A common measurement scale for scores from self-report instruments in mental health care: T scores with a normal distribution

Beurs, E. de; Oudejans, S.; Terluin B.

European Journal of Psychological Assessment

A Common Measurement Scale for Self-Report Instruments in Mental Health Care: T Scores With a Normal Distribution

Edwin de Beurs, Suzan Oudejans, and Berend Terluin

Online First Publication, December 16, 2022. https://dx.doi.org/10.1027/1015-5759/a000740

Original Article

A Common Measurement Scale for Self-Report Instruments in Mental Health Care

T Scores With a Normal Distribution

Edwin de Beurs[1,2], Suzan Oudejans[3], and Berend Terluin[4]

[1]Department of Clinical Psychology, Faculty of Social Sciences, Leiden University, The Netherlands
[2]Arkin Mental Health Institute, Amsterdam, The Netherlands
[3]Mark Bench, Amsterdam, The Netherlands
[4]EMGO Institute, VU Medical Center, Amsterdam, The Netherlands

Abstract: The diversity of measures in clinical psychology hampers a straightforward interpretation of test results, complicates communication with the patient, and constitutes a challenge to the implementation of measurement-based care. In educational research and assessment, it is common practice to convert test scores to a common metric, such as *T* scores. We recommend applying this also in clinical psychology and propose and test a procedure to arrive at *T* scores approximating a normal distribution that can be applied to individual test scores. We established formulas to estimate normalized *T* scores from raw scale scores by regressing IRT-based θ scores on raw scores. With data from a large population and clinical samples, we established crosswalk formulas. Their validity was investigated by comparing calculated *T* scores with IRT-based *T* scores. IRT and formulas yielded very similar *T* scores, supporting the validity of the latter approach. Theoretical and practical advantages and disadvantages of both approaches to convert scores to common metrics and alternative approaches are discussed. Provided that scale characteristics allow for their computation, *T* scores will help to better understand measurement results, which makes it easier for patients and practitioners to use test results in joint decision-making about the course of treatment.

Keywords: IRT, *T* scores, percentile ranks, common metric, norms, clinical

The importance of measurement for clinical management and quality improvement of Mental Health Care (MHC) is widely acknowledged (Kilbourne et al., 2018; Lambert & Harmon, 2018). Inspired by the recovery movement (Slade et al., 2008) and developments such as shared decision-making (Patel et al., 2008), feedback informed treatment (Miller et al., 2015), and measurement-based care (Harding et al., 2011; Kilbourne et al., 2018), patients' needs and preferences are granted a more prominent role in MHC. Increased patient involvement requires being well-informed about the severity of one's condition at the onset of treatment and about the progress made in the journey toward recovery. Therefore, a good understanding of measurement results is needed when findings of Routine Outcome Monitoring (ROM; de Beurs et al., 2011) are shared. However, the huge diversity of measurement instruments in clinical practice, each with its measurement scale, complicates straightforward communication among professionals and between professionals and patients about their measurement results. This might hamper further implementation of measurement-based care. Using common, measure-independent metrics for the results of clinical tests may solve this problem (de Beurs, Fried, et al., 2022).

Two metrics have been proposed and researched in recent years: *T* scores and percentile rank. *T* scores, first proposed by McCall (1922), are standardized scores (*Z*-scores) multiplied by 10 and 50 added, resulting in a metric with $M = 50$ ($SD = 10$). The *T* score denotes the commonness of a test result by its distance from the mean of a reference group in standard units. *T* scores require the assumption that test results are measured on an interval scale. *Z* scores (and *T* scores) based on data from a reference population may have a highly skewed frequency distribution. It is advisable first to transform the raw scores to have a normal frequency distribution before converting them to *T* scores, which results in *normalized T* scores. When the normalized *T* score metric is calibrated on the

general population, most psychiatric patients will score before treatment around 65–70 on measures of psychopathology, and when treated successfully, their score will decrease to 55–60 over time. Percentile rank scores have been proposed as an alternative way to express how common or exceptional a test result is (Crawford & Garthwaite, 2009; Ley, 1972). They quantify the rarity of a tested person's score in a percentage. However, percentile scores are not on an interval scale but ordinal and are, especially at the extremes, easily misunderstood (Bowman, 2002).

There is quite some literature on converting measurement instruments to a common metric (Dorans et al., 2007; Holland et al., 2006) and the subject is closely linked to norming instruments (Mellenbergh, 2011). Most frequently used are distribution-based approaches, such as percentile conversion (Kolen & Brennan, 2014), and methods based on Item Response Theory (IRT; Embretson & Reise, 2013). Following the latter approach, several researchers have published reports on linking measures for the same constructs and have proposed population-based *T* scores as a common metric for the severity of depression (Choi et al., 2014; Fischer et al., 2011; Schalet et al., 2015; Wahl et al., 2014), anxiety (Schalet et al., 2014), pain (Cook et al., 2015), physical functioning (Schalet et al., 2015), fatigue (Friedrich et al., 2019), psychological distress (Batterham et al., 2018), and quality of life measures (ten Klooster et al., 2013). Usually, general population samples are used to easy interpretability of the resulting metric (Wahl et al., 2014). If appropriately sampled, such samples reflect the general population. In contrast, clinical samples vary in composition and severity or complexity of the disorder, and as such, they are less useful as a reference group for general psychopathology measures. Various clinical measurement instruments are administered to the same group of respondents, and an IRT estimate, usually the expected a posteriori (EAP) method (Lord & Wingersky, 1984) is used to estimate θ in order to express scores in the common metric. This endeavor is also known as the PROsetta Stone project (Schalet et al., 2015).

A potential drawback of the IRT approach is that it requires a comprehensive dataset and dedicated software to calculate θ-scores. In clinical practice, this is not always feasible: a clinician may only have a raw scale score from an individual patient, and he/she does not have access to the algorithm to obtain an IRT-based θ-score. To help out, several authors provide crosswalk tables to translate the raw test score into a *T* score (Batterham et al., 2018; Choi et al., 2014; Schalet et al., 2014). However, reading from crosswalk tables is cumbersome and prone to error. The relation between raw scores and *T* scores can also be modeled and expressed in a function. Such a function to calculate *T* scores can easily be used or implemented in ROM software and provides an alternative method to obtain scores on the common metric for individual patients. In line with international developments, *T* scores may be based on IRT models and stem from data from general population samples (de Beurs, Fried, et al., 2022). However, for everyday clinical practice, we propose an approach in which *T* scores are calculated with a conversion function. This will be feasible, even if only a test score from a single individual is available.

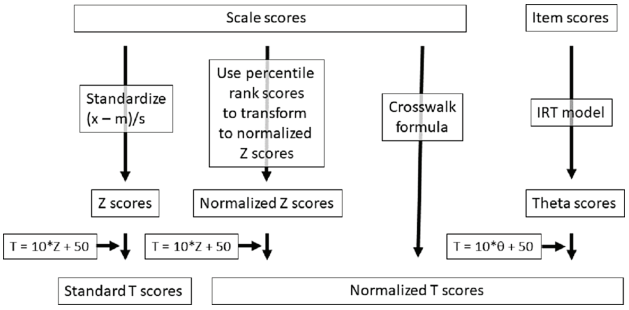

Figure 1. Overview of approaches to establish *T*-scores.

Figure 1 presents an overview of various approaches obtaining *T* scores. The first approach (on the left) entails standardizing the sum score of items of a scale into *Z* scores and converting these to *T* scores with $T = 10 \times Z + 50$ (standard *T* scores). The second approach is based on percentile rank scores, which are converted to normalized *Z* scores and subsequently to normalized *T* scores and take the frequency distribution of scores into account (percentile-based *T*-scores). The third approach is advocated in the present paper, with crosswalk formulas derived from the regression of *T* scores based on the factor scores resulting from an IRT model on sum scores (calculated *T* scores). The fourth approach is the IRT-based approach itself, practiced by – among others – the PROMIS group (θ-based *T* scores).

In this article we present crosswalk formulas for three frequently used clinical measurement instruments: the Brief Symptom Inventory (BSI; Derogatis, 1975), the Four-Dimensional Symptom Questionnaire (4DSQ; Terluin et al., 2006), and the Outcome Questionnaire (OQ-45; Lambert et al., 2004). A necessary step is to investigate the validity of calculated *T* scores by comparing them with θ-based *T* scores. If there is a high agreement, the transformation of raw scale scores with the conversion function is a valid approach for obtaining a proxy for θ-based *T* scores. For each measure, data were used from two samples: the general population and patients. IRT-based *T* scores were founded on both samples, with the general population as the reference population. The patient samples allowed us to investigate whether the *T* scores were normally distributed in clinical samples. Measures for ROM, where patients are repeatedly assessed to determine change over time, preferably have an interval scale of measurement

with a normal distribution. Finally, percentile ranks are provided based on the population and clinical samples.

Methods

Datasets

BSI and 4DSQ data from the general population were collected on separate occasions in a large sample of the Dutch population, called the LISS panel (Long-term Internet Studies for the Social Sciences). The LISS panel is maintained by CentERdata, a research institute located on the Tilburg University campus, and includes about 5,300 household respondents who were approached with the help of the municipal population register (van der Laan, 2009). The sample is representative of the Dutch population (Scherpenzeel, 2018; Scherpenzeel & Bethlehem, 2011). In 2007/2008, one-third of all households were approached, of which one person each completed the BSI (n = 1,662). This sample was stratified by age (four strata: 18–29, 30–49, 50–64, 65+ years), gender, and ethnic origin. In 2013, normative data on the 4DSQ were collected from the entire LISS panel (n = 5,273). The sample and the procedure for collecting 4DSQ data were described by Terluin and colleagues (2016).

For the OQ-45, data were used from n = 1,810 respondents, which have been described by Timman and colleagues (2017): 1,000 respondents came from a panel of the TNS-NIPO research agency and were stratified by gender, age, socioeconomic status, and education level; 810 came from an earlier validation study (de Jong et al., 2007), 448 came from a sample drawn from the telephone directory, and 362 were invited via internal mail from 14 companies or non-commercial organizations (Timman et al., 2017).

