Universele schalen voor testuitslagen: een pleidooi voor T-scores en percentielrangordescores

Beurs, E. de; Boehnke, J., Fried; E.I.

Boom

info@boomamsterdam.nl
www.boomuitgeversamsterdam.nl

Universele schalen voor testuitslagen

Een pleidooi voor T-scores en percentiel rangorde scores[1]

EDWIN DE BEURS, JAN BOEHNKE & EIKO FRIED

Samenvatting

De diversiteit aan meetinstrumenten die in gebruik zijn binnen de huidige ggz-praktijk bemoeilijkt de interpretatie van testuitslagen en de implementatie van Routine Outcome Monitoring (ROM). In dit artikel raden we aan bij het bespreken van testuitslagen met cliënten scores om te zetten naar universele schalen. De berekening van T-scores en percentiel rangorde scores wordt uitgelegd. We stellen voor om testresultaten uit te drukken als T-scores, met de algemene bevolking als referentiegroep. T-scores kunnen aangevuld worden met percentiel rangorde scores die gebaseerd zijn op een relevante klinische referentiegroep, om ze begrijpelijker te maken voor clienten. We maken de voordelen van deze benadering inzichtelijk met gegevens uit drie veelgebruikte depressievragenlijsten. We bieden een Excel-file met oversteekformules voor diverse vragenlijsten om eenvoudig ruwe scores naar T-scores en percentiel rangorde scores om te zetten. Recent is voorgesteld om bij gesubsidieerd onderzoek een beperkt aantal meetinstrumenten verplicht te stellen. Dit kan de ontwikkeling en toepassing van nieuwe instrumenten belemmeren. Als alternatief raden we aan om de voorgestelde universele schalen te gebruiken, teneinde testuitslagen te harmoniseren en zo ROM in de dagelijkse praktijk te vergemakkelijken en te bevorderen.
Trefwoorden: universele schaal, T-score, percentiel rangorde score, zelfrapportagevragenlijsten, testuitslag, depressie, ROM

1 Dit artikel is een bewerkte vertaling van de Beurs, Boehnke en Fried (2022).

Kernboodschappen voor de klinische praktijk

- Het gebruik van universele schalen helpt om de uitslag van tests over de ernst van klachten aan cliënten te verduidelijken en kan hun betrokkenheid bij ROM vergroten.
- Met behulp van universele schalen kan de vooruitgang die bereikt is met de behandeling (of het ontbreken daarvan) gemakkelijker worden vastgesteld, begrepen en besproken.
- Twee universele schalen worden uitgelegd: genormaliseerde T-scores en percentiel rangorde scores, en van beide worden de voor- en nadelen besproken.

INLEIDING

Meten is de basis van wetenschappelijk onderzoek. Sinds de tijd van Wundt en Thurstone heeft meten een prominente rol gespeeld in de psychologie. Er is in de psychologische wetenschap een aparte tak gewijd aan onderzoek naar meetmethoden en meetinstrumenten: de psychometrie. De gereedschapskist van de psychometrist is goed gevuld. De afgelopen decennia laten een wildgroei aan meetinstrumenten zien voor een breed scala aan psychologische concepten.

Enerzijds weerspiegelt het grote aantal beschikbare meetinstrumenten het belang van psychometrie in de klinische psychologie. Nieuwe meetinstrumenten komen vaak voort uit nieuwe theoretische perspectieven op de fenomenologie of behandeling van psychopathologie (bijvoorbeeld uit de verschuiving van aandacht voor gedragsmatige aspecten naar cognitieve). Of ze duiden op verandering in wat we als de belangrijkste kenmerken van een stoornis beschouwen (Böhnke & Croudace, 2015). Voorbeelden zijn het onderscheid tussen internaliserende en externaliserende symptomen bij kinderen en jeugdigen, de opkomst van het paniekconcept bij angststoornissen of het dimensionele model voor persoonlijkheidspathologie. Er zijn meetinstrumenten gevalideerd voor verschillende populaties (de algemene populatie versus klinische groepen) en voor verschillende doeleinden (diagnostiek versus het volgen van de voortgang van de behandeling; McPherson & Armstrong, 2021; Patalay & Fried, 2020). Ook weerspiegelen nieuwe meetinstrumenten innovaties in de meettechnologie zelf (bijvoorbeeld *computer adapted testing*; Williams et al., 2017).

Anderzijds vormt het grote aantal meetinstrumenten een uitdaging voor psychologisch wetenschappelijk onderzoek, want de diversiteit ervan bemoeilijkt het vergelijken van onderzoeksresultaten. Zo zijn er minstens 19 instrumenten om woede te meten (Weidman et al., 2017) en zijn er meer dan 280 instrumenten voor depressie ontwikkeld, waarvan er vele nog steeds in gebruik zijn (Fried, 2017; Santor et al., 2006). Er is weinig consistentie in de meetinstrumenten die worden gebruikt in therapie-effectonderzoek, zoals opgemerkt door Ogles (2013): in 163 onderzoeken rapporteerden on-

derzoekers resultaten op 435 unieke uitkomstmaten, waarvan er 371 slechts eenmalig in een onderzoek werden gebruikt. Deze overvloed aan meetinstrumenten – in combinatie met andere twijfelachtige meetpraktijken, zoals het wisselen van uitkomstmaten (Weston et al., 2016) en een gebrek aan transparantie over hoe scores worden afgeleid (Weidman et al., 2017) – helpt niet om een consistent wetenschappelijk fundament (*knowledge base*) onder de klinische psychologie tot stand te brengen. Het bemoeilijkt de vergelijking van resultaten van verschillende onderzoeken en vertraagt voortgang in de wetenschap (Flake & Fried, 2020).

In de klinische praktijk worden meetinstrumenten steeds vaker gebruikt om bij de start van de behandeling de ernst van de klachten, het functioneren en het welbevinden van cliënten te meten, en om vervolgens periodiek de voortgang van de behandeling te monitoren. In Nederland en Vlaanderen noemen we dit Routine Outcome Monitoring (ROM). (Een mooie introductie tot ROM is te vinden op https://kennisnet.vgct.nl/bericht/een-goede-introductie-van-rom; Schaffrat en collega's (2022) beschrijven een casus met een ROM-systeem dat men in Trier gebruikt.) Het meten van de resultaten van zorgverlening in de klinische ggz-praktijk wordt bepleit in tal van redactionele bijdragen en overzichtsartikelen (Boehnke & Rutherford, 2021; Fortney et al., 2017; Harding et al., 2011; Lambert, 2007; Lewis et al., 2019; Snyder et al., 2012). De Nederlandse ggz was rond 2000 een voorloper in deze beweging met ROM (de Beurs et al., 2011), dat inmiddels navolging heeft gevonden met PROMs en PREMs in de medische zorg in ziekenhuizen (http://iqprom.nl), gespecialiseerde klinieken (zie de Nederlandse Kanker Federatie: https://nfk.nl/promprem) en de eerstelijnszorg (www.nivel.nl/nl/dossiers/dossier-prems-en-proms-kwaliteit-meten-vanuit-patientenperspectief; zie ook de website van het Zorginstituut Nederland: www.zorginzicht.nl/ondersteuning/prom-wijzer/over-de-prom-wijzer).

Paradoxaal genoeg vormt de overvloed aan meetinstrumenten een belemmering voor het gebruik ervan in de klinische praktijk (Fried, 2017; Santor et al., 2006). Ten eerste bemoeilijkt de grote verscheidenheid aan instrumenten, elk met hun eigen schaal, de communicatie tussen professionals. Zo is de ernst van de symptomen van depressie uitgedrukt in een score op de Beck Depression Inventory (BDI-II; Beck et al., 1996), met een spreiding in scores van 0 tot 63, niet compatibel met de ernst uitgedrukt in een score op de Patient Health Questionnaire (PHQ-9; Kroenke & Spitzer, 2002) met een spreiding in scores van 0 tot 27. Dit compliceert de doorverwijzing van een cliënt door een behandelaar die de BDI-II gebruikt naar een collega die meer bekend is met de PHQ-9. Ten tweede wordt de communicatie over testresultaten tussen behandelaars en cliënten bemoeilijkt door het gebruik van verschillende schalen (Snyder et al., 2019). Wanneer alleen clinici testresultaten goed kunnen interpreteren, vergroot dat ongewild de kenniskloof tussen clinici en cliënten. Tegenwoordig pleit men juist voor meer betrokkenheid van de cliënt en voor gezamenlijke besluitvorming over de koers van de behandeling (Patel et al.,

2008), ook omdat dit leidt tot betere resultaten (Lambert & Harmon, 2018). Wanneer we cliënten een gelijkwaardiger rol willen geven in het therapeutische proces op hun weg naar hun herstel, werkt een kenniskloof over de betekenis van testuitslagen belemmerend. Duidelijke informatie over de betekenis van een testuitslag en hoeveel vooruitgang er is geboekt in de richting van behandeldoelen zal cliënten helpen om hen beter bij de behandeling te betrekken, zal hun een actievere rol geven en zal hun betrokkenheid bij de verstrekte informatie versterken (Goetz, 2010). Ten slotte wordt de implementatie van meten in de dagelijkse klinische praktijk gehinderd wanneer veel ervaring met meetinstrumenten en kennis over hun specifieke psychometrische eigenschappen nodig is voor een juiste interpretatie van testuitslagen. Dit kan ertoe leiden dat professionals in de ggz hun klinische oordeel blijven stellen boven de meetresultaten van gestandaardiseerde meetinstrumenten, ondanks het feit dat dit klinische oordeel vaak vertekend wordt door wensdenken en notoir onbetrouwbaar is (Kahneman et al., 2021; Meehl, 1954). Overigens zijn we niet van mening dat een testuitslag het klinische oordeel zou moeten *vervangen*, maar vinden we dat naast de testuitslag zowel het oordeel van de clinicus als dat van de cliënt zelf van belang blijft bij het nemen van beslissingen over het continueren, opschalen, afschalen of beëindigen van de behandeling.

Een strategie om de interpretatie van testresultaten in de klinische praktijk te vergemakkelijken bestaat al tientallen jaren: het gebruik van universele schalen. Met een schaal bedoelen we hier een systeem dat de eenheden beschrijft waarin meetresultaten worden uitgedrukt en geïnterpreteerd. Een voorbeeld is het metrische systeem dat rond 1790 is ingevoerd, met schalen om afstand, volume en gewicht te beschrijven. In de medische wereld zijn universele schalen vaak gebaseerd op fysieke entiteiten, zoals de meter, kilogram, mol en Kelvin. In de psychologie meten we echter meestal niet-fysieke entiteiten. We meten abstracte concepten of latente trekken, zoals depressie, angst en zelfvertrouwen – allemaal zaken die we niet direct kunnen observeren (Flake & Fried, 2020; Kellen et al., 2021). Bijgevolg kunnen we niet terugvallen op fysieke entiteiten, maar drukken we meetresultaten uit op een schaal die geen fysieke tegenhanger heeft, zoals wanneer we de ernst van depressie uitdrukken als de afstand tot de gemiddelde score van een referentiegroep in standaardeenheden (de standaardafwijking). Bij psychologische beoordelingen zetten we zo ruwe testuitslagen om naar een algemeen toepasbare universele schaal, zoals standaardscores of percentiel rangorde scores, die gebaseerd zijn op normatieve steekproeven (Kendall et al., 1999).

Tal van leerboeken over psychologisch testen beschrijven het omzetten van ruwe scores in universele schalen (Anastasi, 1968; Cronbach, 1984; Urbina, 2014), maar het wordt nog onvoldoende benut in klinisch onderzoek en bij psychologische behandelingen. Een voorbeeld van hoe het gebruik van een universele schaal goed heeft uitgepakt, kennen we uit een ander vakgebied, namelijk dat van de psychologische beoordeling of psychodiagnostiek: het meten van intellectuele vaardigheid. De IQ-score is meer dan een eeuw geleden

ontwikkeld door Binet en Simon (1907) en heeft universele toepassing gevonden; zowel professionele psychologen als leken begrijpen (in grote lijnen) de betekenis van IQ-scores, ongeacht het specifieke meetinstrument dat wordt gebruikt om intellectuele vaardigheid te meten. De interpretatie van testresultaten in de klinische praktijk zou veel gemakkelijker worden wanneer de betekenis van testscores net zo vanzelfsprekend zou zijn als die van IQ-scores.

