2024 ■ Vol. 20 ■ no. 3

Test MultiDimensionnel de Personnalité (TMDP): Norms in T-scores and Percentile Rank order scores for the general and clinical population

Edwin de Beurs [a] and Jacques Cloutier [b]

[a]Leiden University, Netherlands & Arkin Mental Health Institute, Amsterdam
[b]Private consultant, Montréal

Abstract ■ Converting raw scores to T-scores simplifies the interpretation of test results by providing a standardized metric. A simple linear conversion of raw scores to T-scores can produce misleading T-scores if the raw score distribution deviates significantly from normality. The aim of the present research was to provide normalized T-scores and percentile rank scores for the Test MultiDimensionnel de Personnalité (TMDP). T-scores were established by first transforming raw scale scores to a normal distribution with either an approach based on the Item Response Theory or with an approach based on Rankit percentile-based normalization. The three resulting T-scores (Linear-, IRT- and Rankit-based T-scores) were compared. Results revealed that both normalization approaches yielded similar normalized T-scores, which deviated substantially from linear T-scores for most scales, especially in the higher range of the scores. Benefits of normalized T-scores and the need for gender- and age-based norms are discussed.

Keywords ■ Test MultiDimensionnel de Personnalite, raw scores, T-scores. **Tools** ■ R.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

cloutierja@gmail.com

10.20982/tqmp.20.3.p205

Introduction

The TMDP (Test MultiDimensionnel de Personnalité) was developed some 30 years ago and was inspired by the MMPI. The MMPI-3 and TMDP have in common that both use a true-false response format to measure psychopathology and psychiatric dimensions. When analyzing principal components in the instruments, both questionnaires appear to have a similar structure with three main dimensions: anxiety-depression, psychotic ideation, and antisocial behavioral tendencies. The original 410 items of the TMDP were written in universal French by a group of clinical psychologists in the Montréal area, Canada (Cloutier, 1997). An English version of the items was also developed. Initially, scales were created using a 2 by 3 chi-square test on item-scores, comparing the two response options (true-false) among three subgroups of respondents: a general population-based group, a clinical group with the focus pathology and all other clinical respondents. This chi-square approach was an improvement over merely comparing the clinical group with the general population sample. Items chosen more frequently or less frequently by the focus pathological group would be considered and retained for a specific scale. When this chi-square procedure was applied to the TMDP, it created some overlap among scales in item content. For instance, hospitalized psychotic patients tended to endorse items measuring depression. Consequently, items concerning depression symptoms and allocated to the depression scale were included in other clinical scales as well. The chi square technique was used by many researchers in order to identify items from the MMPI to construct a specific scale; for example, close-head trauma by Gass and Wald (1997).

After reanalyzing MMPI data, Butcher et al. (1989) proposed to revise the MMPI and keep the clinical scales as they were modified over the years and corrected with the K scale. As an example, the Sc scale (for Schizophrenia) had 78 items and many items were also included in other scales. Tellegen et al. (2003) were critical of the overlap of items among the MMPI-2 scales. Later on, the clinical scales of the

TMDP were revised in order to allocate items to scales with no item overlap among scales. For instance, all 'depressive items' were removed from clinical scales other than the Depression scale.

In line with the procedure suggested by Tellegen and Ben-Porath (2003), clinical scales without overlap were revised for the TMDP as well (Cloutier, 2024, paper submitted). This reduced the number of items per clinical scale to 15-28 items for the TMDP. A Principal Component Analysis was applied to the TMDP, utilizing a tetrachoric correlation matrix for dichotomous items, resulting in ten revised clinical scales. Items with substantial cross-loading were omitted from the final allocation of items to scales, yielding scales with items predominantly loading on a single component. This resulted in scales for Pessimism (15 items Cronbach $\alpha = .91$), Anxiety (17 items, $\alpha = .89$), Depression (24 items, $\alpha = .91$), Social isolation (22 items, $\alpha = .87$), Perceived parental care (12 items, $\alpha = .86$), Psychoticism (20 items, $\alpha = .83$), Paranoid ideation (16 items, $\alpha = .72$), Antisocial conduct (17 items, $\alpha = .76$), Impulsivity (18 items, $\alpha = .84$), and Hypomania (20 items, $\alpha = .78$). The questionnaire also has validity scales and complementary scales (Cloutier, 1997) and personality disorder scales based on the DSM-5 (Cloutier & Renaud, 2012).