Patient data for the BSI came from a dataset of patients referred for treatment of depression, anxiety disorders, or somatoform disorders at RijnVeste, the outpatient clinic of GGZ Rivierduinen in the city of Leiden. Data from n = 4,853 patients collected between 2002 and 2013 were used. The procedure of data collection has been described by de Beurs and colleagues (2011). For the 4DSQ, patient data were used from 199 patients from primary care physicians. All patients were on sick leave, reported elevated levels of stress, and participated in a trial evaluating an intervention for stress-related mental disorders (Bakker et al., 2007). The patient data for the OQ-45 came from 12,436 patients described by Timman and colleagues (2017). This comprised a mixed sample: patients in daycare (n = 481) or inpatient care (n = 484) from various MHC institutes; patients in outpatient care (basic and specialized MHC, n = 1,581 and n = 9,433 respectively), and patients treated by private practitioners (n = 457). According to Dutch law, anonymized questionnaire data collected to support treatment may be used for scientific research and such use is exempt from an informed consent procedure.

Measures

The Brief Symptom Inventory (BSI; Derogatis, 1975; Dutch version: de Beurs & Zitman, 2006) consists of 53 items describing symptoms. Respondents can indicate to what extent they were bothered by each symptom on a Likert-type scale from 0 (= *not at all*) to 4 (= *very much*) in the past week. In this study, we limited ourselves to investigating the three most important scales of the BSI: depression (BSI-DEP), anxiety (BSI-ANX), and somatic complaints (BSI-SOM). Scale scores are calculated as the mean score of the comprising items and range from 0 to 4. In addition, the global score for the severity of psychopathology was analyzed: the mean score on all 53 items (Global Severity Index or BSI-GSI, range: 0–4).

The Four-Dimensional Symptom Questionnaire (4DSQ; Terluin et al., 2006) consists of 50 items, each describing one symptom. Respondents can indicate how often they experienced the symptom in the past week on a scale from 0 (= *no*) to 4 (= *very often or constantly*). The 4DSQ comprises four scales: general distress (4DSQ-DIS, 16 items, range: 0–32), depression (4DSQ-DEP, 6 items, range: 0–12), anxiety (4DSQ-ANX, 12 items, range 0–24), and somatic complaints (4DSQ-SOM, 16 items, range 0–32). All scores are sum scores (after recoding the two response options for high frequency: 4 = 2 and 3 = 2), as recommended in the scoring instruction of this instrument (Terluin et al., 2006).

The Outcome Questionnaire (OQ-45; Lambert et al., 2004; Dutch version: de Beurs et al., 2005; de Jong et al., 2007) comprises 45 items describing symptoms or problems. The respondent is asked to indicate how often these emerged during the past week on a scale from 0 (= *never*) to 4 (= *almost always*). A total score can be calculated (OQ-TOTAL, 45 items, range: 0–180) and four subscale scores: Symptom Distress (OQ-SD, 25 items, range: 0–100), Interpersonal Relations (OQ-IR, 11 items, range: 0–4), Social Role, (OQ-SR, 9 items, range: 0–36), and Anxiety and Social Distress (OQ-ASD, 13 items, range: 0–52). All scores are sum scores.

Statistical Analysis

Calculations According to Item Response Theory

The "Graded Response Model for polytomous items" and its Expected A Posteriori (EAP) score was used as the estimator for the θ-score with the multidimensional IRT (mirt) package (Chalmers, 2012) version in R. There are other

Table 1. Raw scores, θ-based, and calculated *T*-scores of the population sample and patients on BSI, 4DSQ, and OQ-45

	Raw scores					θ-based *T* scores					Calculated *T* scores				
	M	*Mdn*	*SD*	Skew.	Kurt.	*M*	*Mdn*	*SD*	Skew.	Kurt.	*M*	*Mdn*	*SD*	Skew.	Kurt.
BSI															
Population															
GSI	0.38	0.28	0.34	2.00	5.97	49.81	49.95	9.17	0.16	−0.07	50.37	50.38	9.26	0.24	−0.06
DEP	0.39	0.17	0.48	2.20	6.75	49.94	49.35	8.48	0.65	0.05	51.09	48.05	8.74	0.76	0.32
ANX	0.34	0.17	0.43	2.02	5.23	49.97	48.50	8.17	0.66	−0.34	52.04	50.02	7.94	0.58	−0.41
SOM	0.33	0.14	0.42	2.30	7.50	49.98	48.03	7.96	0.75	0.19	51.67	49.17	8.22	0.77	0.33
Patients															
GSI	1.21	1.09	0.73	0.67	−0.02	66.71	66.62	11.36	0.02	0.25	66.57	66.36	11.49	−0.02	0.26
DEP	1.57	1.50	1.03	0.37	−0.80	67.94	67.95	12.10	−0.01	−0.41	67.54	68.33	12.42	−0.11	−0.41
ANX	1.35	1.17	0.96	0.63	−0.41	66.14	66.18	9.94	0.11	−0.41	65.46	65.77	10.55	−0.11	−0.24
SOM	0.95	0.71	0.83	1.06	0.67	62.17	61.83	11.06	0.32	−0.23	61.60	60.98	11.41	0.28	−0.24
4DSQ															
Population															
DIST	5.52	3	6.40	1.72	2.87	50.00	49.63	9.20	0.50	−0.11	50.03	49.07	9.20	0.53	−0.01
DEP	0.70	0	1.92	3.69	14.71	50.00	46.57	7.05	1.88	2.44	50.17	46.77	6.97	1.87	2.38
ANX	1.05	0	2.53	4.14	21.64	50.00	45.48	7.46	1.56	1.77	50.13	45.58	7.42	1.52	1.69
SOM	4.92	4	4.91	1.58	3.07	50.00	49.68	8.87	0.42	−0.10	50.11	50.79	8.87	0.42	−0.05
Patients															
DIST	19.02	20	8.32	−0.24	−0.93	66.90	66.77	8.32	0.12	−0.38	66.19	66.43	8.55	−0.16	−0.32
DEP	3.45	2	3.66	1.06	0.05	62.77	61.97	6.45	0.65	−0.24	60.39	61.66	9.52	−0.16	−0.92
ANX	5.21	4	5.14	1.06	0.43	63.38	63.16	7.28	0.37	−0.58	61.28	63.12	9.59	−0.16	−0.67
SOM	12.77	11	7.00	0.46	−0.51	63.82	62.90	8.20	0.35	−0.04	62.36	60.96	8.95	0.14	−0.12
OQ-45															
Population															
OQ-TOT	40.83	39	17.80	0.82	1.15	49.62	49.49	9.03	0.26	0.75	50.43	50.25	9.41	0.32	0.42
SDIS	23.76	22	11.43	0.83	1.09	49.76	49.58	8.98	0.30	0.78	50.84	50.13	9.50	0.39	0.37
IR	8.96	8	4.97	0.81	0.93	49.83	49.69	8.68	0.31	−0.02	51.78	50.69	8.69	0.38	0.03
SR	8.12	8	3.88	0.64	0.71	49.81	50.44	7.94	0.10	−0.22	50.46	50.39	8.39	0.49	0.42
ASD	14.37	14	6.82	0.59	0.45	49.84	49.73	8.60	0.11	0.38	51.67	51.81	9.52	0.26	0.21
Patients															
OQ-TOT	78.23	79	24.00	0.00	−0.24	68.52	68.89	10.96	−0.10	0.17	68.34	68.95	10.85	−0.17	0.07
SDIS	48.24	49	15.51	−0.06	−0.28	69.63	69.88	11.35	−0.07	0.14	69.45	70.09	11.33	−0.13	0.14
IR	16.54	16.5	6.79	0.11	−0.35	64.31	64.93	10.46	−0.12	−0.15	63.95	64.62	10.35	−0.22	−0.19
SR	13.35	13	5.42	0.20	−0.27	61.58	62.13	11.77	−0.05	−0.27	61.42	61.05	11.12	0.04	−0.40
ASD	25.62	26	8.86	0.02	−0.29	66.81	66.91	11.37	0.01	0.17	66.43	67.00	11.40	−0.03	0.09

Note. Brief Symptom Inventory (BSI): GSI = Global Severity Index; DEP = Depression; ANX = Anxiety; SOM = Somatic complaints; Four-Dimensional Symptom Questionnaire (4DSQ): DIST = Distress; DEP = Depression; ANX = Anxiety; OM = Somatization; Outcome Questionnaire (OQ): TOT = OQ45 Total score; SDIS = Symptomatic Distress; IR = Interpersonal Relationships; SR = Social Role; ASD = Anxiety and Social Distress; *M* = Mean; *Mdn* = Median; *SD* = Standard deviation; Skew. = Skewness; Kurt. = Kurtosis.

estimates available for the latent variable scores in IRT, such as Maximum Likelihood (ML) and Weighted Likelihood Estimates (WLE). We performed a sensitivity analysis, comparing EAP with these alternatives, and found sufficiently similar results regarding the mean θ's, with their 95% confidence intervals (CI95) mostly overlapping and highly correlated (results are provided in Table C in the supplementary materials; de Beurs, Oudejans, et al., 2022). Item parameters were established in the combined general population and clinical samples. Thus, θ scores were estimated with the multigroup mirt option (Smits, 2016). We fixed the item parameters to be equal across groups. The latent trait (θ) was standardized to a scale with a mean of 0 and a standard deviation of 1 for the general population. The unidimensionality of scales, a requirement for IRT, was investigated with confirmatory factor analysis using the R package lavaan (version 06.5; Rosseel, 2012). We used the DWLS estimator based on the polychoric correlation matrix for ordinal items and inspected (scaled) fit statistics and set as requirements for unidimensionality:

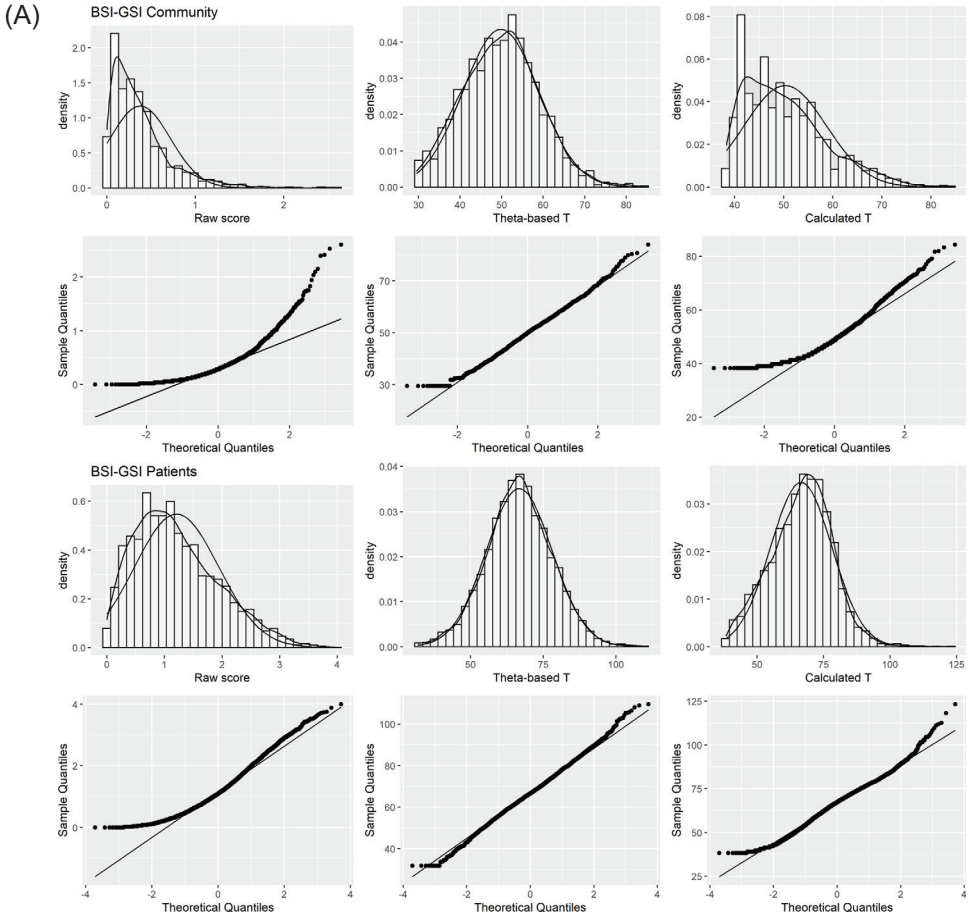

Figure 2. Frequency distribution (histograms with density and normal curves) and QQ-plots of scores for raw scores above, θ-based *T* scores, and calculated *T* scores of the BSI General Severity Index (BSI-GSI; A), 4DSQ-Distress (4DSQ-DIS; B), and OQ-Symptomatic Distress (OQ-SD; C) of the general population (above) and patient samples (below).

CFI > .95, TLI ≥ 0.95, RMSEA < .06, and SRMR < .08. If insufficient fit of a unidimensional model is found, IRT based scores are potentially flawed, and alternatives (e.g., percentile conversion or regression-based norming) can be utilized.

We used non-linear least squares modeling of R (nls) to establish the best fitting function (which had the lowest AIC value and/or the most parsimonious number of coefficients) for the relation between raw scores and θ-based *T*-scores. Linear, polynomial, exponential, logarithmic, power, division, rational, sigmoid, and hyperbolic equations were evaluated. We cross-validated each equation by randomly splitting each dataset in two and using the first dataset to establish the best-fitting function and using the second dataset as a validation sample (Camstra & Boomsma, 1992). Applying the conversion formula to the raw scale score results in a calculated *T* score. The distributions of the resulting scores (mean, median, standard deviation, skewness, and kurtosis) were investigated for the population and the patient samples and visually inspected on normality with histograms/density plots and QQ-plots. ICC estimates for absolute agreement and consistency of θ-based and calculated *T* scores and their CI95 were determined using R and based on a two-way mixed-effects model. We established "bias" (mean difference between both *T* scores (van Stralen et al., 2012), as well as the "percentage error" (the width of the CI95 interval proportional to the population mean; Van Hoeck et al., 2000). If the CI95 interval was within 5 *T* score points, we would conclude that both approaches yielded sufficiently similar results since 5 *T* score points are the proposed limit for statistically reliable change in score over time (de Beurs et al., 2019). We also inspected the agreement between θ-based *T* scores and calculated *T* scores with Bland-Altman plots for the full range of severity (Bland & Altman, 1986). We did this for the entire sample and for the population and clinical samples separately. Finally, to establish the effect of normalization, we also compared standard

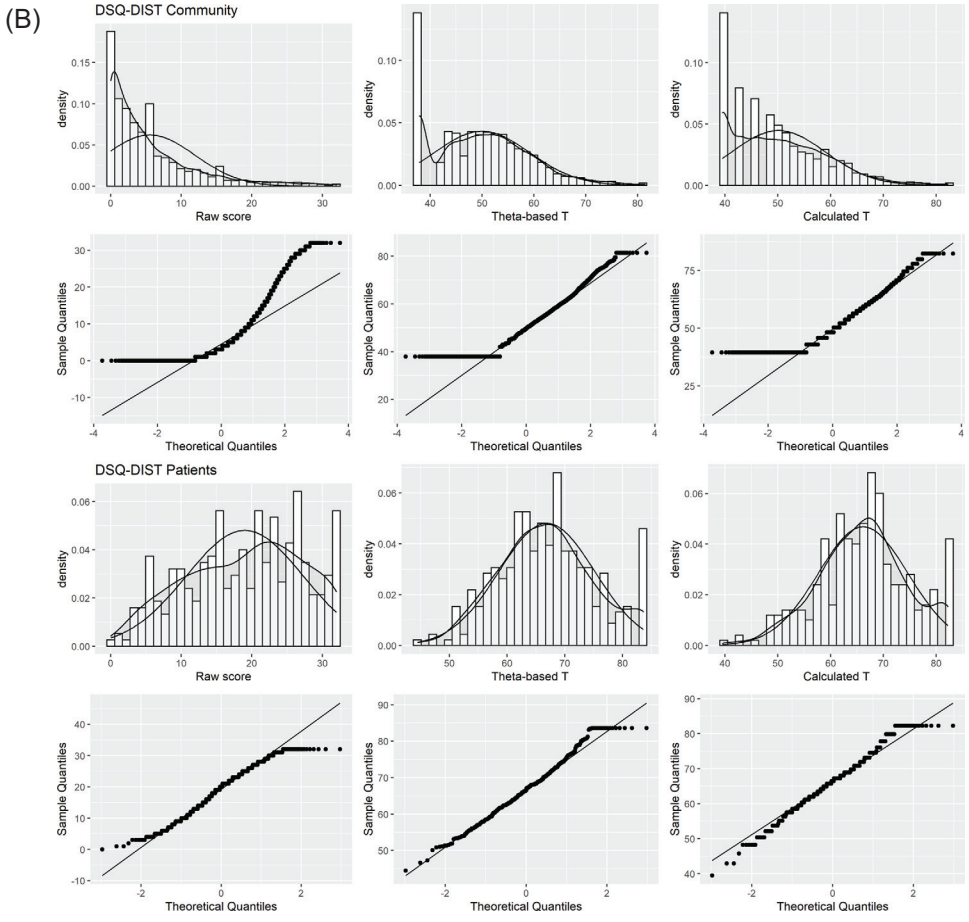

Figure 2. Continued.

T scores (based on the standard conversion formula "standard $T = 10 \times Z + 50$") with θ-based *T* scores and calculated *T* scores.

Results

Unidimensionality of Scales

Most scales met criteria for unidimensionality, using CFI $\geq$ 0.95, TLI $\geq$ 0.95, RMSEA < 0.06 to 0.08; SRMR $\leq$ .08 (Schreiber et al., 2006). Table A in the supplementary materials provides CFA results for all scales (de Beurs, Oudejans, et al., 2022). The CFI, TLI, and SRMR requirements were met by all scales of the BSI and the 4DSQ, except for the BSI-GSI and the DIS-SOM. The OQ scales did not meet CFI and TLI requirements. RMSEA was larger than 0.06 for most scales but not substantially larger, except for the BSI-ANX and – again – the OQ-SR (RMSEA > 0.12). As a consequence, the precision of the θ's for the OQ scales, in particular, may be compromised, as these scales appear to lack sufficient unidimensionality (Crişan et al., 2017).

Distribution of Raw Scores and *T* Scores

Table 1 presents an overview of raw scores, θ-based *T* scores, and calculated *T* scores and characteristics of the frequency distribution for scales of the BSI, 4DSQ, and OQ-45. Figure 2 shows the main score on each instrument, the frequency distribution of the raw score, θ-based *T* score, and calculated *T* score in the general population sample (upper half) and the clinical sample (lower half), and QQ-plots. The Similarity of the frequency distribution, the density curve, and the normal curve and accordance with the dots in the QQ-plot and the straight diagonal line, indicate how close the distributions approximate normality. (Figure A1–A3 in the supplementary materials presents plots for all scales; de Beurs, Oudejans, et al., 2022). The average BSI-GSI score in the population sample was $M = 0.38$ ($SD = 0.34$). For the BSI-GSI, score skewness was 2.00; kurtosis was 5.97, showing substantial deviation from the normality