De rest van dit artikel is als volgt opgebouwd. Eerst beschrijven we twee universele schalen: T-scores en percentiel rangorde scores (PR-scores). Beide schalen worden vaak gebruikt in neuropsychologisch en educatief onderzoek, en worden ook steeds vaker gebruikt voor klinische beoordeling. We beschrijven T-scores en PR-scores in detail en presenteren hun belangrijkste voor- en nadelen. In een technische bijlage bij dit artikel beschrijven we hoe deze schalen op diverse manieren kunnen worden afgeleid. We bieden ook een Excel-file met formules voor diverse meetinstrumenten om ruwe schaalscores om te zetten naar T-scores en PR-scores. We illustreren het gebruik ervan met meetinstrumenten voor depressie en bespreken de keuze van de juiste referentiegroep voor deze schalen. Ten slotte stellen we verdere stappen voor die nodig zijn voor een brede implementatie van deze schalen en doen we suggesties voor toekomstig onderzoek.

KANDIDAATSCHALEN

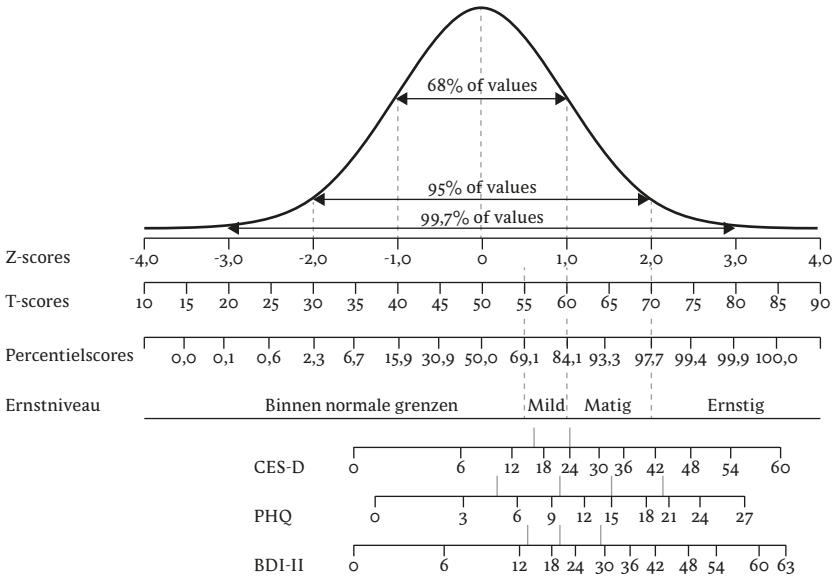
Als universele schalen voor psychologische concepten zijn twee opties voorgesteld: (1) standaardscores (Z-scores en alternatieven, zoals T-scores, stanines en stens) en (2) PR-scores (Crawford & Garthwaite, 2009; Ley, 1972). Tabel 1 geeft een overzicht van deze schalen, hoe ze berekend kunnen worden en enkele eigenschappen.

TABEL 1 *Universele schalen om testscores te standaardiseren en harmoniseren*

	Berekening	M	SD	Theoretisch bereik
Z-score	Z = (X-M)/SD	0	1	-INF tot INF
Stanine	S = Z * 2 + 5	5	2	1 tot 9
Sten	S = Z * 2 + 5,5	5,5	2	1 tot 10
T-score	T = Z * 10 + 50	50	10	-INF tot INF
PR-score	$PR = \frac{CF - (0{,}5 * F)}{N}$	50	n.v.t.	0 tot 100

Noot. M = gemiddelde; SD = standaarddeviatie; CF = cumulatieve frequentie; F = frequentie; N = totaal aantal observaties; INF = oneindig; n.v.t. = niet van toepassing

Figuur 1, geïnspireerd op Seashore (1955), toont de verschillende schalen onder de normaalverdeling en geeft labels voor de interpretatie van de scores bij het gebruik als ernstindicator of als screener. Om bijvoorbeeld betekenis te geven aan de ernst van depressie, heeft de PROMIS-groep de conventie gevolgd om de algemene bevolking in vier segmenten op te delen, van een lage naar een hoge score, respectievelijk 69,1%, 15,0%, 13,6% en 2,3%. Labels en bijbehorende T-scores zijn: 'binnen normale grenzen' (T < 55, 0), 'mild' (T = 55, 0-59,9), 'matig' (T = 60, 0-69,9) en 'ernstig' (T ≥ 70,0) (www.healthmeasures.net/score-and-interpret/interpret-scores/promis/promis-score-cut-points).



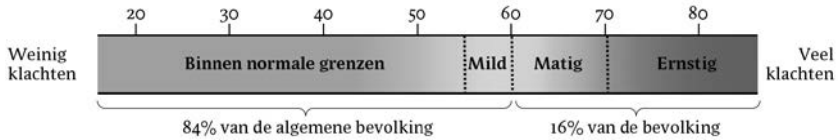
FIGUUR 1 *Normaalverdeling met Z-scores, T-scores en PR-scores, en overeenkomstige ruwe scores van de Amerikaanse algemene bevolking op de Center for Epidemiological Studies – Depressieschaal (CES-D), Patient Health Questionnaire (PHQ-9) en Beck Depression Inventory (BDI-II)*

Figuur 1 bevat ook ruwe scores op de PHQ-9 (Kroenke & Spitzer, 2002), de Center for Epidemiological Studies – Depressieschaal (CES-D; Radloff, 1977) en de BDI-II (Beck & Steer, 1987) van de Amerikaanse bevolking (Choi et al., 2014) om te laten zien hoe die zich verhouden tot deze universele schalen. Voor alle drie instrumenten geldt dat de intervallen tussen ruwe scores aan de lage kant van de schaal verder uit elkaar liggen dan in het midden of aan de hoge kant van de schaal. Dit illustreert de rechts scheve frequentieverdeling van ruwe scores in de algemene bevolking: lage scores komen vaker voor dan hoge scores.

T-scores

T-scores (zo genoemd door McCall (1922) om de psychometrische pioniers Thorndike, Terman en Thurstone te eren) zijn gebaseerd op gestandaardiseerde of Z-scores. Z-scores zijn ruwe scores omgezet naar een standaardschaal met een gemiddelde van 0 en een standaarddeviatie van 1. Ze worden berekend op basis van het gemiddelde en de standaarddeviatie van een referentiegroep (zie tabel 1). Standaardisatie naar Z-scores levert echter onhandige scores op met een bereik van -3 tot 3, hetgeen negatieve scores impliceert, en meerdere decimalen zijn vereist om voldoende precisie aan te geven. Er zijn alternatieven voorgesteld met een handiger formaat, zoals stans, stanines en T-scores. Stans en stanines geven een nogal grove indeling van scoreniveaus en we laten ze hier verder buiten beschouwing. T-scores zijn Z-scores vermenigvuldigd met 10 en daaraan 50 punten toegevoegd. Ze hebben zo een gemiddelde van 50 en een standaarddeviatie van 10, en hebben in de praktijk een bereik van 20 tot 80. Het theoretisch bereik van T-scores loopt van -oneindig tot +oneindig, maar T-scores lager dan 20 of hoger dan 80 zijn zeldzaam; Figuur 1 laat zien dat een T-score van 80 drie standaarddeviaties boven het gemiddelde is, een score die volgens de cumulatieve normale verdeling door slechts 0,13% van de respondenten wordt behaald; 99,7% scoort in het bereik van 20-80.

Al geruime tijd zijn T-scores de voorkeursschaal binnen psychodiagnostiek of een assessment. Om Cronbach (1984) te citeren: ‘Confusion results from the plethora of scales. In my opinion, test developers should use the system with mean 50 and s.d. 10 unless there are strong reasons for adopting a less familiar scale’ (p. 100). (Vertaling: ‘De overdaad aan schalen kan tot verwarring leiden. Naar mijn mening zouden testontwikkelaars de schaal moeten gebruiken met een gemiddelde van 50 en een SD van 10, tenzij er heel goede redenen zijn om een minder bekende alternatieve schaal te gebruiken.’) Voor de interpretatie van T-scores zijn praktische richtlijnen opgesteld. Bij het begin van de behandeling zullen de meeste cliënten een T-score hebben tussen 65 en 75; met de behandeling kan men streven naar een score onder de 55, wat als een redelijke grenswaarde geldt voor herstel bij veel meetinstrumenten die de ernst van psychopathologie meten (Aschenbrand et al., 2005; Cella et al., 2014; de Beurs, Carlier, & van Hemert, 2019; Recklitis & Rodriguez, 2007). Uit onderzoek blijkt dat veel cliënten de voorkeur geven aan kleurcodering van scoreniveaus volgens een *heat map* van groen via geel naar rood (Brundage et al., 2015; van Muilekom et al., 2021). Figuur 2 laat zien hoe de betekenis van T-scores kan worden overgebracht naar cliënten of collega’s.



FIGUUR 2 *De betekenis van T-scores voor het ernstniveau van klachten; bij weergave in kleur wordt links een groene en rechts een rode achtergrond kleur toegepast (bron: www.healthmeasures.net/score-and-interpret/interpret-scores/promis/promis-score-cut-points)*

T-scores zijn goed ingeburgerd in de klinische psychometrie. Ze werden gekozen als schaal door het PROMIS-initiatief, dat gericht is op het ontwikkelen van een nieuwe reeks meetinstrumenten in het gezondheidszorgonderzoek (Cella et al., 2010). Op dit gebied is al veel werk verzet, bijvoorbeeld in het Prosetta Stone-initiatief (Choi et al., 2014; Schalet et al., 2015). Hier zijn ruwe scores omgezet naar de PROMIS T-scoreschaal voor veelgebruikte meetinstrumenten van psychologische concepten, waaronder depressie (Choi et al., 2014), angst (Schalet et al., 2014) en pijn (Cook et al., 2015). Ook andere onderzoekers hebben T-scores gekozen als universele schaal, bijvoorbeeld voor depressie (Wahl et al., 2014), angst (Rose & Devine, 2014), fysiek functioneren (Oude Voshaar et al., 2019), vermoeidheid (Friedrich et al., 2019) en persoonlijkheidspathologie (Zimmermann et al., 2020).

PR-scores

De PR-score geeft voor elke ruwe score het percentage gevallen aan dat een lagere (of gelijke) score heeft dan die ruwe score. PR-scores geven zo met een waarde van 0 tot 100 de relatieve positie aan van de geteste persoon ten opzichte van zijn lotgenoten uit een referentiegroep (Kurtz & Mayo, 1979). In de klinische context laat zich dit vertalen in een boodschap aan de cliënt, zoals: 'Ten minste 75% van onze cliënten heeft een lagere score bij aanvang van de behandeling.' PR-scores geven zo de uitzonderlijkheid van de ruwe testuitslag aan met een percentage, wat helpt om de betekenis op een eenvoudige en directe manier aan collega's en cliënten over te brengen. Dit verklaart ook waarom PR-scores veel worden gebruikt in het onderwijs, bijvoorbeeld bij de Cito-toets. PR-scores worden weergegeven in figuur 1 onder de T-scoreschaal.

De literatuur over het gebruik van PR-scores in de klinische praktijk is beperkt. Crawford en Garthwaite (2009) hebben het gebruik ervan gepropageerd voor klinische (neuro)psychologische toepassingen. Crawford en collega's hebben PR-scores en betrouwbaarheidsintervallen gepubliceerd voor verschillende depressie- en angstmetingen op basis van Australische steekproeven (Crawford et al., 2011) en Britse steekproeven (Crawford et al., 2009) en een computerprogramma beschikbaar gesteld om deze scores te berekenen. Ze publiceerden ook oversteektabellen om ruwe scores om te

zetten in PR-scores (Crawford et al., 2011; Crawford & Henry, 2003). Recenter zijn ruwe scores en PR-scores gepubliceerd voor de Inventory of Depression- and Anxiety Symptoms (IDAS-II; Nelson et al., 2015).