Another similarity with the MMPI is that a multidimensional profile of scores on the scales can be made. To make scores among scales compatible, raw score are converted into a common metric: T-scores (McCall, 1922). Initially, T-scores were derived for the clinical scales through linear conversion of Z-scores:

$$T = 50 + 10 \times \big((\text{Raw Score} - M_{\text{Raw Score}})/SD_{\text{Raw Score}}\big).$$

A potential problem arises with standardization of scores (to Z-scores and eventually to T-scores) when raw test scores have a non-normal distribution. In the general population raw scores on clinical scales, such as the MMPI or the TMDP, are usually not normally distributed, but skewed to the right, due to an overrepresentation of low scores. With a skewed frequency distribution, the mean is no longer an adequate statistic to represent the midpoint of the frequency distribution. Furthermore, with a peaked distribution the variation in scores is smaller than expected under a normal distribution. Thus, under non-normal conditions, a simple linear transformation of raw scale scores into standardized scores will yield aberrant Z- or T-scores, and raw scores need to be normalized first. In the context of the MMPI, Colligan et al. (1984) proposed a normalizing transformation of MMPI raw scale scores. With the introduction of the MMPI-2 raw scale scores were transformed into "uniform" T-scores accounting for non-normality of the raw scores using a procedure described by Tellegen and Ben-Porath (1992). They proposed to convert MMPI raw scores towards "uniform" T-scores with a polynomial equation found by regressing raw scores onto percentile based T-scores with a normal distribution.

One of the aims of the present research is to provide T-scores for the TMDP following a similar approach, but this time in the form of *normalized* T-scores (de Beurs, Oudejans, & Terluin, 2022). Previously, we have proposed an alternative method to establish T-scores, by first transforming raw scale scores to a normal distribution with either (1) an approach based on the Item Response Theory (Embretson & Reise, 2013) or with (2) an approach based on Rankit percentile-based normalization (Soloman & Sawilowsky, 2009). With the IRT-based approach, first IRT factor scores (thetas) are established. To obtain for each raw score an IRTbased T-score, raw scores can be regressed onto the IRT factor scores with nonlinear regression. The fitted values of the regression model are used as TIRT.. The other method of normalizing raw scores is the Rankit approach. With Rankit normalization, percentile rank order scores are established for raw scores and these are converted with the `probit` function (the inverse cumulative distribution function) into corresponding Z-scores (Bliss, 1967; Soloman & Sawilowsky, 2009). Subsequently, these scores can be transformed to normalized T-scores (T_{Rankit}) (de Beurs, Oudejans, & Terluin, 2022). IRT- and Rankit-based methods have been compared and appear to yield quite similar results (Schalet et al., 2021). In this research, we compared normalized T-scores with linear T-scores to illustrate the benefit of normalization. Furthermore, we compared T_{IRT} and T_{Rankit} again to explore which approach yields the best results for the TMDP.

An alternative approach to provide meaning to raw test results is to establish information on the relative position of the raw score in the frequency distribution of the reference group. This can be done by calculating percentile rank order (PR) scores (Crawford & Garthwaite, 2009; Crawford et al., 2009). PR-scores are calculated as the cumulative proportion of scores in the reference group that are equal or lower that the raw score minus half of the proportion with exactly the raw score ($PR = CF - 0.5 \times F/N$; in which CF is the cumulative frequency, F is the frequency of the raw score and N is the sample size of the reference group; see also Crawford & Garthwaite, 2009; Kurtz & Mayo, 1979). They allow for an interpretation, such as: "65% of respondents to this questionnaire have a similar or lower score than you". Percentile scores are intuitive and more easily understood compared to T-scores by clients (and professionals). However, a drawback of percentile scores is that they are ordinal, whereas T-scores are equidistant and have an interval scale, allowing for basic arithmetic operations, such as subtraction of scores to establish (therapeutic) gain. Percentile scores tend to be more spread out