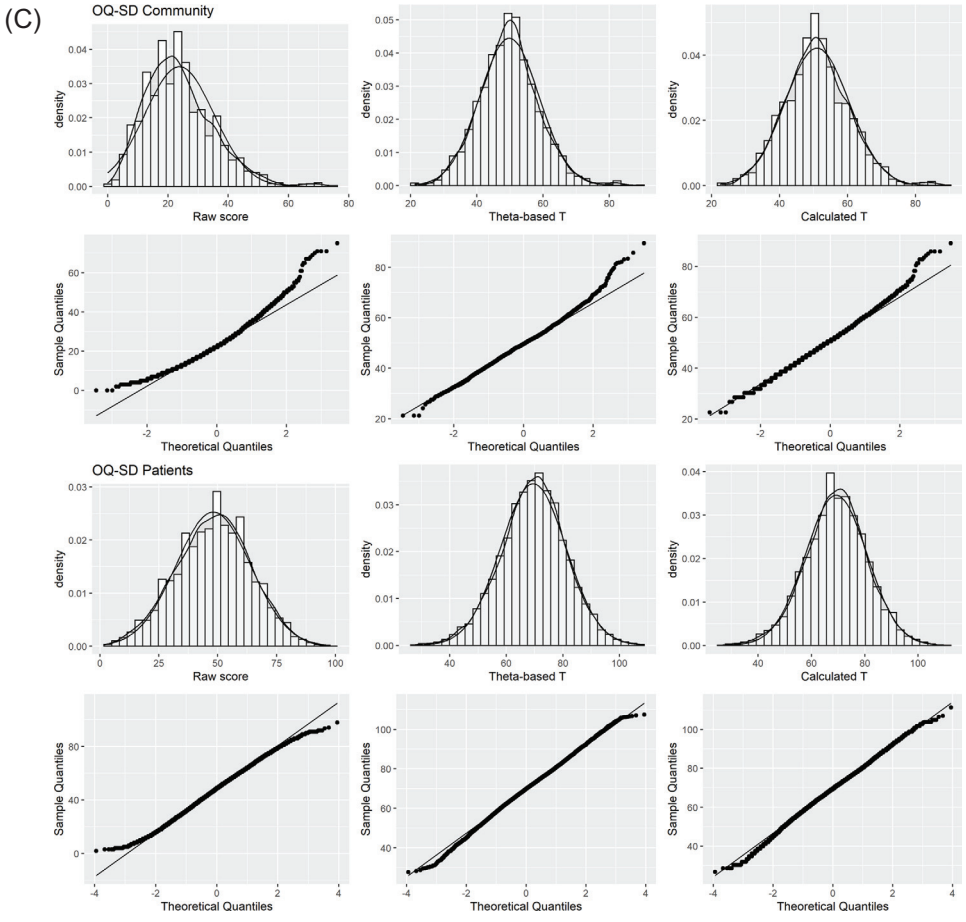

Figure 2. Continued.

of the scores with a tail to the right. Many respondents had a low score, but this is to be expected when a symptom list is completed by a population sample. The raw scores on the 4DSQ scales in the population were also skewed. For the depression and anxiety scale, values for skewness and kurtosis were extreme, as 78.7% and 67.6% of the respondents had the lowest possible score. In contrast, raw scores of the general population on the OQ-45 had an almost normal distribution; only the total score and the OQ-SD score showed marginal kurtosis in the population sample (some surplus of scores below the average score). Most *T* scores based on θ's and calculated *T* scores had a normal distribution.

Conversion functions for the BSI, 4DSQ, and OQ-45 are shown in Table 2. For BSI scores rational functions provided the best fit; indices of skewness and kurtosis decreased considerably compared to the raw scores. Raw scale scores in the patient sample had a normal distribution, and this was preserved in the calculated *T* scores.

For 4DSQ scales, also rational functions best fit the relation between raw scores and θ-based *T* scores. Again, raw scores were skewed and peaked, whereas θ-based and calculated *T* scores approximated the normal distribution better, although the depression score and the anxiety score of the population sample were still skewed and showed kurtosis due to a large proportion of respondents with the lowest possible score on these scales. Thus, transformation to a normal distribution was successful for only two of the four subscales of the 4DSQ. Patient scores had a normal distribution. Finally, for the OQ-45, two rational functions, a cubic, a quadratic, and a hyperbolic function, provided the best fit.

In the supplementary materials, Figure A1–A3 (de Beurs, Oudejans, et al., 2022) presents histograms with density lines and QQ-plots for all scales. These graphs reveal a sufficiently normal distribution for most scales, except for the 4DSQ depression and anxiety scales, but also reveal some surplus of extreme high and extreme low scores for all scales.

Comparison of θ-Based and Calculated *T* Scores

We cross-validated the equations shown in Table 2 by applying a random split of each dataset into a calibration

Table 2. Formulas to calculate *T*-scores and indicators of correspondence from cross-validation

Scale	Formula $y = \ldots$	RMSE	R^2	MAE
BSI-GSI	$31.1 + (138.641x + 22.4779x^2)/(1 + 4.089x - 0.3392x^2)$	1.456	0.988	1.116
BSI-DEP	$41.6 + (47.593x + 0.7650x^2)/(1 + 1.444x - 0.1909x^2)$	1.627	0.986	1.304
BSI-ANX	$43.4 + (50.294x - 1.9425x^2)/(1 + 1.564x - 0.2368x^2)$	2.658	0.950	2.189
BSI-SOM	$42.9 + (56.622x + 9.3246x^2)/(1 + 2.233x - 0.1947x^2)$	1.725	0.978	1.396
4DSQ-DIST	$38.0 + (6.460x - 0.0284x^2)/(1 + 0.258x - 0.0050x^2)$	1.087	0.988	0.788
4DSQ-DEP	$46.8 + (27.876x - 0.8495x^2)/(1 + 1.411x - 0.0766x^2)$	1.497	0.961	0.854
4DSQ-ANX	$45.6 + (15.735x + 0.0349x^2)/(1 + 0.718x - 0.0155x^2)$	1.875	0.946	1.222
4DSQ-SOM	$37.4 + (5.746x + 0.0336x^2)/(1 + 0.203x - 0.0031x^2)$	1.208	0.984	0.845
OQ-TOT	$20.9 + (1.157x + 0.0017x^2)/(1 + 0.018x - 0.0001x^2)$	2.434	0.963	1.895
OQ-SD	$25.8 + (1.487x - 0.0051x^2)/(1 + 0.014x - 0.0001x^2)$	2.287	0.970	1.796
OQ-IR	$33.5 + 2.46x - 0.0447x^2 + 0.00058x^3$	3.370	0.914	2.649
OQ-SR	$32.1 + 2.39x - 0.0121x^2$	3.879	0.896	3.016
OQ-ASD	$68.2 + 1.094 \times (x - 27.0) + \sinh((x - 27.0)/8.533)$	2.350	0.965	1.840

Note. Brief Symptom Inventory (BSI): GSI = Global Severity Index; DEP = Depression; ANX = Anxiety; SOM = Somatic Complaints; Four-Dimensional Symptom Questionnaire (4DSQ): DIST = Distress; DEP = Depression; ANX = Anxiety; SOM = Somatization; Outcome Questionnaire (OQ): TOT = Total score; SD = Symptomatic Distress; IR = Interpersonal Relationships; SR = Social Role; ASD = Anxiety and Social Distress; y = calculated *T*-score; x = raw scale score; RMSE = Root Mean Squared Error; R^2 = coefficient of determination; MAE = Mean Absolute Error.

Table 3. Indicators of correspondence between θ-based and calculated *T*-scores

Scale	ICC-A	CI95	ICC-C	CI95	Bias	Perc. Err.	CI95+	CI95−	> 5 %	Difference −5 and 5 %	< −5 %
BSI-GSI	.99	.99–.99	.99	.99–.99	−0.02	12.3	−3.87	3.83	2.37	97.60	0.03
BSI-DEP	.99	.99–.99	.99	.99–.99	0.01	11.4	−3.61	3.62	0.82	99.04	0.14
BSI-ANX	.97	.97–.97	.97	.97–.97	−0.02	17.8	−5.56	5.51	3.45	93.54	3.01
BSI-SOM	.99	.98–.99	.99	.98–.99	−0.01	13.1	−3.88	3.86	1.36	98.63	0.02
4DSQ-DIST	.99	.99–.99	.99	.99–.99	0.02	11.3	−2.88	2.91	0.00	99.98	0.02
4DSQ-DEP	.98	.98–.98	.98	.98–.98	−0.03	11.8	−3.03	2.98	0.00	98.36	1.64
4DSQ-ANX	.99	.99–.99	.99	.99–.99	0.00	9.9	−2.52	2.52	0.41	98.56	1.03
4DSQ-SOM	.99	.99–.99	.99	.99–.99	−0.02	9.9	−2.54	2.50	0.00	99.82	0.18
OQ-TOT	.98	.98–.98	.98	.98–.98	0.02	14.7	−4.79	4.83	1.77	95.44	2.79
OQ-SD	.98	.98–.99	.98	.98–.99	−0.02	13.5	−4.52	4.49	1.32	96.89	1.79
OQ-IR	.96	.95–.96	.96	.95–.96	−0.01	21.3	−6.61	6.60	6.89	86.44	6.67
OQ-SR	.95	.95–.95	.95	.95–.95	0.03	25.1	−7.47	7.52	10.14	81.33	8.53
OQ-ASD	.98	.98–.98	.98	.98–.98	0.02	14.7	−4.68	4.72	1.92	96.03	2.05

Note. Brief Symptom Inventory (BSI): GSI = Global Severity Index; DEP = Depression; ANX = Anxiety; SOM = Somatic complaints; Four-Dimensional Symptom Questionnaire (4DSQ): DIST = Distress; DEP = Depression; ANX = Anxiety; OM = Somatization; Outcome Questionnaire (OQ): TOT = OQ45 Total score; SD = Symptomatic Distress; IR = Interpersonal Relationships; SR = Social Role; ASD = Anxiety and Social Distress; ICC-A = Absolute agreement; ICC-C = consistency; Bias = difference between both *T*-scores (positive bias values indicates that θ-based *T*-scores are higher than calculated *T*-scores); Perc. Err. = Percentage error, the width of the limits of agreement interval divided by the mean *T*-score of the population [(CI95− + CI95+)/*M*]; "Difference −5 and 5 %" = The percentage of subjects for whom both *T*-scores differ less than 5 points; "Difference < −5 and > 5 %" = The percentage of subjects for whom the difference between both methods is greater than 5 points.

sample and a cross-validation sample. Table 2 provides the Root Mean Squared Error (RMSE), the coefficient of determination (R^2), and the Mean Absolute Error (MAE) for the correspondence of predicted scores (calculated *T* scores with formulas based on the calibration sample) with observed *T* scores (θ-based *T* scores in the cross-validation sample). Overall, correspondence was high; the lowest $R^2 = .869$ for the OQ-SR scale.