OVERSTEEKTABELLEN

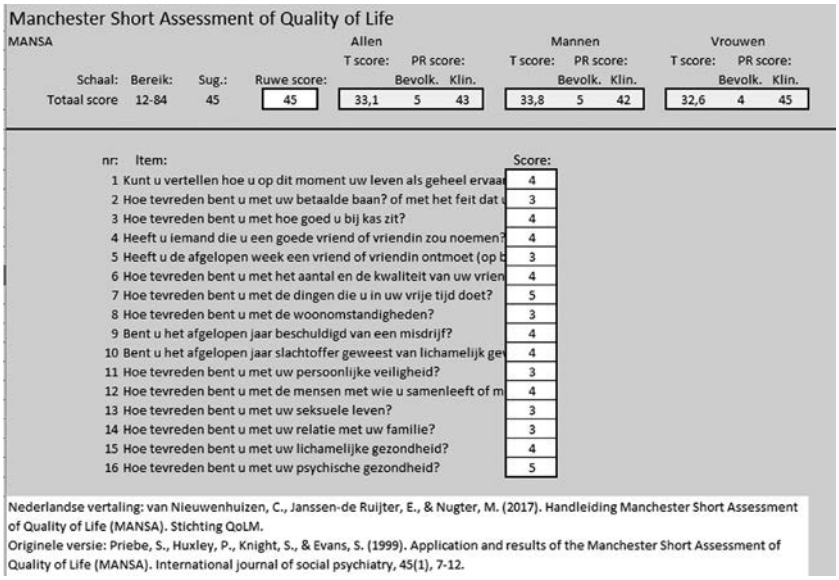
Om clinici te helpen de ruwe testscore van een individuele cliënt op een universele schaal uit te drukken, zijn voor veel meetinstrumenten oversteektabellen gepubliceerd (bijvoorbeeld: Batterham et al., 2018; Choi et al., 2014; Zimmermann et al., 2020). Ook handleidingen voor meetinstrumenten bieden soms oversteektabellen, waaronder de Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher et al., 1989), de Brief Symptom Inventory (BSI; Derogatis, 1975) en de Child Behaviour Checklist (CBCL; Achenbach, 1991). Het Prosetta Stone-initiatief (www.prosettastone.org) heeft oversteektabellen gepubliceerd voor veel meetinstrumenten naar de PROMIS T-score, die is gebaseerd op de Amerikaanse algemene bevolking. T-scores kunnen ook worden afgeleid met behulp van de webapp www.commonmetrics.org (Fischer & Rose, 2016).

T-scores of PR-scores opzoeken in tabellen is omslachtig en foutgevoelig. Daarom zijn er ook oversteek*formules* opgesteld voor de rekenkundige conversie van ruwe scores naar T-scores (of PR-scores). De relatie tussen de twee kolommen getallen in een oversteektabel (zoals tabel 2, van ruwe score naar T-score of van ruwe score naar PR-score) kan gemodelleerd worden met statistische software, zoals *curve-fitting* in de regressiemodule van SPSS, of de nls- en nlstools-pakketten voor *non-linear least squares* modellering in R (Baty et al., 2015). Uit een dergelijke analyse rolt een formule om de ruwe score om te zetten naar een T-score of PR-score. Een voorbeeld van conversie met een formule wordt gegeven door Roelofs en collega's (2013) en door de Beurs, Oudejans en Terluin (2022). In het laatste artikel wordt verwezen naar R software om formules te verwerven waarmee T- en PR-scores kunnen worden bepaald (www.psycharchives.org/en/item/86e598e9-4828-4127-86ae-5f0d18e9586a). Men dient dan wel te beschikken over normatieve gegevens van cliënten en de algemene Nederlandse bevolking. Recent hebben we in dit tijdschrift deze methode ook gedemonstreerd voor de Mental Health Continuum (MHC-SF) (de Beurs, Kosterman et al., 2022). Voor een aantal in Nederland gebruikte vragenlijsten hebben we een Excel-bestand samengesteld met oversteekformules (zie figuur 3 en https://github.com/EdwindeBeurs/ScoringQuestionnaires/blob/main/Scoring_Vragenlijsten.xlsx). Een gebruiker kan hier ruwe schaalscores (of itemscores) invoeren en het Excel-sheet berekent vervolgens T- en PR-scores voor cliënten, vaak apart voor mannen en vrouwen of leeftijdsgroepen. Het is de bedoeling dat dit

TABEL 2 *Ruwe schaalscores en PROMIS T-scores op de CESD, PHQ-9 en de BDI-II (Choi et al., 2014), en BDI-II PR-scores van de Nederlandse bevolking en van een klinische steekproef*

CES-D			PHQ-9			BDI-II				
RS		T	RS		T	RS		T	PR-pop	PR-clin
0	afwezig tot mild	34,5	0	minimaal	37,4	0	minimaal	34,9	10,1	0,5
2		41,1	1		42,7	2		42,3	26,7	1,7
4		44,7	2		45,9	4		46,2	38,5	3,4
6		47,5	3		48,3	6		48,9	48,6	5,5
8		49,8	4		50,5	8		51,0	57,3	7,9
10		51,7	5	mild	52,5	10		52,7	64,1	10,8
12		53,4	6		54,2	12		54,2	69,7	14,5
14		54,8	7		55,8	14	mild	55,6	74,4	18,4
16	matig	56,2	8		57,2	16		56,9	78,2	22,6
18		57,4	9		58,6	18		58,2	81,3	27,7
20		58,6	10	matig	59,9	20	matig	59,3	84,0	33,3
22		59,7	11		61,1	22		60,5	86,3	38,7
24	ernstig	60,8	12		62,3	24		61,6	88,5	44,8
26		61,8	13		63,5	26		62,7	90,3	51,2
28		62,9	14		64,7	28		63,8	91,6	57,5
30		63,9	15	matig ernstig	65,8	30	ernstig	64,8	92,7	63,6
32		64,9	16		66,9	32		65,8	94,0	69,2
34		66,0	17		68,0	34		65,8	95,2	74,0
36		67,0	18		69,2	36		67,6	96,0	78,3
38		68,1	19		70,3	38		68,9	96,8	82,5
40		69,2	20	ernstig	71,5	40		69,9	97,4	86,3
42		70,4	21		72,7	42		70,9	98,0	89,5
44		71,7	22		74,0	44		71,9	98,5	92,3
46		73,0	23		75,3	46		72,9	98,9	94,6
48		74,4	24		76,7	48		74,0	99,2	96,3
50		76,0	25		78,3	50		75,2	99,5	97,7
52		77,7	26		80,0	52		76,4	99,8	98,6
54		79,7	27		82,3	54		77,7	99,8	99,2
56		82,0				56		79,1	99,9	99,6
58		84,3				58		80,8	99,9	99,8
60		86,4				60		82,9	100,0	99,9
						62		85,1	100,0	100,0
						63		86,3	100,0	100,0

Noot. BDI-II = Beck Depression Inventory; CES-D = Center for Epidemiological Studies – Depressie-schaal; PHQ-9 = Patient Health Questionnaire; RS = ruwe score; T = PROMIS Depression T-score (gebaseerd op de algemene bevolking van de VS). T-scores kunnen berekend worden uit ruwe scores (RS) met een rationele functie: (T = 35,7 + (3,83*RS-0,0023*RS**2)/(1+0,13*RS-0,0012*RS**2); PR-scores kunnen berekend worden uit T-scores met: PR-pop = -2,9+103,7/(1+exp(-0,162(T-49,6))); PR-cl = 1,0+100,7/(1+exp(-0,232*(T-62,7))).



Manchester Short Assessment of Quality of Life

MANSA

Schaal:	Bereik:	Sug.:	Ruwe score:	Allen T score:	Allen PR score: Bevolk.	Allen PR score: Klin.	Mannen T score:	Mannen PR score: Bevolk.	Mannen PR score: Klin.	Vrouwen T score:	Vrouwen PR score: Bevolk.	Vrouwen PR score: Klin.
Totaal score	12-84	45	45	33,1	5	43	33,8	5	42	32,6	4	45

nr:	Item:	Score:
1	Kunt u vertellen hoe u op dit moment uw leven als geheel ervaar	4
2	Hoe tevreden bent u met uw betaalde baan? of met het feit dat u	3
3	Hoe tevreden bent u met hoe goed u bij kas zit?	4
4	Heeft u iemand die u een goede vriend of vriendin zou noemen?	4
5	Heeft u de afgelopen week een vriend of vriendin ontmoet (op b	3
6	Hoe tevreden bent u met het aantal en de kwaliteit van uw vrien	4
7	Hoe tevreden bent u met de dingen die u in uw vrije tijd doet?	5
8	Hoe tevreden bent u met de woonomstandigheden?	3
9	Bent u het afgelopen jaar beschuldigd van een misdrijf?	4
10	Bent u het afgelopen jaar slachtoffer geweest van lichamelijk ge	4
11	Hoe tevreden bent u met uw persoonlijke veiligheid?	3
12	Hoe tevreden bent u met de mensen met wie u samenleeft of m	4
13	Hoe tevreden bent u met uw seksuele leven?	3
14	Hoe tevreden bent u met uw relatie met uw familie?	3
15	Hoe tevreden bent u met uw lichamelijke gezondheid?	4
16	Hoe tevreden bent u met uw psychische gezondheid?	5

Nederlandse vertaling: van Nieuwenhuizen, C., Janssen-de Ruijter, E., & Nugter, M. (2017). Handleiding Manchester Short Assessment of Quality of Life (MANSA). Stichting QoLM.
Originele versie: Priebe, S., Huxley, P., Knight, S., & Evans, S. (1999). Application and results of the Manchester Short Assessment of Quality of Life (MANSA). International journal of social psychiatry, 45(1), 7-12.

FIGUUR 3 *Voorbeeld van scoring van de MANSA: ruwe score, T-score en twee PR-scores voor allen (ongeacht sekse), en voor mannen en vrouwen apart. In de kolom 'Ruwe score' kunnen schaalscores worden ingevuld. Als alternatief kunnen achter de items in de kolom 'Score' de itemscores worden ingevuld. Bij invullen van itemscores wordt de schaalscore berekend en in de kolom 'Sug.' weergegeven. Om de T- en PR-score te krijgen, moet die schaalscore vervolgens overgebracht worden naar de kolom voor de ruwe score (knippen en plakken als waarde of overtypen).*

een levend document wordt om in de toekomst uit te breiden met extra vragenlijsten zodra daarvan de benodigde normgegeven voorhanden zijn.

T-SCORES EN PR-SCORES VERGELEKEN

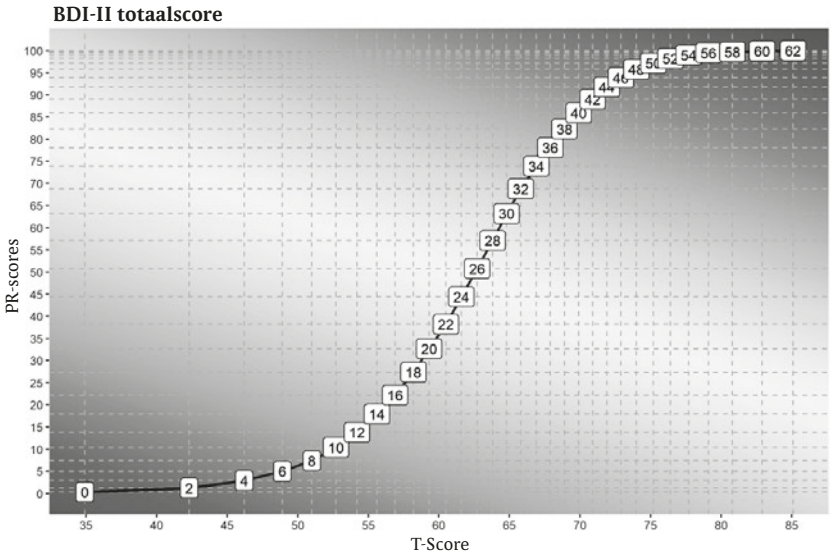
Een minimumeis voor een universele schaal is dat T-scores gemakkelijk te interpreteren zijn en op een eenvoudige manier uitdrukken hoe uitzonderlijk een testuitslag is (Snyder et al., 2012). In de COSMIN-checklist wordt interpreteerbaarheid gedefinieerd als 'the degree to which one can assign a qualitative meaning to a score' (Prinsen et al., 2018). Veel van de problemen met meetinstrumentgebonden scoringsmethoden die in de inleiding werden genoemd kunnen zo opgelost worden. De T-score geeft de universele betekenis van een testuitslag weer met de afstand tot het gemiddelde van een referentiegroep in standaardeenheden, waarbij 50 staat voor het gemiddelde en 10 punten een standaardafwijking betekenen. De interpretatie van de T-score veronderstelt bij de testgebruiker echter wel enige kennis van de normaalverdeling, zoals de '68-95-99,7'-regel, het ezelsbruggetje om te onthouden dat

68% van de scores binnen 1 SD van het gemiddelde valt, 95% binnen 2 SD's en 99,7% binnen 3 SD's (zie figuur 1). Dit betekent dat een T-score hoger dan 80 of lager dan 20 vrij uitzonderlijk is en alleen behaald wordt door de hoogste en laagste 0,13% van de referentiegroep. In de praktijk loopt de T-score dus van 20 tot 80. PR-scores verlenen op een meer intuïtieve manier een universele betekenis aan een testuitslag door de score uit te drukken als het percentage (0-100) respondenten met dezelfde of een lagere ruwe score.