around the mean and bunched together at the ends of the scale, compared to T-scores, which may result in a misinterpretation of their meaning (Bowman, 2002). Consequently, we provide for each possible raw scale score both normalized T-scores and PR-scores. Finally, two PR-scores were established: one based on data from the general population and the other from a clinical population. All in all this allows for interpretation of test results, such as: "Your score on the Pessimism scale is T=65.2; 93% of the general population has a lower score, but compared to other patients this score is less uncommon as only 49% of our patients have a lower score"

Finally, we investigated the need for separate norms for genders and age groups. The authors of the MMPI-3 opted for non-gendered norms, after finding that differences in T scores between men and women were quite small, never exceeding 3 T-score points. However, in the normative group used for the TMDP, it was noticed that women have a different response to items regarding antisocial behavior and perceived parental care than men. Furthermore, it is well-established that the genders respond differently in self-report instruments to items regarding depression and anxiety, both in the general population as well as in clinical samples (Cavanagh et al., 2016; McLean & Anderson, 2009; Pesce et al., 2016; Sigmon et al., 2005). If gender or age differences are found, more precision in scores can be achieved with specific norms for the genders and age groups.

Methods

This study's design and its analysis were not preregistered. The deidentified data on which study conclusions are based and the materials described in the method section are available from the first author upon request. Some of the analytic code from R needed to reproduce analyses is available at https://www.psycharchives.org/en/item/a4a7c591-af33-4463-b828-6b41b9ea0f7d. R packages used and their references are described in the statistical analysis section.

Samples

Data were obtained from 2156 respondents comprising a normative sample from the general population (n=356) and a clinical sample ($n = 1800$). Mean age for the entire group of respondents was $M = 35.19$ ($SD = 12.66$; range=16-79). The population sample contained 155 men (43.5%) and 201 women (56.5%); the clinical sample contained 780 men (43.3%) and 1020 women (56.7%). Thus, compared to the general population women were somewhat overrepresented in both samples. To investigate the effect of gender on TMDP scale scores, we compared the scores of respondents who identified themselves as men (n=935) and women ($n = 1221$), combining both samples. The effect of age was established by dividing the total sample in four age groups of roughly equal size: 16-26 years, $n = 557$; 27-26 years, $n = 620$; 37-44 years, $n = 498$; 45-79 years; $n = 481$.

Procedure

Data from the general population were collected by psychologists in the Montreal area. An inclusion criterion was language since items were written in French. For the second criterion, participants should have never been hospitalized in psychiatry and be over 18-year-old. Also excluded were respondents working in psychiatry, but other staff members of a general hospital in Saint-Jean-sur-Richelieu, Québec, were solicited as respondents. An effort was made to recruit a representative sample regarding educational level: Salesmen in the automotive branch, workers in a bakery, workers in a kindergarten, members of a karate school and students in engineering were solicited as well. All were asked to complete the questionnaire truthfully and honestly about themselves. They could put their names on the questionnaire or complete it anonymously. Respondents aged 16 and 17 were added to the normative adult group, proportionally to the size of this age group according to the last census for the Canadian population.

For the clinical sample data were collected from 1800 psychiatric patients seeking treatment at various hospitals in the Montreal area. Psychologists working in private practice who used the scoring software for the TMDP, also contributed data to the clinical sample. Most clinical respondents were outpatients, but around 15% completed the questionnaire during their hospitalization. Again, the ability to read and understand French was an inclusion criterion. A very small proportion of English-speaking patients (< 5%) completed the English version of the TMDP). The most common pathologies were psychotic episodes, anxiety disorders, major depression, and borderline personality disorders. For the present analysis, only age and gender were kept in the dataset to preserve anonymity of the respondents.