Furthermore, we calculated ICCs for absolute agreement between θ-based and calculated *T* scores and consistency (similar ranking of subjects according to both scores; van Stralen et al., 2012). Table 3 presents the ICCs and

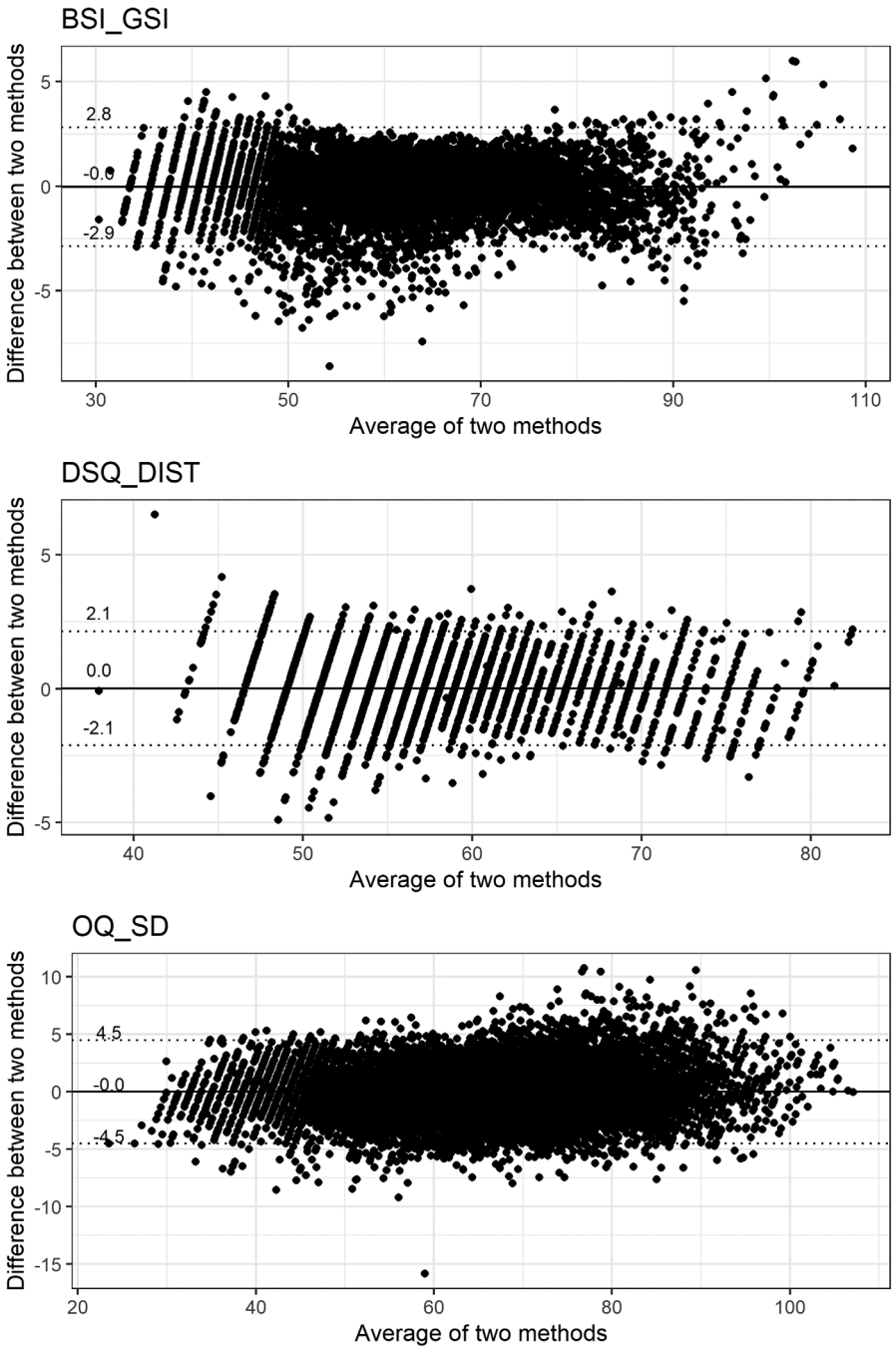

Figure 3. A Bland-Altman plot displays the difference between two methods for the full range of scores; the *x*-axis displays the range based on the average of both methods; the *y*-axis displays for each subject the difference in score between both methods. BSI-GSI = Global Severity Index; DSQ-DIST = 4DSQ Distress; OQ-SD = Symptomatic Distress.

additional information on the association between both scores. All ICCs were high, suggesting excellent agreement and consistency. Bias and percentage error were also low, with the exception of a higher percentage error for BSI-ANX, OQ-IR, and OQ-SR, and for these scales, the CI95 extended beyond 5 *T* score points. Figure 3 presents Bland-Altman plots for selected scales. Figures B1–B3 in the supplementary materials present plots for all scales (de Beurs, Oudejans, et al., 2022).

For the BSI, θ-based *T* corresponded well to calculated *T* scores (mean difference *M* = 0.01–0.02, the solid gray line in the BA plot); only for the BSI-ANX, less than 95%

Table 4. Indicators of correspondence of standardized *T* scores with θ-based and calculated *T* scores for the main scale of each self-report measure

Scale	Pair	ICC-A	CI95	ICC-C	CI95	Bias	Perc. Err.	CI95+	CI95−	> 5 %	Difference −5 and 5 %	< −5 %
BSI-GSI	ST-TT	.93	.93–.93	.93	.93–.93	−0.29	35.6	−11.42	10.83	18.89	67.92	13.19
	ST-CT	.93	.93–.94	.93	.93–.94	−0.26	34.3	−11.01	10.49	18.90	75.59	5.51
4DSQ-DIST	ST-TT	.93	.93–.94	.94	.93–.94	−0.78	29.7	−8.40	6.85	10.22	89.16	0.62
	ST-CT	.94	.93–.94	.94	.94–.94	−0.79	28.5	−8.10	6.52	9.72	90.28	0.00
OQ-SD	ST-TT	.97	.96–.98	.98	.97–.98	−1.00	18.3	−7.12	5.12	11.27	87.13	1.60
	ST-CT	.99	.98–.99	.99	.99–.99	−0.99	12.5	−5.16	3.19	1.00	99.00	0.00

Note. BSI-GSI = Brief Symptom Inventory – Global Severity Index; 4DSQ-DIST = Four-Dimensional Symptom Questionnaire – Distress; OQ-SD = Outcome Questionnaire – SD = Symptomatic Distress; ICC-A = Absolute agreement; ICC-C = Consistency; Bias = Difference between both *T*-scores (positive bias values indicates that θ-based *T*-scores are higher than calculated *T*-scores); Perc. Err. = Percentage error, the width of the limits of agreement interval divided by the mean *T*-score of the population [(CI95− + CI95+)/*M*]; "Difference −5 and 5 %" = The percentage of subjects for whom both *T*-scores differ less than 5 points; "Difference < −5% and > 5 %" = The percentage of subjects for whom the difference between both methods is greater than 5 points; ST = Standardized *T*-score; CT = Calculated *T*-score; TT = θ-based *T*-score.

Table 5. Crosswalk table for a selection of raw scores to calculated *T*-scores and percentile scores for the BSI global severity index, depression, anxiety, and somatic complaints scale

Global severity index				Depression				Anxiety				Somatic complaints			
RS	*T*-score	PRn	PRcl	RS	*T*-score	PRn	PRcl	RS	*T*-score	PRn	PRcl	RS	*T*-score	PRn	PRcl
0.00	31.1	13.2	2.1	0.00	41.6	16.1	2.6	0.00	43.4	17.1	2.9	0.00	42.9	45.9	27.4
0.17	45.5	52.9	8.3	0.17	48.2	41.5	7.5	0.17	50.1	44.0	8.9	0.17	50.1	91.8	54.8
0.33	52.0	79.3	12.5	0.33	52.5	58.4	12.2	0.33	54.4	62.0	15.3	0.33	54.4	91.8	54.8
0.50	56.4	79.3	12.5	0.50	55.9	71.5	17.7	0.50	57.7	75.5	22.0	0.50	57.7	91.8	54.8
0.67	59.8	79.3	12.5	0.67	58.7	80.6	23.7	0.67	60.3	83.4	28.8	0.67	60.4	91.8	54.8
0.83	62.5	83.3	21.7	0.83	61.0	86.7	29.4	0.83	62.3	88.2	36.1	0.83	62.5	91.8	54.9
1.00	65.0	90.2	40.2	1.00	63.1	90.6	35.1	1.00	64.2	91.7	43.4	1.00	64.6	95.6	69.3
1.17	67.4	96.0	59.8	1.17	65.0	92.9	40.9	1.17	65.8	94.1	50.1	1.17	66.5	99.3	83.6
1.33	69.5	98.9	70.2	1.33	66.6	94.5	46.4	1.33	67.2	95.8	56.2	1.33	68.2	99.3	83.7
1.50	71.7	98.9	70.2	1.50	68.3	96.1	51.6	1.50	68.6	97.2	61.6	1.50	69.9	99.3	83.9
1.67	73.9	98.9	70.2	1.67	70.0	97.2	56.6	1.67	70.0	98.1	66.6	1.67	71.7	99.3	84.0
1.83	75.9	99.4	75.5	1.83	71.5	97.9	61.7	1.83	71.2	98.6	71.2	1.83	73.3	99.3	84.0
2.00	78.1	100.0	84.9	2.00	73.1	98.5	66.6	2.00	72.6	99.1	75.2	2.00	75.0	99.7	89.7
2.17	80.3	100.0	92.1	2.17	74.7	98.9	71.0	2.17	73.9	99.4	78.9	2.17	76.7	100.0	95.4
2.33	82.4	100.0	95.1	2.33	76.2	99.1	75.1	2.33	75.1	99.5	82.4	2.33	78.3	100.0	95.4
2.50	84.6	100.0	95.1	2.50	77.8	99.4	78.9	2.50	76.5	99.7	85.5	2.50	80.1	100.0	95.4
2.67	86.9	100.0	95.1	2.67	79.5	99.5	82.5	2.67	77.9	99.8	88.3	2.67	81.9	100.0	95.4
2.83	89.2	100.0	95.5	2.83	81.2	99.6	86.0	2.83	79.3	99.9	90.7	2.83	83.6	100.0	95.5
3.00	91.6	100.0	97.0	3.00	83.0	99.8	89.1	3.00	80.8	100.0	92.9	3.00	85.5	100.0	97.4
3.17	94.2	100.0	98.9	3.17	84.9	99.9	91.6	3.17	82.5	100.0	94.7	3.17	87.5	100.0	99.3
3.33	96.6	100.0	99.6	3.33	86.8	99.9	94.1	3.33	84.1	100.0	96.2	3.33	89.4	100.0	99.3
3.50	99.3	100.0	99.6	3.50	89.0	99.9	96.1	3.50	86.0	100.0	97.6	3.50	91.4	100.0	99.3
3.67	102.1	100.0	99.6	3.67	91.2	100.0	97.6	3.67	88.0	100.0	98.6	3.67	93.6	100.0	99.3
3.83	104.8	100.0	99.6	3.83	93.5	100.0	98.6	3.83	90.1	100.0	99.3	3.83	95.7	100.0	99.3
4.00	107.8	100.0	99.8	4.00	96.1	100.0	99.5	4.00	92.4	100.0	99.8	4.00	98.0	100.0	99.6