T-scores worden gemeten op een intervalmeetschaal, ervan uitgaande dat dit een redelijke aanname is, gegeven de frequentieverdeling van de oorspronkelijke ruwe scores: de afstand van 40 naar 50 is even groot als de afstand van 50 naar 60. Dit geldt echter niet voor PR-scores. Die zitten niet op gelijke afstanden van elkaar, maar moeten als ordinale of volgordelijke scores beschouwd worden (een ordinale meetschaal). De relatie tussen T-scores en PR-scores is S-vormig. Een hoge PR-score (PR = 75) komt bijvoorbeeld overeen met een T-score van 57, wat op een relatief bescheiden afstand van het gemiddelde van 50 ligt. In het hoge bereik komen de T-scores 65, 70, 75 en 80 overeen met respectievelijk 93,3, 97,7, 99,4 en 99,9 (zie figuur 1). Tussen PR-scores zitten dus duidelijk ongelijke intervallen, vooral aan de extremen, wat bij gebruik van PR-scores leidt tot een overdrijving van verschillen tussen scores nabij het gemiddelde en onderschatting van verschillen aan de extremen.

Figuur 4 toont, voor een selectie van ruwe BDI-II-scores, hun relatie met T-scores en klinische PR-scores. (Voor de leesbaarheid van de figuur zijn alleen de even ruwe scores opgenomen.) Nederlandse gegevens zijn gebruikt om T-scores en PR-scores te verkrijgen uit een klinische steekproef (de Beurs et al., 2011). De figuur laat de S-vormige relatie tussen normaal verdeelde T-scores en PR-scores zien, wat de uitrekking van de PR-score weerspiegelt ten opzichte van de T-score aan de extremen. (Vanwege de klokvorm van de normaalverdeling zijn PR-score-intervallen smaller in het midden van de schaal en breder aan de uitersten; zie ook figuur 1.)

Als mens zijn we het meest gewend aan schalen met gelijke intervallen, zoals meters, kilogrammen, enzovoort. Als de PR-score echter ten onrechte wordt beschouwd als een schaal met gelijke intervallen, zijn conclusies over de testuitslag vaak onjuist. Deze vertekening in perceptie van PR-scores werd door Bowman (2002) aangetoond in een onderzoek met derdejaars psychologiestudenten, die ertoe neigden om PR-scores boven 80 of onder 20 als een vrij extreme testuitslag te beschouwen, terwijl onder de normaalverdeling PR = 80 overeenkomt met T = 58,4, dat wil zeggen: minder dan 1 SD verwijderd van het gemiddelde en nog niet een derde op weg naar de maximale score van T = 80 (zie figuur 1). In verband met de heersende opvatting over toelaatbare transformaties en rekenkundige en statistische bewerkingen, zijn met PR-scores eenvoudige bewerkingen niet mogelijk. Zo is een gemiddelde PR-score of een verschilscore tussen herhaalde PR-scores betekenisloos. Voor een verdere bespreking van dit onderwerp, zie: Meijer en Oos-



FIGUUR 4 *Relatie tussen geselecteerde ruwe scores op de Beck Depression Inventory (BDI-II), met T-scores (x-as) en PR-scores binnen de klinische populatie (y-as); horizontale en verticale rasterlijnen corresponderen met de ruwe scores.*

terloo (2008). Het grote voordeel van PR-scores, namelijk dat ze over het algemeen intuïtiever zijn, wordt door dit nadeel overschaduwd.

T-scores zijn voor de interpretatie van testresultaten en verdere verwerking van gegevens de beste keuze, maar voor een heldere communicatie met collega's en cliënten raden we aan om T-scores aan te vullen met PR-scores. T-scores zijn veelzijdiger, gezien hun psychometrische eigenschappen, maar PR-scores zijn intuïtiever te interpreteren. We wijzen er verder op dat met name extreme PR-scores vatbaar blijken voor incorrecte interpretatie. Voorzichtigheid is geboden en bij gebruik moet goede uitleg gegeven worden.

KEUZE VAN DE JUISTE REFERENTIEGROEP

De twee gesuggereerde universele schalen, T-scores en PR-scores, hebben als doel om aan te geven hoe gewoon of hoe uitzonderlijk een testuitslag is. Hierbij is een passende referentiegroep nodig. Wat passend is, hangt af van de vraag die voorligt: willen we de ernst van de depressie van een geteste persoon weten in vergelijking met de algemene bevolking of in vergelijking met cliënten die geestelijke gezondheidszorg ontvangen? We doen hierbij uiteenlopende aanbevelingen voor T-scores en PR-scores. Voor T-scores is de algemene bevolking de beste keuze. We streven bij voorkeur naar een breed toepasbare schaal, zowel voor bijvoorbeeld epidemiologische studies

als voor toepassing in de ggz-context. Gebruik van een referentiesteekproef uit de algemene populatie maakt het mogelijk om bij diverse subgroepen binnen de ggz de ernst van het gemeten concept aan te duiden en onderling te vergelijken, omdat we voor iedereen van dezelfde referentiestandaard gebruikmaken. Een vergelijking met IQ-scores maakt deze voorkeur evident: het zou vreemd zijn om IQ-scores te normeren op personen met een verminderd intellectueel functioneren of, aan het andere uiterste, voor hoogbegaafden.

Net als voor T-scores is voor PR-scores de algemene bevolking de meest geschikte referentiegroep voor experimenteel en epidemiologisch onderzoek. Echter, voor toepassing in de klinische praktijk raden we aan om een *klinische* referentiegroep te gebruiken. Dit advies hangt samen met de eerdergenoemde vervorming van de PR-scoreschaal aan de extremen in vergelijking met de gelijke intervallen van de T-scoreschaal. Cliënten zullen, zeker aan het begin van de behandeling, vaak tot de hoogste 5 tot 10% van de algemene populatie behoren (Löwe et al., 2008; Schulte-van Maaren et al., 2013). Bijgevolg zijn PR-scores op basis van de algemene populatie vrij beperkt in bereik wanneer ze worden toegepast op cliënten die zich aanmelden voor behandeling. Hierdoor wordt het moeilijk om met PR-scores onderscheid te maken tussen cliënten met milde, matige en ernstige klachten. De S-vormige relatie tussen T-scores en PR-scores heeft ook tot gevolg dat een substantiële verandering in T-score zich aan het uiteinde van de schaal vertaalt in een kleine verandering in PR-score. Verandering in score die verbetering tijdens de behandeling weerspiegelt, wordt dus onderschat met de PR-schaal. Het uitdrukken van de score als klinische PR-score (dat wil zeggen: de PR-waarde ten opzichte van een vergelijkbare klinische groep) zal bruikbaardere informatie opleveren.

Wanneer klinische PR-scores worden gebruikt, moet wel de klinische populatie worden gespecificeerd waarvan ze zijn afgeleid (bijvoorbeeld opgenomen cliënten, poliklinische cliënten, poliklinische cliënten gezien in een privépraktijk, enzovoort). Een illustratief voorbeeld zijn de normatieve gegevens voor de Outcome Questionnaire (OQ-45; Lambert et al., 2004) die zijn gepubliceerd door Timman en collega's (2017), die normatieve gegevens leverden voor verschillende cliëntengroepen. Ook Tingey en collega's (1996) bespreken de kwestie van klinische subgroepen die variëren naar ernst. Zij stellen een oplossing voor in het kader van de benadering van Jacobson en Truax (1991) van klinisch significante verandering.

Voor een meer gedetailleerde vergelijking met subklinische cliëntgroepen (bijvoorbeeld aan het einde van een succesvolle behandeling, waarbij beide referentiegroepen gerechtvaardigd zijn), zou men zowel een algemeen populatiepercentiel als een klinisch percentiel kunnen presenteren. Men kan bijvoorbeeld rapporteren dat een cliënt een T-score van 60 heeft, wat hoog is in vergelijking met de algemene bevolking (van wie 84% een lagere score heeft), maar wel onder het gemiddelde ligt van andere cliënten die

de kliniek bezoeken (van wie 67% een hogere score heeft). De aangeboden informatie is afhankelijk van de context. Wel willen we ervoor waarschuwen de cliënt niet te overladen met te veel informatie, omdat dit verwarrend kan zijn en indruist tegen het principe van eenvoud en minimalisme: 'less is more' (Peters et al., 2007).

ILLUSTRATIE

We illustreren het gebruik van T-scores en PR-scores met gegevens van drie verschillende meetinstrumenten die veel worden gebruikt om de ernst van depressie te meten. De oversteektabellen in tabel 2 dienen als voorbeeld om het nut van universele schalen te demonstreren. De gegevens voor tabel 2 zijn gebaseerd op de tabellen A1 tot A3 uit Choi en collega's (2014), die oversteektabellen bieden voor drie veelgebruikte depressievragenlijsten: (1) de CES-D, (2) de PHQ-9 en (3) de BDI-II. Deze oversteektabellen dienen om ruwe scores om te zetten naar de PROMIS-depressie T-scoreschaal. Tabel 2 bevat een selectie van ruwe scores, hun betekenis en T-scores, die zijn gebaseerd op de Amerikaanse algemene bevolkingssteekproef van PROMIS.

Volgens de richtlijnen voor interpretatie van ruwe schaalscores van de CES-D (Radloff, 1977), duidt een score van 0-15 op afwezige tot milde depressie, 16-23 op matige depressie, en 24 of meer op ernstige depressie (zie tabel 2). De grens voor *caseness* (een positieve testuitslag die suggereert dat de cliënt aan depressie lijdt) is ≥ 16, wat overeenkomt met een T-score van 56,2. Voor de PHQ-9 duiden scores van 0 tot 4 op minimale depressie, 5-9 op milde, 10-14 op matige, 15-19 op matig ernstige en 20-27 op ernstige depressie (Kroenke & Spitzer, 2002). De voorgestelde grens voor caseness is ≥ 10, wat overeenkomt met een T-score van 59,9. Interpretatierichtlijnen uit de handleiding van de BDI-II duiden een totaalscore van 0-13 aan als minimale depressie, 14-19 als mild, 20-28 als matig en 29-63 als ernstig. Een recente meta-analyse suggereerde als optimale afkapwaarde voor klinische depressie de waarde ≥ 18,2 (von Glischinski et al., 2019). Voor deze drie maten begint caseness dus bij T = 56 tot 60, en de afkapwaarden tussen milde en matige depressie vallen over het algemeen samen met T = 60, wat door de PROMIS-groep ook werd voorgesteld voor de T-score (www.promis.org). Bij het vertalen van meerdere depressiematen naar één gemeenschappelijk concept en T-scoreschaal, suggereerden Wahl en collega's (2014) voor de BDI-II T = 54,0, T = 66,2 en T = 72,9 als drempels voor milde, matige en ernstige depressie. Deze afkapwaarden voor de drie vragenlijsten komen onderling overeen (maar zijn niet exact gelijk), wat mogelijk te wijten is aan verschillen in operationalisering van het concept 'depressie' of aan verschillen tussen de normatieve steekproeven. Maar de afkapwaarden in tabel 2 laten zien dat de grenswaarden 55, 60 en 70 voor mild, matig en ernstig in grote lij-

nen overeenkomen met wat auteurs voor de verschillende instrumenten als grenswaarden voor ruwe scores hebben voorgesteld.

Tabel 2 toont voor de BDI-II ook PR-scores met als referentiegroep de algemene bevolking (PR-pop) en een klinische groep (PR-clin). De gegevens zijn afkomstig uit een steekproef (N = 7500) van de Nederlandse algemene bevolking (Roelofs et al., 2013) en een steekproef van cliënten (N = 9844) die zich hadden aangemeld voor psychiatrische zorg en die deelnamen aan Routine Outcome Monitoring bij GGZ Rivierduinen (de Beurs et al., 2011). Hier wordt het verschil in bruikbaarheid duidelijk van PR-scores op basis van de algemene populatie en PR-scores van een klinische steekproef. Een persoon met een ruwe score van 20 op de BDI-II heeft PR-pop = 84,0 (en behoort daarmee tot de hoogste 16% van de algemene bevolking), maar een PR-clin van 33,3 (de laagste 33% van de klinische populatie). Een andere persoon met een ruwe score van 26 heeft een PR-pop = 90,3 (de hoogste 10% van de populatie), maar de klinische PR-score is PR-clin = 51,2, wat aangeeft dat een BDI-II-score van 26 dicht ligt bij de klinische mediane score. Bij BDI-II-scores in het hogere segment, zoals we die vinden bij cliënten, differentiëren de PR-clin-scores veel beter in ernstniveaus dan de PR-pop-scores.