Measure

A description of the TMDP is provided in the introduction. The self-report instrument is intended to assess the nature and severity of psychopathology. It comprises 410 items with a (true=1/false=0) response format. The number of endorsed items is counted (after reversing the scores for some items) and, generally, a higher scale score implies more of the measured construct. For example, a higher score on Pessimism means more pessimism. However, a higher score on Perceived parental care means a more negative view on care received from one's parents.

Table 1 ■ Basic Psychometrics of the Raw Scale Scores and T-scores of the General Population Sample ($n = 356$).

	Raw scores				T-scores			
scale	M	SD	skew.	kurt.	M	SD	skew.	kurt.
Pessimism	2.83	3.32	1.50	1.66	49.97	8.68	0.72	-0.09
Anxiety	4.80	3.59	0.85	0.03	49.98	8.82	0.40	-0.43
Depression	6.14	5.03	1.06	0.54	49.97	9.06	0.45	-0.23
Social isolation	6.25	4.84	0.55	-0.60	49.97	9.07	0.04	-0.77
Perceived parental care	2.60	2.69	1.63	2.84	49.98	8.37	0.71	-0.18
Psychoticism	2.18	2.24	1.54	2.31	49.98	7.97	0.65	-0.68
Paranoid ideation	2.64	2.03	0.94	0.70	45.59	9.04	0.31	-0.32
Antisocial behavior	3.95	3.06	0.93	0.48	49.99	8.42	0.40	-0.34
Impulsivity	4.93	3.33	0.77	-0.16	49.98	8.91	0.36	-0.57
Hypomania	3.77	2.66	0.78	-0.05	49.98	8.26	0.37	-0.50

Note. skew.: skewness, kurt.: kurtosis; values > 1.00 or < -1.00 suggest important deviations from normality.

Statistical analyses

All statistical analyses were done with R. Psychometric properties (M, SD, Skewness, and Kurtosis) were established for the raw scale scores from the general population sample. Linear T-scores were compared to "normalized" T-scores based on either (1) the Rankit approach (Soloman & Sawilowsky, 2009) with the RankNorm function in the R `RNOmni` package version 1.0.1 (McCaw et al., 2019) or (2) by means of an IRT model established with the multipleGroup function of the R `mirt` package (Chalmers, 2012) for dichotomous response options (Bock & Lieberman, 1970) with the EAPsum score as estimator as proposed by Fischer and Rose (2019). PR scores were also established with the RankNorm function of the `RNOmni` package. Cross-walk tables are presented which provide T- and PR-scores for each possible raw score. Finally, formulas were established through non-linear modelling (the Nonlinear Least Squares nls function in the R `stats` (version 3.6.2) package and the `glsnls` package version 1.2.0 to convert raw scores to normalized T-scores and PR-scores. These formulas can be built into an Excel file or questionnaire scoring software. In addition, we investigated gender and age differences in TMDP scale scores and, where needed, we provided normative values separately for men and women and for age groups.