Note. Brief Symptom Inventory (BSI): GSI = Global Severity Index; DEP = Depression; ANX = Anxiety; SOM = Somatic complaints; RS = Raw Score; *T* = Calculated *T*-score; PRn = Percentile rank in the normal population; PRcl = Percentile rank in the clinical population.

of the cases fell within the −5 to 5 interval for acceptable difference. For the 4DSQ θ-based *T* scores and calculated *T* scores corresponded well (mean difference *M* = 0.02), 99% of the cases fell within the −5 to 5 interval. For the OQ-SD also, excellent correspondence was found (mean difference *M* = 0.02); however, on average, 91.2% of the

Table 6. Cross walk table for a selection of raw scores to calculated *T*-scores and percentile scores for the 4DSQ scales for distress, depression, anxiety, and somatization

4DSQ-DIST				4DSQ-DEP				4DSQ-ANX				4DSQ-SOM			
RS	*T*	PRn	PRcl	RS	*T*	PRn	PRcl	RS	*T*	PRn	PRcl	RS	*T*	PRn	PRcl
0	38.0	10.4	0.2	0	46.8	39.5	13.3	0	45.6	42.6	16.2	0	37.4	14.6	0.3
2	46.6	37.7	1.1	1	58.3	82.8	33.8	2	58.9	89.4	41.6	2	45.8	39.5	2.5
4	51.0	54.9	4.0	2	61.7	88.7	46.7	4	63.1	94.9	55.6	4	50.8	57.8	9.2
6	54.0	67.2	8.1	3	63.5	91.9	57.5	6	65.7	96.9	66.2	6	54.4	71.5	18.7
8	56.2	75.5	11.8	4	64.8	93.9	66.5	8	67.9	98.0	75.4	8	57.3	81.2	29.9
10	58.1	81.7	17.7	5	66.0	95.4	72.5	10	69.9	98.7	82.1	10	59.8	87.9	42.7
12	59.8	86.0	22.8	6	67.1	96.6	77.8	12	71.9	99.1	88.0	12	62.1	91.9	51.7
14	61.4	89.4	29.7	7	68.3	97.4	83.1	14	74.0	99.4	92.2	14	64.3	94.5	58.9
16	63.0	92.0	36.5	8	69.6	97.8	86.1	16	76.2	99.5	94.7	16	66.5	96.6	68.2
18	64.7	93.8	42.4	9	71.1	98.3	87.7	18	78.7	99.7	96.6	18	68.7	97.8	76.5
20	66.4	95.2	50.0	10	72.8	98.7	89.8	20	81.6	99.8	98.3	20	70.9	98.6	83.0
22	68.3	96.3	59.6	11	74.9	99.2	92.4	22	84.8	99.9	99.4	22	73.3	99.2	88.5
24	70.3	97.3	68.3	12	77.5	99.7	96.7					24	75.8	99.5	93.6
26	72.6	98.2	76.3									26	78.4	99.7	96.1
28	75.1	98.8	83.6									28	81.2	99.8	97.8
30	78.0	99.5	89.0									30	84.3	99.9	99.2
32	81.3	99.9	96.8									32	87.6	100.0	99.7

Note. Four-Dimensional Symptom Questionnaire (4DSQ): DIST = Distress; DEP = Depression; ANX = Anxiety; OM = Somatization; RS = Raw Score; *T* = Calculated *T*-score; PRn = Percentile rank in the normal population; PRcl = Percentile rank in the clinical population.

cases fell within the −5 to 5 interval with less correspondence for the OQ-IR and OQ-SR scales. Generally, in the low scoring range (< 40) and in the high score range (> 70), calculated *T* scores were somewhat higher; in the mid-range θ-based, *T* scores were higher. We also established ICC's denoting correspondence between θ-based and calculated *T* scores for the population and the clinical sample separately. Correspondence was somewhat lower in the clinical sample, especially for the 4DSQ-DEP, 4DSQ-ANX, OQ-IR, and OQ-SR, but was still very high (see Table B in the supplementary materials; de Beurs, Oudejans, et al., 2022).

We also compared the θ-based and the calculated *T* scores for the BSI-GSI, 4DSQ-DIS, and OQ-SD with standard *T* scores (based on the simpler linear equation $T = 10 \times Z + 50$). ICCs ranged from ICC = .81 for the BSI-GSI to ICC = .99 for the OQ-SD (see Table 4). Correspondence between standard *T* scores and calculated *T* scores was still substantial, especially for the OQ-SD. However, for the BSI-GSI, there was a large subgroup with higher standard *T* scores than θ-based or calculated *T* scores (for both comparisons $\Delta < -5$; 18.9%), which is understandable, as negative skewness (a low mean score relative to the maximum scale score) results in more respondents with extremely high standard *T* scores. After all, the non-linear conversion formula yielding calculated *T* scores corrects for precisely this undesirable effect.

Crosswalk From Raw Scores to *T* Scores and Percentiles

Table 5, 6, and 7 present crosswalk tables for the conversion of a selection of raw scores to calculated *T* scores and percentile ranks for the general population and the clinical population. Per measurement instrument, the first column gives raw scores (RS) and the second *T* scores are calculated according to the functions in Table 2 for all respondents. The conversion functions stretch the scales at the extremes: a change in raw score of one scale point in the low or high score area is larger than a change of one scale point in the mid-range score area, and the normalized *T* score reflects this appropriately. Figure 4 depicts the relationship between raw scores on the instruments and *T* scores.

Discussion

We analyzed community data of frequently used generic outcome measures in the Netherlands to link raw scores to two common metrics: *T* scores and percentile ranks. In line with previous research (Cook et al., 2015; Fischer & Rose, 2016; Friedrich et al., 2019; Schalet et al., 2015; Wahl et al., 2014), we applied methods based on IRT, which

Table 7. Crosswalk table for a selection of raw scores to calculated *T*-scores and percentile scores for the OQ total score, symptomatic distress, interpersonal relations, social role functioning, and anxiety and social distress

OQ-TOT				OQ-SD				OQ-IR				OQ-SR				OQ-ASD			
RS	*T*	PRn	PRcl	RS	*T*	PRn	PRcl	RS	*T*	PRn	PRcl	RS	*T*	PRn	PRcl	RS	*T*	PRn	PRcl
0	20.9	0.0	0.0	0	25.8	0.4	0.0	0	33.5	1.0	0.2	0	32.1	1.0	0.2	0	26.9	0.3	0.0
10	30.9	2.5	0.0	5	32.6	3.2	0.2	2	38.3	5.7	1.1	2	36.8	5.4	1.0	2	31.5	1.4	0.2
20	38.7	11.0	0.8	10	38.5	11.7	0.8	4	42.7	15.8	3.1	4	41.4	15.3	3.5	4	35.7	4.3	0.6
30	45.2	30.7	2.5	15	43.7	26.4	1.9	6	46.8	29.7	6.6	6	46.0	31.3	9.7	6	39.4	10.7	1.3
40	50.8	54.6	6.0	20	48.4	43.8	3.8	8	50.7	46.0	11.6	8	50.4	52.4	18.7	8	42.9	19.7	2.4
50	55.8	73.9	12.7	25	52.6	60.0	7.0	10	54.3	61.7	18.6	10	54.7	72.3	29.7	10	46.0	29.8	4.1
60	60.6	86.2	22.5	30	56.6	72.7	12.0	12	57.7	75.2	27.4	12	59.0	85.6	42.6	12	49.0	41.0	6.6
70	65.0	93.9	35.5	35	60.3	83.1	19.3	14	60.9	85.7	37.2	14	63.1	92.4	56.3	14	51.8	51.7	10.0
80	69.4	98.2	50.2	40	63.9	90.2	29.4	16	63.9	91.8	48.0	16	67.2	95.9	69.4	16	54.5	62.3	14.5
90	73.7	98.5	67.1	45	67.4	93.7	40.8	18	66.8	95.4	59.2	18	71.1	98.1	80.5	18	57.1	71.6	20.4
100	77.9	99.1	82.8	50	70.8	96.3	53.8	20	69.6	97.6	69.8	20	75.0	99.2	88.9	20	59.6	79.4	27.3
110	82.3	99.7	92.1	55	74.2	98.2	66.7	22	72.3	98.7	79.5	22	78.7	99.8	94.4	22	62.1	86.2	35.0
120	86.7	100.0	96.7	60	77.6	98.8	77.5	24	74.9	99.3	86.9	24	82.4	100.0	97.4	24	64.6	90.6	43.3
130	91.2	100.0	99.0	65	81.2	99.1	85.5	26	77.5	99.7	91.5	26	85.9	100.0	98.9	26	67.0	93.8	51.9
140	95.9	100.0	99.9	70	84.9	99.6	90.7	28	80.1	99.9	95.0	28	89.4	100.0	99.7	28	69.4	96.5	60.6
150	100.8	100.0	100.0	75	88.8	99.9	95.1	30	82.8	100.0	97.4	30	92.8	100.0	99.9	30	71.9	98.0	68.8
160	106.0	100.0	100.0	80	93.0	100.0	98.2	32	85.5	100.0	98.8	32	96.0	100.0	100.0	32	74.3	98.7	76.4
170	111.4	100.0	100.0	85	97.6	100.0	99.4	34	88.3	100.0	99.5	34	99.2	100.0	100.0	34	76.8	99.1	83.0
180	117.2	100.0	100.0	90	102.6	100.0	99.9	36	91.2	100.0	99.9	36	102.3	100.0	100.0	36	79.3	99.4	88.1
				95	108.3	100.0	100.0	38	94.3	100.0	100.0					38	81.9	99.6	92.0
				100	114.7	100.0	100.0	40	97.5	100.0	100.0					40	84.6	99.7	94.7
								42	100.9	100.0	100.0					42	87.4	99.8	96.7
								44	104.5	100.0	100.0					44	90.4	99.8	98.1
																46	93.6	99.9	99.0
																48	97.0	100.0	99.5
																50	100.8	100.0	99.8
																52	104.9	100.0	100.0