DISCUSSIE

De huidige praktijk van psychologisch onderzoek, met een grote diversiteit aan instrumenten voor vergelijkbare concepten, draagt bij aan zwak gedefinieerde metingen, vergeleken met de sterk gedefinieerde metingen zoals we die kennen uit de natuurwetenschappen (Finkelstein, 2003). Dit vormt een bedreiging voor de kwaliteit van onderzoek en belemmert het tot stand komen van een robuust kennisfundament onder ons klinisch handelen. Onder professionals kan het bijdragen aan verwarring en onduidelijkheid over de betekenis van klinische testresultaten. Ten slotte – en de focus van dit artikel – belemmert dit een heldere communicatie tussen clinici en hun cliënten over de betekenis van testuitslagen.

Universele schalen zijn beschikbaar voor implementatie, zoals blijkt uit de uitvoerige lijst van referenties over hun toepassing. Ze zijn al geïmplementeerd in de Nederlandse klinische praktijk (de Beurs, 2010), waar 'Delta T' een veelgebruikte term is geworden om de voortgang aan te geven van cliënten van pre- tot posttest. We raden T-scores aan als universele schaal om de ernst van een score op meetinstrumenten voor klinische fenomenen uit te drukken. We herhalen hierbij ook nog eens de eerdere aanbeveling om de algemene bevolking als referentiegroep voor de T-score te gebruiken (de Beurs, Flens, & Williams, 2019). Bij cliënten die zich aanmelden voor behandeling zal de T-score hoog zijn (T > 50) bij klachten en symptomen, of laag (T < 50) bij functioneren, welbevinden of kwaliteit van leven. De keuze voor de algemene bevolking als referentiegroep is logisch: dezelfde keuze is ge-

maakt bij bijvoorbeeld de IQ-score. Een PR-score gebaseerd op een relevante klinische referentiegroep kan aan cliënten duidelijk maken hoe hun ernstniveau zich verhoudt tot andere cliënten. Een PR-score spreekt intuïtief aan, maar heeft beperkingen. Het is bijvoorbeeld niet mogelijk om PR-scores van elkaar af te trekken als maat voor wat er tijdens de behandeling is bereikt.

Verder willen we twee belangrijke onderwerpen onderstrepen, elk met een aantal suggesties voor verdere ontwikkeling. Het eerste onderwerp is puur praktisch: het gebruik van universele schalen zou gemeengoed moeten zijn en de standaard voor rapportage van testuitslagen. Daar moeten verschillende stappen voor worden gezet. Ten eerste is een betere uitleg van universele schalen nodig en een demonstratie van hun nut. We hopen dat dit artikel hier een bijdrage aan levert. Een tweede belangrijke stap voor de praktische implementatie is dat voor zoveel mogelijk meetinstrumenten oversteektabellen, -formules en -figuren beschikbaar komen waarmee de testgebruiker ruwe schaalscores op een gemakkelijke manier kan omzetten naar T-scores en PR-scores, bijvoorbeeld als standaardonderdeel van de documentatie van een meetinstrument. Het Excel-bestand dat we voor dit artikel hebben samengesteld, is een eerste aanzet hiertoe. Verder moeten oversteekformules beschikbaar komen die kunnen worden ingebouwd in software die binnen de klinische praktijk wordt gebruikt om vragenlijsten af te nemen, zoals VIP-live, QuestManager, BergOp, enzovoort.

Een tweede onderwerp voor toekomstige doorontwikkeling is psychometrisch onderzoek. Een uitputtend overzicht van de resterende obstakels valt buiten het bestek van dit artikel, maar we noemen hier vier uitdagingen. Ten eerste vereist de conversie van ruwe scores naar T-scores op basis van de algemene populatie dat we kunnen beschikken over normatieve gegevens van bevolkingssteekproeven. Momenteel zijn dergelijke gegevens niet voor alle meetinstrumenten voorhanden. Een tweede uitdaging hangt samen met het feit dat beschikbare normatieve gegevens van meetinstrumenten afkomstig zijn uit specifieke landen. De PROMIS T-score is bijvoorbeeld gebaseerd op normatieve gegevens van de Amerikaanse bevolking. We zullen moeten onderzoeken of we deze normen ook internationaal kunnen toepassen en daartoe moeten gegevens in andere landen worden verzameld en scores worden vergeleken (Terwee et al., 2021). Voor Nederlandse vertalingen van de PROMIS-instrumenten is hiermee een begin gemaakt door Elsman en collega's (Elsman, Flens et al., 2022; Elsman, Roorda et al., 2022). Dit zal de internationale vergelijking van uitkomstonderzoek naar behandeling in de ggz vergemakkelijken en uitkomstgegevens ook relevanter maken voor de klinische praktijk. Een derde onderzoeksuitdaging is de mogelijke invloed van de vertaling van meetinstrumenten en aanpassing aan de specifieke cultuur (Sousa & Rojjanasrirat, 2011). We moeten tevens onderzoeken of meetinstrumenten gevoelig zijn voor andere factoren, zoals geslacht, leeftijd, opleidingsniveau en/of sociaaleconomische status. Invloed van deze factoren zou impliceren dat verschillende normen (en T-scores en PR-scores) gelden

voor subgroepen van respondenten (Teresi et al., 2009; Weinfurt, 2021). Dat impliceert aanvullende conversieformules, met coëfficiënten voor geslacht, leeftijd, enzovoort, bijvoorbeeld voortvloeiend uit regressiegebaseerde normering (Timmerman et al., 2021) of alternatieve benaderingen (Lenhard & Lenhard, 2021). Voorts moet mogelijk rekening worden gehouden met andere externe variabelen, zoals de culturele achtergrond van cliënten of aspecten van hun psychopathologie (Böhnke & Croudace, 2015). In het Excelbestand hebben we daar alvast een begin mee gemaakt door aparte formules te leveren voor mannen en vrouwen, en – waar relevant – voor leeftijdsgroepen. De vierde uitdaging ten slotte is dat conversie naar T-scores zou moeten corrigeren voor niet-normaliteit van onbewerkte testscores, om te eindigen met T-scores met echt gelijke intervallen. Om dit te bewerkstelligen, kunnen conversietabellen en -formules gebaseerd worden op de itemresponstheorie (IRT)-benadering, op de frequentieverdeling van ruwe scores (op PR-scores gebaseerde normalisatie) of op andere normalisatiebenaderingen. Welke aanpak de beste T-score oplevert, verdient nader onderzoek (Kolen & Brennan, 2014).

Het harmoniseren van het meten van psychopathologie en psychische gezondheid is een belangrijk methodologisch probleem. We hebben hier praktische oplossingen voorgesteld om meetresultaten in de klinische praktijk te rapporteren en te bespreken: twee universele schalen. Maar dat is maar een deel van de oplossing van het onderliggende probleem: de wildgroei aan meetinstrumenten in wetenschappelijk psychologisch onderzoek. Om harmonisatie te stimuleren is recent gesuggereerd een beperkte set aan meetinstrumenten verplicht te stellen om subsidie te kunnen verwerven (Wolpert, 2020). Vervolgens hebben het National Institute of Mental Health en de Wellcome Trust plannen gelanceerd (Farber et al., 2020) om een beperkte set meetinstrumenten voor onderzoek en uitkomstmeting voor te schrijven: de PHQ-9 voor depressie, de General Anxiety Disorder (GAD-7; Spitzer et al., 2006) voor angst, de Revised Child Anxiety and Depression Scale (RCADS-22; Chorpita et al., 2000) voor depressie en angst bij kinderen en adolescenten, en de World Health Organisation Disability Assessment Schedule (WHODAS; Üstün et al., 2010) voor de invloed van beperkingen op het functioneren bij volwassenen. Hoewel dit de vergelijkbaarheid en interpreteerbaarheid van onderzoeksresultaten zal vergroten, kan het ook onbedoelde negatieve gevolgen hebben (Patalay & Fried, 2020). De geselecteerde instrumenten zijn niet zonder gebreken of nadelen. De PHQ-9 en GAD-7 bijvoorbeeld zijn ontwikkeld om te screenen op depressie en angst bij de algemene bevolking, niet om de voortgang tijdens de behandeling te volgen. Beide vragenlijsten zijn vrij kort, wat hun bereik en meetprecisie beperkt, en hun betrouwbaarheid en nauwkeurigheid vermindert, waardoor ze minder geschikt zijn om (statistisch) betrouwbare verandering (Jacobson & Truax, 1991) in de klinische praktijk te volgen, wat op zijn beurt kan leiden tot de voorbarige conclusie dat er geen klinische verandering is opgetreden. Ook

kan het verplichte gebruik van slechts een paar meetinstrumenten de verdere ontwikkeling van meetinstrumenten belemmeren (zoals het PROMIS-initiatief met computer adapted testing; Cella et al., 2010) of de verbetering van bestaande meetinstrumenten of meetmethoden tegenwerken. Toch is enige harmonisatie van uitkomstmaten in de klinische praktijk op zijn plaats en heeft dit belangrijke initiatieven gestimuleerd, zoals het International Consortium for Health Outcome Measurement (ICHOM), dat standaard meetsets heeft voorgesteld voor allerlei aandoeningen, waaronder depressie en angst (Obbarius et al., 2017). Dit kan helpen om de wildgroei aan meetinstrumenten voor *measurement-based care* te beperken en bijdragen aan een verbeterslag in de kwaliteit van de geestelijke gezondheidszorg (Fortney et al., 2017; Kilbourne et al., 2018).

In een ander deel van de literatuur wordt getracht beter uit te leggen wat we nu feitelijk nastreven met universele schalen: er is een verschil tussen enerzijds het vergelijkbaar maken van testresultaten die met uiteenlopende instrumenten zijn gemeten, en anderzijds het louter uitdrukken van ruwe schaalscores op een universele schaal. Testscores uitdrukken op een universele schaal kan twee doelen dienen: (1) het vergelijkbaar maken van scores van twee tests die bedoeld zijn om hetzelfde te meten (*parallel tests*), en (2) het uitdrukken van ruwe scores van meetinstrumenten die verschillende dingen meten op dezelfde schaal. Deze twee doelstellingen moeten niet met elkaar worden verward. Bij het vergelijkbaar maken van testuitslagen van verschillende tests is een achterliggende aanname dat hetzelfde concept wordt gemeten (Kolen & Brennan, 2014). Bij het uitdrukken van scores op een universele schaal is deze aanname niet vereist en is de operatie alleen bedoeld om testuitslagen naar een gemeenschappelijke schaal te transponeren. In het laatste geval blijven we ervan uitgaan dat we concepten meten die intrinsiek verschillend zijn, hoewel ze op dezelfde schaal worden uitgedrukt. Een T-score van 60 voor depressie betekent bijvoorbeeld niet hetzelfde als een T-score van 60 voor angst, want de twee concepten verschillen van elkaar. Sterker nog, gezien de aanzienlijke verschillen in inhoud van diverse meetinstrumenten voor depressie (zoals de CES-D en de BDI), kan het zijn dat twee dezelfde T-scores op deze meetinstrumenten niet dezelfde ernst van depressie aangeven (Fried, 2017). Daarom moeten we terughoudend zijn met het onderling vergelijken van T-scores voor verschillende concepten: ze drukken uit hoe uitzonderlijk de score is in vergelijking met een referentiegroep, zoals de algemene bevolking, maar niet meer dan dat. Hoewel T-scores belangrijke informatie bevatten, waarschuwen we dus voor het vergelijken of gelijkstellen van T-scores (of PR-scores, waar feitelijk hetzelfde voor geldt) die niet dezelfde concepten representeren. Voor een persoon die T = 65 scoort op depressie en T = 55 op angst, zou de juiste interpretatie en boodschap zijn: uw depressiescore wijkt meer af van het gemiddelde van de algemene bevolking (namelijk 1,5 standaardafwijking) dan uw angstscore (0,5 standaardafwijking). Anders gezegd: uw depressiescore is uitzonderlij-

ker dan de angstscore. Dit wijkt subtiel maar wezenlijk af van de foutieve interpretatie dat de cliënt meer depressief dan angstig is.

