan het principe van 'fairness'. Binnen de testversies zelf leiden verfijnde indeling naar opleidingsniveau ook niet tot verschillen. Zo volgt uit onderzoek met de Nederlandse Verify 'academische starters/graduate' testen dat het verschil tussen deelnemers met WO en HBO opleiding verwaarloosbaar is (d-maat 0.2 of kleiner).

Verify: Korte testen, ongesuperviseerde afname mogelijk

De items van Verify zijn onafhankelijk van elkaar, van elk item is bekend welke bijdrage deze heeft aan de meting van de latente trek, en hoe dit item geplaatst moet worden in relatie tot de referentiegroep. Zo kan worden volstaan met de kortste test die nodig is om de latente trek optimaal te meten. Verify testen zijn korte testen met een geringe afnameduur¹⁰. Het is mogelijk om uit een zeer grote itemdatabase steeds een testversie samen te stellen voor elke deelnemer, welke dezelfde kenmerken bezit in termen van betrouwbaarheid, validiteit in vergelijking met referentiegroep. Verify kan ongesuperviseerd afgenomen worden waarbij de kans op bekendwording van de test vrijwel nihil is. Deelnemers kunnen via een link de test thuis (of op een andere door hen gekozen locatie) maken; voor de testgebruiker is direct het resultaat beschikbaar. Om eventueel fraude bij invulling te controleren is verificatie van de test mogelijk. De ervaring leert dat fraudering in het Nederlandstalig gebied gering is. Zo men wil kan men volstaan met de mededeling dat de resultaten steekproefgewijs worden geverifieerd en kan men dit doen per bijvoorbeeld elke 10e deelnemer. De verificatietesten kennen een zeer korte duur (zie Verify user guide).

Praktisch gebruik: Verify profielen voor individuele deelnemers

Na afname komt voor elke deelnemer een rapport beschikbaar met toelichting op de gemaakte testversie en referentiegroep. De resultaten worden op verschillende indices gerapporteerd: percentielen, T-scores en Sten scores.



De Normaal Verdeling: basis voor Verify rapportage

¹⁰ Desgewenst is het mogelijk om de afnameduur bij te stellen, zie Verify Userguide. Dit kan vanwege de onafhankelijkheid van de items in de testversie.

Verify: Solutions aanpak

Zoals aangeduid, is Verify een krachtig instrument. Het komt met name tot zijn recht binnen een Solutions aanpak waarbij de diagnose/voorspellingstaak gestructureerd wordt opgepakt. Er wordt dan eerst bepaald welke methoden geschikt zijn voor de voorspelling van de relevante aspecten van de beoogde functie. Hiertoe kan de competentieprofiel (online beschikbaar) worden gebruikt: samen met opdrachtgever of relevante personen bepaalt men de competenties. Daar waar nodig en relevant kan werkervaring en specifieke opleiding (mits dit niet tot ´unfaire´ toepassingen leidt) worden gebruikt om taakprestaties te voorspellen. Verify wordt ingezet voor voorspelling van competenties die een beroep doen op cognitieve kwaliteiten. Daarnaast is het veelal ook zinvol om de Solutions aanpak aan te vullen met andere methodieken om tot een optimaal resultaat te komen. Bijvoorbeeld het gebruik van OPQ persoonlijkheidsmeting, MQ motivatiemeting en een gestructureerd competentiebased interview. SHL adviseert gebruikers graag bij het samenstellen van optimale Solutions aanpakken.

Voordelen van Verify

De voordelen van de Verify tests zijn evident. Door gebruik van ´de beste´ items zijn de testversies meer betrouwbaar en valide. De referentiegroepen zijn zeer groot en leiden tot een goede en geëigende vergelijking. In vergelijking met traditionele testen die normen baseren op – in het gunstige geval – enkele honderdtallen deelnemers, hanteert Verify een zeer krachtig model van grote steekproeven met duizendtallen deelnemers. Korte testen en de mogelijkheid tot ongesuperviseerde afname kennen veel praktisch gebruiksgemak, naast opties als preselectie e.d. Dit betekent voor de professionele gebruiker naast psychometrische accuratesse, een sterke reductie in kosten en tijd. Testversies en referentiegroepen zijn wereldwijd beschikbaar. Niet onbelangrijk is dat Verify geschikt is voor ´fair´ testgebruik, en door deelnemers als zeer prettig, acceptabel en modern/´bij de tijd´ wordt gevonden. Met name het laatste werkt ook imago verhogend voor de gebruiker. Dit alles maakt de Verify testen de oplossing voor het huidig en toekomst capaciteiten onderzoek.