Note. Outcome Questionnaire (OQ): TOT = OQ45 Total Score; SD = Symptomatic Distress; IR = Interpersonal Relationships; SR = Social Role; ASD = Anxiety and Social Distress; RS = Raw Score; *T* = Calculated *T*-score; PRn = Percentile rank in the normal population; PRcl = Percentile rank in the clinical population.

resulted in θ-based *T* scores with a normal frequency distribution. We also determined functions to convert sum scores into *T* scores and showed that for most scales, calculated *T* scores approximated θ-based *T* scores very well, as the scores were strongly related (all ICC > .95, see Table 3). Scores were similar across the width of the entire scale and yielded similar mean values for the groups. Correspondence between calculated and θ-based *T* scores supports the validity of our approach to calculate *T* scores with a function based on curve fitting. However, at the extreme end of the scales, the two approaches diverged somewhat. Thus, caution with extreme scores is in order, especially with $T < 40$ and $T > 80$. Furthermore, the findings of two scales of the 4DSQ revealed that if raw scale scores are too skewed or too leptokurtic, to begin with (due to an excess of respondents with the lowest possible score), conversion to θ-based or calculated *T* scores will not yield scores with a normal frequency distribution.

Compared to standard *T* scores ($T = 10 \times Z + 50$), the correspondence of the more complex conversion formulas (correcting for a non-normal frequency distribution of raw scores) with θ-based *T* scores was better. For instance, the ICC between standard *T* scores and θ-based *T* scores for the BSI-GSI was ICC = .93 (see Table 4), whereas calculated *T* scores showed almost perfect correspondence with θ-based *T* scores (ICC = .99; see Table 3). Better correspondence was especially obtained in the higher score range. For each scale, we cross-validated the formulas on subsamples (splitting the samples randomly in half, and we also compared the population and clinical samples). The results revealed a similarity and high correlation between predicted scores from the "learning sample" with obtained scores in the "test sample". However, thus a substantial number of statistical tests were done per scale in each dataset, and the applicability of these formulas still needs further validation with data from other samples.

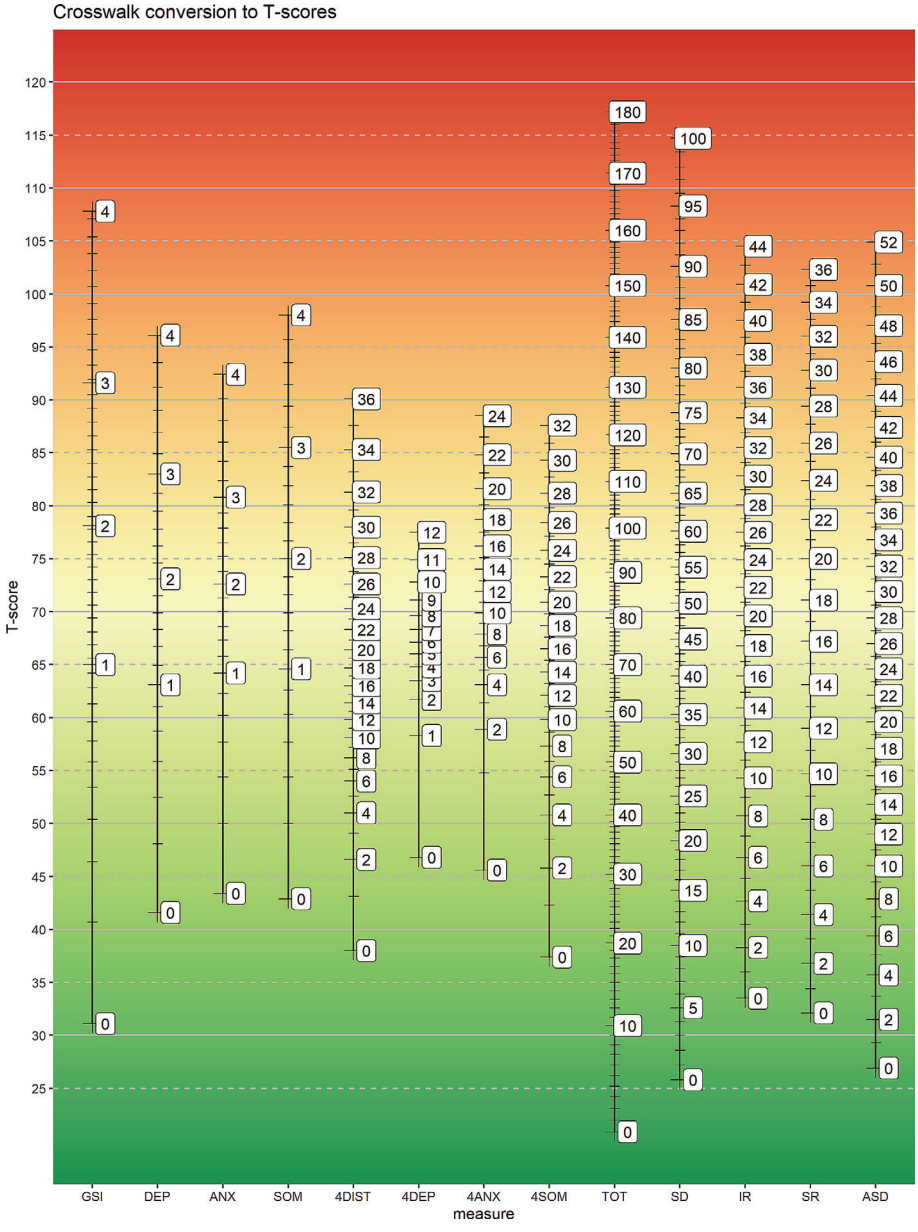

Figure 4. Linking raw scores to the *T* score metric for subscales of the BSI, 4DSQ, and OQ (GSI = Global Severity Index; DEP = Depression; ANX = Anxiety; SOM = Somatic complaints; 4DIST = Distress; 4DEP = Depression; 4ANX = Anxiety; 4SOM = Somatization; TOT = OQ-45 Total score; SD = Symptomatic Distress; IR = Interpersonal Relationships; SR = Social Role; ASD = Anxiety and Social Distress).

The results show that some scales evoked the lowest possible score from many respondents (especially the 4DSQ; see Figure 2). This zero inflation in the data is quite common when measures of psychopathology are administered in the general population. With zero-inflated data, the Graded Response Model may yield biased results of IRT analyses, and alternative models to deal with zero inflation have been proposed (Wall et al., 2015), such as Zero Inflated Mixture GRM (ZIM-GRM). In a simulation study with zero-inflated scores, GRM showed substantial bias (Smits et al., 2020). In future studies, the effect of zero inflation on factor score estimates should be investigated to ascertain the added value of these models.

It should be noted that there are viable alternatives to arrive at normative values and subsequently calculate *T* scores and percentile rank scores instead of the IRT-based methods or the approach advocated in the present paper. When scales are not unidimensional and/or the IRT model does not fit, alternatives can be used, for example, frequency-based percentile rank scores. Especially, regression-based norming (Mellenbergh, 2011) is an interesting option, for example, with the GAMLSS model

(Stasinopoulos et al., 2018), in which the shape of the frequency distribution of raw scores, as well as relevant "norm-predictors", such as gender, age, and educational level can be taken into account. A useful introduction to the continuous norming approach, which corrects for all levels of these demographic variables, is offered by Timmerman and colleagues (2021). However, for the present purposes, we build upon the previous work of the PROsetta stone initiative and the PROMIS group.

The strengths of the study are the use of large datasets with representative samples from the general population and patients, warranting trust in the findings. The data from the Dutch general population were collected by research institutes with a good reputation, and the representativeness of these population samples has been documented by Scherpenzeel (2018) and Timman and colleagues (2017). A relatively simple and straightforward approach is outlined, which leads to calculated *T* scores that approximate θ-based *T* scores well. A potential limitation of the proposed method is that it relies heavily on data from the general population, as this is the basis of the item parameters and θ's used to obtain *T* scores. The clinical measurement instruments were not developed for the population at large but rather for patients suffering from psychological complaints or symptoms of psychiatric disorders. Cronbach (1984) noted that, ideally, instruments should be validated with data from the population for which the instrument was intended. Indeed, the frequency distribution of responses of non-clinical respondents deviated much more from normality than was the case with data obtained in the clinical samples, as the present findings show.

The present results were based on normative data from the Netherlands. Application of the conversion formulas listed in Table 2 or the crosswalk Tables 5–7 is limited to this context, as general population respondents from other countries may score differently on these self-report measures. A further limitation may be the composition of the clinical sample used to evaluate the frequency distribution of the conversion formula in clinical data. All these patients suffered from mild to moderate common mental disorders, and patients with severe mental illness were not included. Future research should investigate other clinical samples.