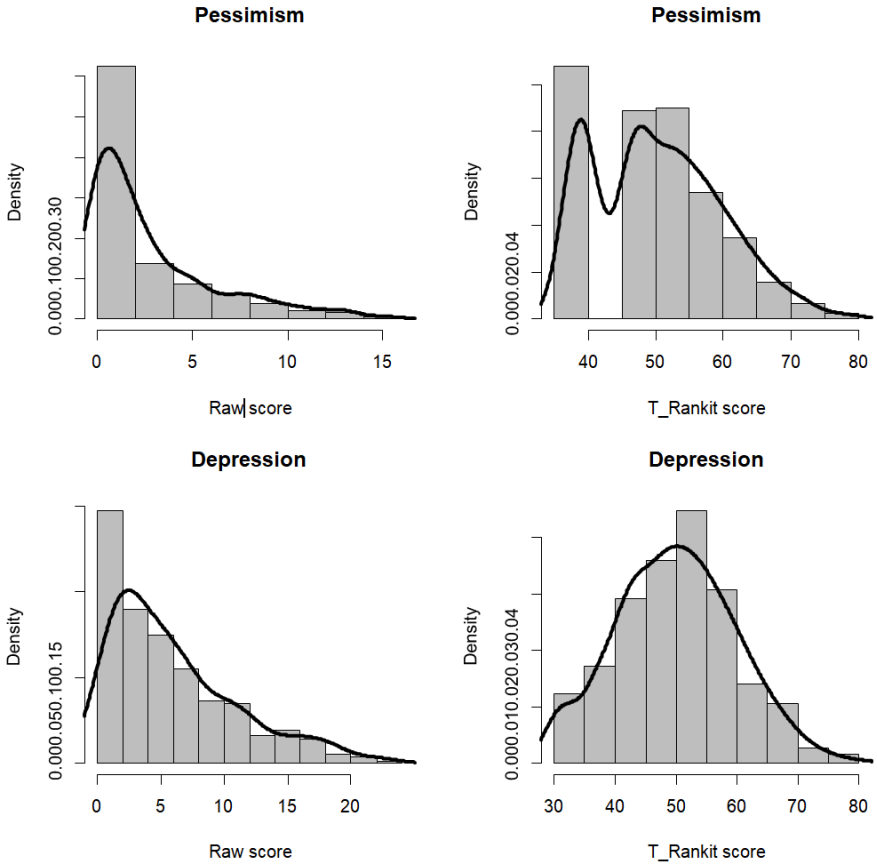
Results

Table 1 shows indicators of the frequency distributions of raw scale scores and T-scores (means, standard deviations, skewness, and kurtosis). All raw scales showed positive skewness (due to an overrepresentation of low scores), in particular the scales for Pessimism, Depression, Perceived parental care, and Psychoticism. Moreover, the frequency distributions of these scales were also peaked. Aften transformation to normalized T-scores some skewness remained, but excessive kurtosis was removed. Figure 1 illustrates this for the Pessimism and Depression scores.

Figure 2 shows the relation between raw scores and T-scores, established by various approaches. The dots represent IRT factor-based T-scores (T_{theta}). There is a range of T_{theta}'s per raw score as in IRT factor scores are established based on the chosen response options ánd the difficulty/severity of the items, resulting in more variant scores than merely summed scale scores. A straight line shows the relation between raw scores and T-scores after a linear transformation (T_{Linear}). Curved lines show the relation between raw scores and T-scores after a Rankit-based normalization (T_{Rankit}), and the fitted scores after regression of raw scores on IRT-based theta scores (T_{IRT}). T_{Rankit} was only available for the general population sample. In this sample raw Psychoticism (PSY) scores ranged from 0 to 11 and were missing for higher scores. Hence, the shorter length of the green line compared to the other lines in Figure 2, right hand side). The linear conversion was based on means and standard deviation of the general population sample and was, again for the Psychoticism scale, only present in the sample for the range in raw scores from 0 to 11. T_{Linear} for higher raw scores is based on extrapolation, which is indicated with a dashed continuation of the line for Psychoticism. In the supplementary materials, similar figures are presented for the other scales.

Figure 2 shows that the linear conversion resulted in too elevated T-scores in the higher range of the scale, especially with Psychoticism; T_{IRT} and T_{Rankit} followed the pattern of the dots (IRT factor scores) better. For ANX, DPR, ISO, and IMP, a linear calculation of T-scores would yield an acceptable approximation of IRT-based T-scores. However, for the other scales normalization of raw scores and a curvi-linear transformation of raw scores to T-scores yielded a much better interpretable test result. Consequently, as basis for the T-scores in crosswalk tables and

Figure 1 ■ The effect of Normalization for Pessimism and Depression: Histograms with Density Lines for Raw Scores and Normalized T-scores



formulas, we have chosen for the IRT-based approach, with the exception of the Paranoid Ideation scale. Due to the high variability in factor scores across raw scores for Paranoid ideation in the IRT model (see Figure 2), indicating a poor model fit, we opted with the Paranoid Ideation scale for the Rankit method for T-score conversion.