Verwarring over universele schalen en universele concepten kan ook ontstaan uit de onderzoeksliteratuur over de PROMIS-meetinstrumenten en de PROMIS T-scoreschaal. Voor diverse bestaande meetinstrumenten zijn oversteektabellen van ruwe schaalscores naar de PROMIS T-scoreschaal opgesteld. Dit onderzoek is meestal uitgevoerd per concept, zoals depressie (Choi et al., 2014), angst (Schalet et al., 2014), pijn (Cook et al., 2015) en fysiek functioneren (Schalet et al., 2015), allemaal met in de titel de formulering 'Establishing a common metric for …' Doel van deze onderzoeken was om telkens *binnen een concept* (bijvoorbeeld depressie) testuitslagen van verschillende meetinstrumenten gelijk te stellen met de T-scoreschaal van de PROMIS-pendant. De PROMIS T-scores zijn allemaal gekalibreerd op de Amerikaanse algemene populatie. Er zijn meer studies die deze benadering per concept gebruiken – zie bijvoorbeeld voor depressie Wahl en collega's (2014), voor fysiek functioneren Oude Voshaar en collega's (2019), voor persoonlijkheidspathologie Zimmermann en collega's (2020), voor psychische problemen Batterham en collega's (2018) en voor vermoeidheid Lai en collega's (2014). Maar ook hier geldt dat het gebruik van dezelfde schaal niet betekent dat identieke concepten zijn gemeten. De T-score drukt voor elk concept op een handzame manier de uitzonderlijkheid van de score uit (voor bijvoorbeeld depressie), maar dit betekent niet dat deze scores tussen concepten vergelijkbaar zijn.

Een andere standaardisatie is dat de scores op PROMIS-schalen voor het gemak allemaal in dezelfde richting gaan, dat wil zeggen: een hogere score vertegenwoordigt meer van het gemeten concept, zoals depressie, pijn, mobiliteit, functioneren of kwaliteit van leven. In visuele presentaties zijn ze allemaal op dezelfde manier kleur gecodeerd: groen is normaal of gezond, rood is disfunctioneel, ziek of ernstig (zie figuur 2).

Ten slotte is het belangrijk om erop te wijzen dat bij het omzetten van ruwe schaalscores naar T-scores de oorspronkelijke ruwe scores niet verloren gaan. De somscore van een stel items kan nog steeds worden berekend en gerapporteerd in onderzoek, bijvoorbeeld om scores te vergelijken met eerder gepubliceerde onderzoeksresultaten.

CONCLUSIE

Concluderend stellen we: (1) dat het gebruik van universele schalen, met name op de algemene bevolking gebaseerde T-scores en klinische PR-scores, nuttig zijn om de meetpraktijk in klinisch psychologisch onderzoek te harmoniseren, (2) dat ze behulpzaam zijn voor een juiste interpretatie van testuitslagen, (3) dat ze de communicatie over testuitslagen tussen professionals onderling kunnen verbeteren, en (4) dat ze het makkelijker maken

om de betekenis van testuitslagen goed uit te leggen aan cliënten in de geestelijke gezondheidszorg.

Een eerste stap om dit te stimuleren is om in handleidingen voor meetinstrumenten oversteektabellen en -formules op te nemen voor de omzetting van ruwe scores naar T-scores. Tegenwoordig worden zelfrapportagevragenlijsten meestal via de computer of smartphone afgenomen en kunnen testuitslagen eenvoudig in T-scores worden uitgedrukt. Aanvullende stappen zijn professionals opleiden en ervaring laten opdoen met de T-scoreschaal, en cliënten voorlichten over de betekenis van T-scores. Gaandeweg zal er zo 'gevoel' ontstaan bij professionals en cliënten voor de betekenis van de T-score. Het algemeen gebruik van IQ-scores en begrip door het grote publiek over wat IQ-scores overbrengen, laat zien dat deze aanpak haalbaar is. We hopen dat dit artikel het gebruik van T-scores stimuleert als de standaard om betekenis te geven aan (neuro)psychologische testuitslagen.

De belangrijkste aanbevelingen voor de klinische praktijk

- Bespreek testuitslagen als T-scores aan de hand van figuur 2, met de algemene bevolking als normgroep ('Uw T-score is 71, wat aangeeft dat u ernstige klachten hebt').
- Illustreer de ernst van de aandoening volgens het testresultaat mede met een klinische PR-score ('Uw percentielscore is 90, wat betekent dat slechts 10% van onze cliënten een hogere score heeft.')
- Belangrijke afkapwaarden voor de T- scoreschaal zijn 70, 60 en 55 (voor de overgangen tussen ernstig, matig, mild en normaal). Zie ook figuur 2.
- Vijf punten verandering in T-score komt overeen met een verandering van een halve standaardafwijking, wat een verandering is die door de meeste cliënten ook als zodanig ervaren wordt. T < 55 kan duiden op herstel.
- T-scores en PR-scores zijn op te zoeken in de handleiding van meetinstrumenten of te benaderen met relatief eenvoudige formules. Voor een aantal meetinstrumenten leveren we ze bij dit artikel in de vorm van een Excel-bestand.

BIJLAGE: DIVERSE METHODEN OM T-SCORES EN PR-SCORES TE VERKRIJGEN

T-score via lineaire omzetting

De meest rechtstreekse benadering is de conversie van ruwe scores naar T-scores met de lineaire formule T = (X - M/SD) * 10 + 50, waarbij X de ruwe score is, en M en SD respectievelijk het gemiddelde en de standaarddeviatie van de referentiepopulatie. Dit is een goede aanpak als de ruwe scores een

normale verdeling hebben. Vaak wordt niet aan deze voorwaarde voldaan en is de verdeling rechts scheef vanwege een oververtegenwoordiging van lage scores. Bij een scheve frequentieverdeling van scores zijn het rekenkundig gemiddelde en de standaarddeviatie ongeschikte descriptoren en is de interpretatie van de testuitslag in standaardscores niet meer eenduidig. Eerst dienen de ruwe scores te worden genormaliseerd en dan omgezet naar T-scores.

Ruwe scores met een afwijkende frequentieverdeling kunnen getransformeerd worden naar een normaalverdeling of genormaliseerd worden (Box & Cox, 1964; Liu et al., 2009). Een alternatieve, grondiger benadering is om eerst genormaliseerde standaardscores vast te stellen met op regressie gebaseerde normering (Zachary & Gorsuch, 1985), waarbij ook rekening kan worden gehouden met achtergrondvariabelen, zoals sekse, leeftijd, opleidingsniveau, enzovoort (Lenhard & Lenhard, 2021). Statistische software is beschikbaar in R (GAMLSS; Stasinopoulos en collega's (2018) en Timmerman en collega's (2021) bieden een gedetailleerde handleiding). Genormaliseerde standaardscores (Z) worden vervolgens omgezet in T-scores met $T = 10 * Z + 50$ (de lineaire formule in tabel 1).

T-scores volgens de benadering van de itemresponstheorie

Recent zijn op de IRT gebaseerde alternatieve benaderingen voorgesteld en beproefd om ruwe scores op bestaande meetinstrumenten om te zetten in T-scores, bijvoorbeeld met behulp van EAP-factorscores (Fischer & Rose, 2019). De IRT-benadering werd toegepast door Choi en collega's (2014) om ruwe scores op de CES-D, PHQ-9 en BDI-II om te zetten naar de PROMIS-schaal (een T-score met de Amerikaanse bevolking als referentiegroep), zoals weergegeven in figuur 1. Deze benaderingen vereisen opnieuw de aanname dat de latente trek die wordt gemeten bij benadering normaal verdeeld is, wat vaak niet het geval is wanneer klinische kenmerken, zoals depressie of angst, worden gemeten in steekproeven van de algemene bevolking. Voor toepassing van IRT bij een niet-normale verdeling van ruwe scores zijn alternatieven ontwikkeld (Reise et al., 2018). Omdat IRT-factorscores vaak worden geschaald als een standaardscore ($M = 0$, $SD = 1$), kunnen ze eenvoudig worden omgezet in T-scores met de lineaire formule uit tabel 1 ($T = 10 * Z + 50$).

Op PR-score gebaseerde omzetting naar T-scores

Een andere methode om ruwe scores naar T-scores te converteren is op basis van PR-scores (*equipercentile linking*). Uitgangspunt is ook hier een ruwe score van een unidimensionele schaal en een frequentieverdeling die de normaalverdeling voldoende benadert (Lord & Wingersky, 1984). De conversie op basis van PR-scores en de conversie die op IRT is gebaseerd leiden tot vergelijkbare resultaten, zoals aangetoond door Choi en collega's (2014) en meer recentelijk door Schalet en collega's (2021) bij het meten van depres-

sie, en door Brady en collega's (2022) bij het meten van de ernst van burnoutklachten onder professionals in de zorg. T-scores zijn gevalideerd aan het oordeel van deskundigen voor de ernst van pijn, vermoeidheid, angst en depressie bij oncologische cliënten (Cella et al., 2014). Ernstniveaus kwamen bijna perfect overeen met de T-score: T = 60 bleek een goede grenswaarde om matige ernst van gemiddelde ernst te onderscheiden, en T = 70 voor de overgang van gemiddeld naar ernstig. Verder zijn T-scores op de BSI (Derogatis, 1975) vergeleken met de Clinical Global Impression verbeteringsscore (CGI; Guy, 1976): een verbetering van 5 T-scorepunten kwam goed overeen met verbetering volgens de CGI-verbeteringsscore (de Beurs, Carlier, & van Hemert, 2019).

Welke methode kiezen?

Bij de keuze uit de besproken methoden om T-scores te verkrijgen (lineaire transformatie, PR-scores, IRT-modellen en regressiegebaseerde normering voor achtergrondvariabelen als geslacht en leeftijd), raden we aan om zo eenvoudig mogelijk te werk te gaan, tenzij complexiteit noodzakelijk is. Bij normaal verdeelde ruwe scores en geen significante verschillen tussen mannen en vrouwen of leeftijdsgroepen, volstaat de lineaire transformatie van ruwe schaalscores (T = 10 * Z + 50). Wanneer de ruwe scores niet normaal verdeeld zijn (wat te zien is aan de frequentieverdeling van de scores, met een qq-plot is te visualiseren of met statistische software is te toetsen) moeten we eerst de scores normaliseren met een transformatie, met PR-scores of op basis van een IRT-model. Wanner er geslachts- of leeftijdsverschillen spelen, leveren aparte conversies voor subgroepen of de regressiebenadering nauwkeurigere T-scores op. De IRT-benadering heeft nog andere voordelen boven eenvoudigweg schaalscores berekenen door itemscores te sommeren (de klassieke testtheoriebenadering), maar die bespreken we hier niet verder.

Berekening van PR-scores

PR-scores worden berekend met de formule uit de laatste kolom van tabel 1. Wanneer normatieve gegevens beschikbaar zijn, wordt de frequentie van elke ruwe score gebruikt om de PR-score te bepalen met de formule PR = (CumF - (0,5 * F)) / N (Crocker & Algina, 1986), waarin de cumulatieve frequentie (CumF) het aantal scores kleiner of gelijk aan de ruwe score is, F de frequentie van de ruwe score zelf is en N het totaal aantal waarnemingen. De PR-score wordt dus berekend als de proportie respondenten met een score net onder de ruwe score plus de helft van de proportie respondenten die precies de ruwe score hebben behaald. Onder de aanname van een normale verdeling kunnen PR-scores ook worden afgeleid uit Z-scores (en T-scores) volgens de formule voor de cumulatieve normaalverdeling, beschikbaar als

statistische functie in veel softwarepakketten (R, SPSS, STATA, MS Excel). De formule kan ook worden benaderd door een exponentiële functie, die een S-vormige curve beschrijft: $PR = 100 / 1 + e^{(-1{,}75*Z)}$, waarbij Z de standaardscore is; of met $PR = 100 / 1 + e^{(-0{,}175*T+8{,}75)}$, waarbij T de T-score is.

Voor dit manuscript zijn voornamelijk eerder gepubliceerde gegevens gebruikt. Voormetingsgegevens van Routine Outcome Monitoring van een klinische steekproef van ggz-instelling Rivierduinen in Leiden werden gebruikt om een kolom toe te voegen aan tabel 2. De medisch-ethische toetsingscommissie van het Leids Universitair Medisch Centrum stemde in met het gebruik voor wetenschappelijk onderzoek van deze geanonimiseerde gegevens, die routinematig werden verzameld in de klinische praktijk. Cliënten gaven toestemming voor geanonimiseerd gebruik van hun gegevens.