Conclusion

The high correlations found in this study between θ-based and calculated *T* scores for assessment with the 4DSQ, the BSI, and the OQ-45, suggest that the proposed approach of using conversion functions provides a good approximation toward a common normalized metric for MHC. The use of such a common metric will make the interpretation of test results easier for therapists and patients (de Beurs, Fried, et al., 2022) and will allow for better involvement of the patient in shared decision-making regarding the treatment (Patel et al., 2008), and will stimulate the future uptake of measurement-based mental health care (Kilbourne et al., 2018).

References

Bakker, I. M., Terluin, B., Van Marwijk, H. W., van der Windt, D. A. M., Rijmen, F., van Mechelen, W., & Stalman, W. A. (2007). A cluster-randomised trial evaluating an intervention for patients with stress-related mental disorders and sick leave in primary care. *PLoS Clinical Trials, 2*(6), Article e26. https://doi.org/10.1371/journal.pctr.0020026

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*(8), 1316–1324. https://doi.org/10.1017/S0033291717002835

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 84767*, 307–310. https://doi.org/10.1016/S140-6736(86)90837-8

Bowman, M. L. (2002). The perfidy of percentiles. *Archives of Clinical Neuropsychology, 17*(3), 295–303. https://doi.org/10.1016/S0887-6177(01)00116-0

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods & Research, 21*(1), 89–115. https://doi.org/10.1177/0049124192021001004

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

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*(2), 513–527. https://doi.org/10.1037/a0035768

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*(10), 2305–2318. https://doi.org/10.1007/s11136-014-0790-9

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*(2), 193–204. https://doi.org/10.1080/13854040801968450

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement, 41*(6), 439–455. https://doi.org/10.1177/0146621617695522

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

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*(4), Article e1797. https://doi.org/10.1002/mpr.1797

de Beurs, E., den Hollander-Gijsman, M., Buwalda, V., Trijsburg, W., & Zitman, F. G. (2005). De Outcome Questionnaire (OQ-45): Een meetinstrument voor meer dan alleen psychische klachten [The Outcome Questionnaire (OQ-45): A measure for psychiatric symptoms and more]. *De Psycholoog, 40*(1), 53–63.

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), 1–12. https://doi.org/10.1002/cpp.696

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

de Beurs, E., Oudejans, S., & Terluin, B. (2022). *A common measurement scale for self-report instruments in mental health care: T* scores with a normal distribution (supplementary materials). https://www.psycharchives.org/en/item/86e598e9–4828-4127–86ae-5f0d18e9586a

de Beurs, E., & Zitman, F. G. (2006). De Brief Symptom Inventory (BSI): De betrouwbaarheid en validiteit van een handzaam alternatief voor de SCL-90 [The Brief Symptom Inventory: Reliability and validity of a handy alternative for the SCL-90]. *Maandblad Geestelijke Volksgezondheid, 61*, 120–141.

de Jong, K., Nugter, M. A., Polak, M. G., Wagenborg, J. E. A., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology & Psychotherapy, 14*(4), 288–301. https://doi.org/10.1002/cpp.529

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

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. Springer.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Erlbaum.

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*(1), Article 142. https://doi.org/10.1186/s12874-016-0241-0

Fischer, H. F., Tritt, K., Klapp, B. F., & Fliege, H. (2011). How to compare scores from different depression scales: Equating the Patient Health Questionnaire (PHQ) and the ICD-10-Symptom Rating (ISR) using item response theory. *International Journal of Methods in Psychiatric Research, 20*(4), 203–214. https://doi.org/10.1002/mpr.350

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*(6), 1615–1626. https://doi.org/10.1007/s11136-019-02147-3

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*(8), 1136–1143. https://doi.org/10.4088/JCP.10r06282whi

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2006). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 169–203). https://doi.org/10.1016/S0169-7161(06)26006-1

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*(1), 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.

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). Erlbaum. 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*(4), Article e12268. https://doi.org/10.1111/cpsp.12268

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*(4), 453–461. https://doi.org/10.1177/014662168400800409

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

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis and application of psychological and educational tests*. Eleven International. https://books.google.es/books?id=jRJJYAAACAAJ

Miller, S. D., Hubble, M. A., Chow, D., & Seidel, J. (2015). Beyond measures and monitoring: Realizing the potential of feedback-informed treatment. *Psychotherapy, 52*(4), 449–457. https://doi.org/10.1037/pst0000031

Patel, S. R., Bakken, S., & Ruland, C. (2008). Recent advances in shared decision making for mental health. *Current Opinion in Psychiatry, 21*(6), 606–6012. https://doi.org/10.1097/YCO.0b013e32830eb6b4

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

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*(1), 88–96. https://doi.org/10.1016/j.janxdis.2013.11.006

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*(10), 1517–1523. https://doi.org/10.1007/s11606-015-3360-0

Scherpenzeel, A. C. (2018). "True" longitudinal and probability-based Internet panels: Evidence from the Netherlands. In M. Das, P. Ester, & L. Kaszmirek (Eds.), *Social and behavioral research and the Internet* (pp. 77–104). Routledge. https://doi.org/10.4324/9780203844922-4

Scherpenzeel, A. C., & Bethlehem, J. G. (2011). How representative are online panels? Problems of coverage and selection and possible solutions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 105–132). Taylor & Francis.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Slade, M., Amering, M., & Oades, L. (2008). Recovery: An international perspective. *Epidemiology and Psychiatric Sciences, 17*(2), 128–137. https://doi.org/10.1017/S1121189X00002827

Smits, N. (2016). On the effect of adding clinical samples to validation studies of patient-reported outcome item banks: A simulation study. *Quality of Life Research, 25*(7), 1635–1644. https://doi.org/10.1007/s11136-015-1199-9

Smits, N., Öğreden, O., Garnier-Villarreal, M., Terwee, C. B., & Chalmers, R. P. (2020). A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Statistical Methods in Medical Research, 29*(4), 1030–1048. https://doi.org/10.1177/0962280220907625

Stasinopoulos, M. D., Rigby, R. A., & Bastiani, F. D. (2018). GAMLSS: A distributional regression approach. *Statistical Modelling, 18*(3–4), 248–273. https://doi.org/10.1177/1471082X18759144

ten Klooster, P. M., Oude Voshaar, M. A. H., Gandek, B., Rose, M., Bjorner, J. B., Taal, E., Glas, C. A. W., van Riel, P. L. C. M., & van de Laar, M. A. F. J. (2013). Development and evaluation of a crosswalk between the SF-36 Physical Functioning Scale and Health Assessment Questionnaire Disability Index in rheumatoid arthritis. *Health and Quality of Life Outcomes, 11*, 199–199. https://doi.org/10.1186/1477-7525-11-199

Terluin, B., Smits, N., Brouwers, E. P. M., & de Vet, H. C. W. (2016). The Four-Dimensional Symptom Questionnaire (4DSQ) in the general population: Scale structure, reliability, measurement invariance and normative data: A cross-sectional survey. *Health and Quality of Life Outcomes, 14*(1), Article 130. https://doi.org/10.1186/s12955-016-0533-4

Terluin, B., van Marwijk, H. W., Adèr, H. J., de Vet, H. C., Penninx, B. W., Hermens, M. L., van Boeijen, C. A., van Balkom, A. J., van der Klink, J. J., & Stalman, W. A. (2006). The Four-Dimensional Symptom Questionnaire (4DSQ): A validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry, 6*(1), Article 1. https://doi.org/10.1186/1471-244X-6-34

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*(1), 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

van der Laan, J. (2009). *Representativity of the LISS panel*. Statistics Netherlands.

Van Hoeck, K. J. M., Lilien, M. R., Brinkman, D. C., & Schroeder, C. H. (2000). Comparing a urea kinetic monitor with Daugirdas formula and dietary records in children. *Pediatric Nephrology, 14*(4), 280–283. https://doi.org/10.1007/s004670050759

van Stralen, K. J., Dekker, F. W., Zoccali, C., & Jager, K. J. (2012). Measuring agreement, more complicated than it seems. *Nephron Clinical Practice, 120*(3), c162–c167. https://doi.org/10.1159/000337798

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*(1), 73–86. https://doi.org/10.1016/j.jclinepi.2013.04.019

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT Modeling in the presence of zero-Inflation with application to psychiatric disorder severity. *Applied Psychological Measurement, 39*(8), 583–597. https://doi.org/10.1177/0146621615588184

History
Received April 19, 2021
Revision received July 6, 2022
Accepted August 3, 2022
Published online December 16, 2022
EJPA Section / Category Clinical Psychology

Acknowledgments
In this paper, we gratefully made use of BSI- and 4DSQ data of the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands) and OQ-45 data from TNS-NIPO, The Netherlands. We would also like to thank MHC providers in the Netherlands for providing patient data on the BSI and the OQ-45 and the VU Medical Center for providing patient data on the 4DSQ.

Conflict of Interest
The authors report no conflict of interest.

Publication Ethics
The data and narrative interpretations of the data/research appearing in the manuscript have not been presented at a conference or meeting, posted on a listserv, shared on a website, including academic social networks like ResearchGate, and so forth.

Open Science
We report on a reanalysis of data about which has been published before. We report and refer to previous publications on how we determined our sample size, all data exclusions, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact *p* values, effect sizes, and 95% confidence or credible intervals.
Open Data: The information needed to reproduce all of the reported results are not openly accessible, but can be requested from the first author.
Open Materials: The information needed to reproduce all of the reported methodology is made available. We provided sufficient information for an independent researcher to reproduce the reported results, including a codebook (https://www.psycharchives.org/en/item/86e598e9-4828-4127-86ae-5f0d18e9586a; de Beurs, Oudejans, et al., 2022). Moreover, we have uploaded an annotated version of our R-code with two practice datasets, which will allow other researchers to apply it to these data (and their own data).
Preregistration of Studies and Analysis Plans: This study was not preregistered.

Edwin de Beurs
Department of Clinical Psychology
Faculty of Social Sciences
Leiden University
Wassenaarseweg 52
2333 AK Leiden
The Netherlands
e.de.beurs@arkin.nl