Edwin de Beurs is als hoogleraar verbonden aan de Sectie Klinische Psychologie, Universiteit Leiden, en aan Arkin GGZ, Amsterdam.

Jan R. Boehnke is als universitair hoofddocent (UHD) werkzaam bij de School of Health Sciences, University of Dundee, UK.

Eiko I. Fried is als universitair hoofddocent (UHD) verbonden aan de Sectie Klinische Psychologie, Universiteit Leiden.

Correspondentieadres: Edwin de Beurs, Universiteit Leiden, Sectie Klinische Psychologie, Wassenaarseweg 52, 2333 AK Leiden. E-mail: edwin.de.beurs@arkin.nl

Summary *Universal metrics for test results: A plea to use T-scores and percentile rank scores*

The diversity of measurement instruments in use in current mental health care practice makes it difficult to interpret test results and hampers implementation of ROM. To discuss test results with clients, we recommend converting them to common metrics. We explain the calculation of T-scores and percentile rank scores and discuss their pros and cons. We propose to express test results as T-scores with the general population as the reference group. T-scores can be supplemented with percentile scores based on a relevant clinical reference group, to make them more understandable for clients. We demonstrate the benefits of this approach with data from three commonly used self-report questionnaires for depression. In addition, we provide an Excel file with crosswalk formulas for several self-report questionnaires to easily convert raw scores to T-scores and percentile scores. It has recently been proposed to require a limited number of measurement instruments in subsidized research. This could hinder the development and application of new instruments. As an alternative, we recommend using the proposed common metrics to harmonize test results and thus facilitate and promote ROM in daily practice.

Keywords *common metric, T-score, percentile rank order score, self-report questionnaires, test result, depression, ROM*

Referenties

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist 4-18 and 1991 profiles*. Department of Psychiatry, University of Vermont.

Anastasi, A. (1968). *Psychological testing*. Macmillan.

Aschenbrand, S. G., Angelosante, A. G., & Kendall, P. C. (2005). Discriminant validity and clinical utility of the CBCL with anxiety-disordered youth. *Journal of Clinical Child & Adolescent Psychology, 34*, 735-746. https://doi.org/10.1207/s15374424jccp3404_15

Batterham, P. J., Sunderland, M., Slade, T., Calear, A. L., & Carragher, N. (2018). Assessing distress in the community: Psychometric properties and crosswalk comparison of eight measures of psychological distress. *Psychological Medicine, 48*, 1316-1324. https://doi.org/10.1017/S0033291717002835

Baty, F., Ritz, C., & Baty, M. F. (2015). *Package 'nlstools'. Tools for nonlinear regression analysis*. https://doi.org/10.18637/jss.v066.i05

Beck, A. T., & Steer, R. A. (1987). *Manual for the revised Beck Depression Inventory*. Psychological Corporation.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II Beck Depression Inventory Manual* (Vol. 2). The Psychological Corporation.

Binet, A., & Simon, T. (1907). Le développement de l'intelligence chez les enfants. *L'Annëe Psychologique, 14*, 1-94. https://doi.org/https://www.persee.fr/doc/psy_0003-5033_1907_num_14_1_3737

Boehnke, J. R., & Rutherford, C. (2021). Using feedback tools to enhance the quality and experience of care. *Quality of Life Research, 30*, 3007-3013. https://doi.org/10.1007/s11136-021-03008-8

Böhnke, J., & Croudace, T. J. (2015). Factors of psychological distress: Clinical value, measurement substance, and methodological artefacts. *Social Psychiatry and Psychiatric Epidemiology, 50*, 515-524. https://doi.org/10.1007/s00127-015-1022-5

Bowman, M. L. (2002). The perfidy of percentiles. *Archives of Clinical Neuropsychology, 17*, 295-303. https://doi.org/10.1016/S0887-6177(01)00116-0

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological), 26*, 211-243. https://doi.org/https://www.jstor.org/stable/2984418

Brady, K. J., Ni, P., Carlasare, L., Shanafelt, T. D., Sinsky, C. A., Linzer, M., Stillman, M., & Trockel, M. T. (2022). Establishing crosswalks between common measures of burnout in US physicians. *Journal of General Internal Medicine, 37*, 777-784. https://doi.org/10.1007/s11606-021-06661-4

Brundage, M. D., Smith, K. C., Little, E. A., Bantug, E. T., Snyder, C. F., & PROD-PSAB. (2015). Communicating patient-reported outcome scores using graphic formats: Results from a mixed-methods evaluation. *Quality of Life Research, 24*, 2457-2472. https://doi.org/10.1007/s11136-015-0974-y

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI–2 (Minnesota Multiphasic Personality Inventory–2): Manual for administration and scoring*. University of Minnesota Press.

Cella, D., Choi, S., Garcia, S., Cook, K. F., Rosenbloom, S., Lai, J.-S., Tatum, D. S., & Gershon, R. (2014). Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Quality of Life Research, 23*, 2651-2661. https://doi.org/10.1007/s11136-014-0732-6

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F.,

Gershon, R., Hahn, E. A., Lai, J. S., Pilkonis, P., Revicki, D., ... PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology, 63*, 1179-1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. *Psychological Assessment, 26*, 513-527. https://doi.org/10.1037/a0035768

Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: A revised child anxiety and depression scale. *Behaviour Research and Therapy, 38*, 835-855. https://doi.org/10.1016/S0005-7967(99)00130-8

Cook, K. F., Schalet, B. D., Kallen, M. A., Rutsohn, J. P., & Cella, D. (2015). Establishing a common metric for self-reported pain: Linking BPI Pain Interference and SF-36 Bodily Pain Subscale scores to the PROMIS Pain Interference metric. *Quality of Life Research, 24*, 2305-2318. https://doi.org/10.1007/s11136-014-0790-9

Crawford, J. R., Cayley, C., Lovibond, P. F., Wilson, P. H., & Hartley, C. (2011). Percentile norms and accompanying interval estimates from an Australian general adult population sample for self-report mood scales (BAI, BDI, CRSD, CES-D, DASS, DASS-21, STAI-X, STAI-Y, SRDS, and SRAS). *Australian Psychologist, 46*, 3-14. https://doi.org/10.1111/j.1742-9544.2010.00003.x

Crawford, J. R., & Garthwaite, P. H. (2009). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *The Clinical Neuropsychologist, 23*, 193-204. https://doi.org/10.1080/13854040801968450

Crawford, J. R., Garthwaite, P. H., Lawrie, C. J., Henry, J. D., MacDonald, M. A., Sutherland, J., & Sinha, P. (2009). A convenient method of obtaining percentile norms and accompanying interval estimates for self-report mood scales (DASS, DASS-21, HADS, PANAS, and sAD). *British Journal of Clinical Psychology, 48*, 163-180. https://doi.org/10.1348/014466508X377757

Crawford, J. R., & Henry, J. D. (2003). The Depression Anxiety Stress Scales (DASS): Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology, 42*, 111-131. https://doi.org/10.1348/014466503321903544

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston. https://books.google.nl/books?id=tfgkAQAAMAAJ

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). Harper & Row.

de Beurs, E. (2010). De genormaliseerde T-score, een 'euro' voor testuitslagen. *Maandblad Geestelijke Volksgezondheid*, *65*, 684-695. www.sbggz.nl

de Beurs, E., Boehnke, J., & Fried, E. I. (2022). Common measures or common metrics? A plea to harmonize measurement results. *Clinical Psychology & Psychotherapy, 29*(5), 1755-1767. https://doi.org/10.1002/cpp.2742

de Beurs, E., Carlier, I. V., & van Hemert, A. M. (2019). Approaches to denote treatment outcome: Clinical significance and clinical global impression compared. *International Journal of Methods in Psychiatric Research, 28*. https://doi.org/10.1002/mpr.1797

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., van der Lem, R., van Fenema, E., & Zitman, F. G. (2011). Routine outcome monitoring in The Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy, 18*, 1-12. https://doi.org/10.1002/cpp.696

de Beurs, E., Flens, G., & Williams, G. (2019). Meetresultaten interpreteren in de klinische psychologie: Een aantal voorstellen. *De Psycholoog, 54*, 10-23.

de Beurs, E., Kosterman, S., Anten, S., Bohlmeijer, E., & Westerhof, G. J. (2022). Psychometrische evaluatie van de Mental Health Continuum – Short Form (MHC-SF): Construct-validiteit, responsiviteit voor verandering, normen en T-scores. *Gedragstherapie, 55*, 131-155.

de Beurs, E., Oudejans, S., & Terluin, B. (2022). A common measurement scale for scores from self-report instruments in mental health care: T-scores with a normal distribution. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000740

Derogatis, L. R. (1975). *The Brief Symptom Inventory*. Clinical Psychometric Research.

Elsman, E. B. M., Flens, G., de Beurs, E., Roorda, L. D., & Terwee, C. B. (2022). Towards standardization of measuring anxiety and depression: Differential item functioning for language and Dutch reference values of the PROMIS item banks v1.0 – Anxiety and Depression. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0273287

Elsman, E. B. M., Roorda, L. D., Smidt, N., de Vet, H. C., & Terwee, C. B. (2022). Measurement properties of the Dutch PROMIS-29 v2. 1 profile in people with and without chronic conditions. *Quality of Life Research, 31*, 3447-3458. https://doi.org/10.1007/s11136-022-03171-6

Farber, G., Wolpert, M., & Kemmer, D. (2020). *Common measures for mental health science: Laying the foundations*. Wellcome Trust. https://wellcome.ac.uk/sites/default/files/CMB-and-CMA-July-2020-pdf.pdf

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement*, *34*, 39-48. https://doi.org/10.1016/S0263-2241(03)00018-6

Fischer, H. F., & Rose, M. (2016). Www.common-metrics. org: A web application to estimate scores from different patient-reported outcome measures on a common scale. *BMC Medical Research Methodology, 16*, 142. https://doi.org/10.1186/s12874-016-0241-0

Fischer, H. F., & Rose, M. (2019). Scoring depression on a common metric: A comparison of EAP estimation, plausible value imputation, and full Bayesian IRT modeling. *Multivariate Behavioral Research, 54*, 85-99. https://doi.org/10.1080/00273171.2018.1491381

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*, 456-465. https://doi.org/10.1177/2515245920952393

Fortney, J. C., Unützer, J., Wrenn, G., Pyne, J. M., Smith, G. R., Schoenbaum, M., & Harbin, H. T. (2017). A tipping point for measurement-based care. *Psychiatric Services, 68*, 179-188. https://doi.org/10.1176/appi.ps.201500439

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders, 208*, 191-197. https://doi.org/10.1016/j.jad.2016.10.019

Friedrich, M., Hinz, A., Kuhnt, S., Schulte, T., Rose, M., & Fischer, F. (2019). Measuring fatigue in cancer patients: A common metric for six fatigue instruments. *Quality of Life Research, 28*, 1615-1626. https://doi.org/10.1007/s11136-019-02147-3

Goetz, T. (2010). *It's time to redesign medical data* [Presentation]. TEDMED.

Guy, W. (1976). *Clinical Global Impressions ECDEU assessment manual for psychopharmacology. Revised* (DHEW publication ADM 76-338). Government Printing Office.

Harding, K. J., Rush, A. J., Arbuckle, M., Trivedi, M. H., & Pincus, H. A. (2011). Measurement-based care in psychiatric practice: A policy framework for implementation. *Journal of Clinical Psychiatry, 72*, 1136-1143. https://doi.org/10.4088/JCP.10r06282whi

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19. https://doi.org/10.1037//0022-006x.59.1.12

Kahneman, D., Sibony, O., & Sunstein, C. (2021). *Ruis: Waarom we zo vaak verkeerde beslissingen nemen en hoe we dat kunnen voorkomen.* Nieuw Amsterdam.

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science, 16*, 767-778. https://doi.org/10.1177/1745691620974771

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 285-299. https://doi.org/10.1037/0022-006X.67.3.285

Kilbourne, A. M., Beck, K., Spaeth-Rublee, B., Ramanuj, P., O'Brien, R. W., Tomoyasu, N., & Pincus, H. A. (2018). Measuring and improving the quality of mental health care: A global perspective. *World Psychiatry, 17*, 30-38. https://doi.org/10.1002/wps.20482

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer Science & Business Media.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*, 509-515. https://doi.org/10.3928/0048-5713-20020901-06

Kurtz, A. K., & Mayo, S. T. (1979). Percentiles and percentile ranks. In *Statistical methods in education and psychology* (pp. 145-163). Springer. https://doi.org/10.1007/978-1-4612-6129-2_6

Lai, J.-S., Cella, D., Yanez, B., & Stone, A. (2014). Linking fatigue measures on a common reporting metric. *Journal of Pain and Symptom Management, 48*, 639-648. https://doi.org/10.1016/j.jpainsymman.2013.12.236

Lambert, M. J. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research, 17*, 1-14. https://doi.org/10.1080/10503300601032506

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire-45. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Volume 3: Instruments for adults* (3rd ed., pp. 191-234). Lawrence Erlbaum Associates Publishers. http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2004-14941-006&site=ehost-live

Lambert, M. J., & Harmon, K. L. (2018). The merits of implementing routine outcome monitoring in clinical practice. *Clinical Psychology: Science and Practice, 25*, e12268. https://doi.org/10.1111/cpsp.12268

Lenhard, W., & Lenhard, A. (2021). Improvement of norm score quality via regression-based continuous norming. *Educational and Psychological Measurement, 81*, 229-261. https://doi.org/10.1177/0013164420928457

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H., Hoffman, M., Scott, K., Lyon, A., & Douglas, S. (2019). Implementing measurement-based care in behavioral health: A review. *JAMA Psychiatry, 76*, 324-335. https://doi.org/10.1001/jamapsychiatry.2018.3329

Ley, P. (1972). *Quantitative aspects of psychological assessment* (Vol. 1). Duckworth.

Liu, H., Lafferty, J., & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research, 10*. https://doi.org/10.1145/1577069.1755863

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement, 8*, 453-461. https://doi.org/10.1177/014662168400800409

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care*, *45*(3), 266-274. https://doi.org/10.1097/MLR.0b013e318160d093

McCall, W. A. (1922). *How to measure in education*. MacMillan.

McPherson, S., & Armstrong, D. (2021). Psychometric origins of depression. *History of the Human Sciences*, 09526951211009085. https://doi.org/10.1177/09526951211009085

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

Meijer, R. R., & Oosterloo, S. J. (2008). A note on measurement scales and statistical testing. *Measurement: Interdisciplinary Research and Perspectives, 6*, 198-204. https://doi.org/10.1080/15366360802324446

Nelson, E. C., Eftimovska, E., Lind, C., Hager, A., Wasson, J. H., & Lindblad, S. (2015). Patient reported outcome measures in practice. *BMJ Case Reports, 350*, g7818. https://doi.org/10.1136/bmj.g7818

Obbarius, A., van Maasakkers, L., Bear, L., Clark, D. M., Crocker, A. G., de Beurs, E., Emmelkamp, P. M. G., Furukawa, T. A., Hedman-Lagerlöf, E., Kangas, M., Langford, L., Legsage, A., Mwesigire, D. M., Nolte, S., Patel, V., Pilkonis, P. A., Pincus, H. A., Reis, R. A., Rojas, G., … Rose, M. (2017). Standardization of health outcomes assessment for depression and anxiety: Recommendations from the ICHOM Depression and Anxiety Working Group. *Quality of Life Research, 26*, 3211-3225. https://doi.org/10.1007/s11136-017-1659-5.

Ogles, B. M. (2013). Measuring change in psychotherapy research. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (pp. 134-166). Wiley.

Oude Voshaar, M., Vonkeman, H., Courvoisier, D., Finckh, A., Gossec, L., Leung, Y., Michaud, K., Pinheiro, G., Soriano, E., & Wulfraat, N. (2019). Towards standardized patient reported physical function outcome reporting: Linking ten commonly used questionnaires to a common metric. *Quality of Life Research, 28*, 187-197. https://doi.org/10.1007/s11136-018-2007-0

Patalay, P., & Fried, E. I. (2020). Editorial perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement. *Journal of Child*

Psychology and Psychiatry, 62, 1032-1036. https://doi.org/10.1111/jcpp.13333

Patel, S. R., Bakken, S., & Ruland, C. (2008). Recent advances in shared decision making for mental health. *Current Opinion in Psychiatry, 21*, 606-6012. https://doi.org/10.1097/YCO.0b013e32830eb6b4

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., & Mertz, C. K. (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review, 64*, 169-190. https://doi.org/10.1177/1077558707060020301

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*, 1147-1157. https://doi.org/10.1007/s11136-018-1798-3

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401. https://doi.org/10.1177/014662167700100306

Recklitis, C. J., & Rodriguez, P. (2007). Screening childhood cancer survivors with the Brief Symptom Inventory-18: Classification agreement with the Symptom Checklist-90-Revised. *Psycho-Oncology, 16*, 429-436. https://doi.org/10.1002/pon.1069

Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative approaches to addressing non-normal distributions in the application of IRT models to personality measures. *Journal of Personality Assessment, 100*, 363-374. https://doi.org/10.1080/00223891.2017.1381969

Roelofs, J., van Breukelen, G., de Graaf, L. E., Beck, A. T., Arntz, A., & Huibers, M. J. H. (2013). Norms for the Beck Depression Inventory (BDI-II) in a large Dutch community sample. *Journal of Psychopathology and Behavioral Assessment, 35*, 93-98. https://doi.org/10.1007/s10862-012-9309-2

Rose, M., & Devine, J. (2014). Assessment of patient-reported symptoms of anxiety. *Dialogues in Clinical Neuroscience, 16*, 197-211.

Santor, D. A., Gregus, M., & Welch, A. (2006). Focus article: Eight decades of measurement in depression. *Measurement: Interdisciplinary Research and Perspectives, 4*, 135-155. https://doi.org/10.1207/s15366359mea0403_1

Schaffrath, J., Weinmann-Lutz, B., & Lutz, W. (2022). The Trier Treatment Navigator (TTN) in action: Clinical case study on data-informed psychological therapy. *Journal of Clinical Psychology*. https://doi.org/10.1002/jclp.23362

Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of Anxiety Disorders, 28*, 88-96. https://doi.org/10.1016/j.janxdis.2013.11.006

Schalet, B. D., Lim, S., Cella, D., & Choi, S. W. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika, 86*, 717-746. https://doi.org/10.1007/s11336-021-09776-z

Schalet, B. D., Revicki, D. A., Cook, K. F., Krishnan, E., Fries, J. F., & Cella, D. (2015). Establishing a common metric for physical function: Linking the HAQ-DI and SF-36 PF subscale to PROMIS® Physical Function. *Journal of General Internal Medicine, 30*, 1517-1523. https://doi.org/10.1007/s11606-015-3360-0)

Schulte-van Maaren, Y. W., Carlier, I. V., Zitman, F. G., van Hemert, A. M., de Waal, M. W., van der Does, A. W., van Noorden, M. S., & Giltay, E. J. (2013). Reference values for major depression questionnaires: the Leiden Routine Outcome Monitoring Study. *Journal of Affective Disorders*, *149*,

342-349. https://doi.org/10.1016/j.jad.2013.02.009

Seashore, H. G. (1955). Methods of expressing Test scores. *Test Service Bulletin, 48*, 7-10. https://eric.ed.gov/?id=ED079347

Snyder, C. F., Aaronson, N. K., Choucair, A. K., Elliott, T. E., Greenhalgh, J., Halyard, M. Y., Hess, R., Miller, D. M., Reeve, B. B., & Santana, M. (2012). Implementing patient-reported outcomes assessment in clinical practice: A review of the options and considerations. *Quality of Life Research, 21*, 1305-1314. https://doi.org/10.1007/s11136-011-0054-x

Snyder, C., Smith, K., Holzner, B., Rivera, Y. M., Bantug, E., Brundage, M., & PRO Data Presentation Delphi Panel. (2019). Making a picture worth a thousand numbers: Recommendations for graphically displaying patient-reported outcomes data. *Quality of Life Research, 28*, 345-356. https://doi.org/10.1007/s11136-018-2020-3

Sousa, V. D., & Rojjanasrirat, W. (2011). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice*, *17*, 268-274. https://doi.org/10.1111/j.1365-2753.2010.01434.x

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine, 166*, 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Stasinopoulos, M. D., Rigby, R. A., & Bastiani, F. D. (2018). GAMLSS: A distributional regression approach. *Statistical Modelling, 18*, 248-273. https://doi.org/10.1177/1471082X18759144

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., Lai, J.-S., Choi, S. W., Hays, R. D., Reeve, B. B., Reise, S. P., Pilkonis, P. A., & Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science, 51*, 148-180.

Terwee, C. B., Crins, M. H. P., Roorda, L. D., Cook, K. F., Cella, D., Smits, N., & Schalet, B. D. (2021). International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores. *Journal of Clinical Epidemiology*, *134*, 1-13. https://doi.org/10.1016/j.jclinepi.2021.01.011

Timman, R., de Jong, K., & de Neve-Enthoven, N. (2017). Cut-off scores and clinical change indices for the Dutch Outcome Questionnaire (OQ-45) in a large sample of normal and several psychotherapeutic populations. *Clinical Psychology & Psychotherapy, 24*, 72-81. https://doi.org/10.1002/cpp.1979

Timmerman, M. E., Voncken, L., & Albers, C. J. (2021). A tutorial on regression-based norming of psychological tests with GAMLSS. *Psychological Methods, 26*(3), 357-373. https://doi.org/10.1037/met0000348

Tingey, R., Lambert, M., Burlingame, G., & Hansen, N. (1996). Assessing clinical significance: Proposed extensions to method. *Psychotherapy Research, 6*, 109-123. https://doi.org/10.1080/10503309612331331638

Urbina, S. (2014). *Essentials of psychological testing*. John Wiley & Sons.

Üstün, T. B., Kosstanjsek, N., Chatterji, S., & Rehm, J. (2010). *Measuring health and disability: Manual for WHO Disability Assessment Schedule WHODAS 2.0.*

van Muilekom, M. M., Luijten, M. A., van Oers, H. A., Terwee, C. B., van Litsenburg, R. R., Roorda, L. D., Grootenhuis, M. A., & Haverman, L. (2021). From statistics to clinics: The visual feedback of PROMIS® CATs. *Journal of Patient-Reported Outcomes, 5*, 1-14. https://doi.org/10.1186/s41687-021-00324-y

von Glischinski, M., von Brachel, R., & Hirschfeld, G. (2019). How depressed is 'depressed'? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). *Quality of Life Research*, *28*, 1111-1118. https://doi.org/10.1007/s11136-018-2050-x

Wahl, I., Löwe, B., Bjorner, J. B., Fischer, F., Langs, G., Voderholzer, U., Aita, S. A., Bergemann, N., Brähler, E., & Rose, M. (2014). Standardization of depression measurement: A common metric was developed for 11 self-report depression measures. *Journal of Clinical Epidemiology, 67*, 73-86. https://doi.org/10.1016/j.jclinepi.2013.04.019

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion, 17*, 267. https://doi.org/10.1037/emo0000226

Weinfurt, K. P. (2021). Constructing arguments for the interpretation and use of patient-reported outcome measures in research: An application of modern validity theory. *Quality of Life Research, 30*, 1715-1722. https://doi.org/10.1007/s11136-021-02776-7

Weston, J., Dwan, K., Altman, D., Clarke, M., Gamble, C., Schroter, S., Williamson, P., & Kirkham, J. (2016). Feasibility study to examine discrepancy rates in prespecified and reported outcomes in articles submitted to The BMJ. *BMJ Open*, *6*, e010075. https://doi.org/10.1136/bmjopen-2015-010075

Williams, G., Flens, G., & de Beurs, E. (2017). Computer adaptief testen: Moderne meettechniek voor uitkomstmaten. *PsyXpert, 3*, 14-21. https://www.psyxpert.nl/tijdschrift/editie/artikel/t/computergestuurd-adaptief-testen-cat

Wolpert, M. (2020). *Funders agree first common metrics for mental health science*. www.linkedin.com/pulse/funders-agree-first-common-metrics-mentalhealth-science-wolpert

Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology, 41*, 86-94. https://doi.org/10.1002/1097-4679(198501)41:1<86::AID-JCLP2270410115>3.0.CO;2-W

Zimmermann, J., Müller, S., Bach, B., Hutsebaut, J., Hummelen, B., & Fischer, F. (2020). A common metric for self-reported severity of personality disorder. *Psychopathology, 53*, 168-178. https://doi.org/10.1159/000507377