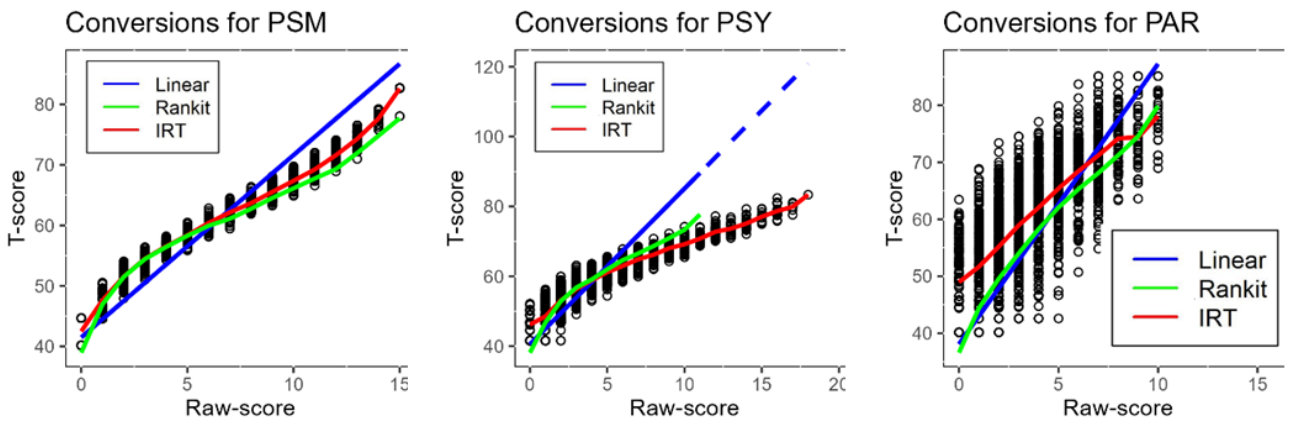
Table 4 presents T_{Linear} and TIRT scores (calculated with the CW-formulas presented in Supplementary materials, Appendix D) for all scales, allowing to establish the benefit of the normalization approach. Comparing columns in Table 4 shows that linear and normalized T-scores diverged substantially. A linear conversion to T-scores yields for all scales exaggerated low T-scores for scores below average and too high T-scores in the upper region of the scale. The difference in T-scores was generally five scale point, but with some scales the difference between T_{Linear} and T_{IRT} was much larger, such as for the Psychoticism and Antisocial behavior scales. Also, around the average score, differences were observed for Pessimism (4 points difference at T =50), perceived Parental care (3,5 points at T=50), Psychoticism (3,4 points at T=50), and Depression (2,5 points at T=50). For Social Isolation both T-scores were more similar (1.0 point at T=50). Normalized T scores were generally higher because in a right-skewed frequency distribution the mean underestimates the true center of the distribution.

Table 2 presents an example of a cross-walk table for raw scores to T-scores according to the normalized conversion for the Psychoticism scale. It also presents PR-scores for the general population and the clinical population. Generally, T-scores range from 20 to 80; PR scores range from 1 to 100. Cross-walk tables for the other scales can be found in the supplementary materials.

Gender and age group specific norming

We also investigated whether separate norms are in order for both genders and for age groups. Table 3 presents means (and SD) of raw scores for men and women. Ac-

Figure 2 ■ The Relation Between Raw Scores (x-axis) and T-scores (y-axis) for Pessimism (PSM), Psychoticism (PSY) and Paranoid Ideation (PAR), Based on Linear, Rankit and IRT Conversions with Dots Representing IRT Factor Scores (Thetas).



cording to t-tests, scores of men and women differed on most scales, justifying separate norms for the genders. Men scored higher on Social Isolation, Paranoid ideation, Antisocial behavior, and Hypomania; Women scored higher on Pessimism, Anxiety, Depression, and Perceived parental care. The effect size of most differences was small to medium, but the difference in Antisocial behavior was large. Obviously, if no differences are found between man and women, there is no need for separate norms either.

Based on these results, we decided to also establish T-scores for men and women separately. In the mid-range of the scale this results in a difference in T-score between men and women of 0.5 to 1.0 T-points (See supplementary materials Appendix C) with the exception of Antisocial behavior where the T-score for men was 13.2 points lower than for women at RS=11, reflecting that a high raw score is much more common for men than for women. For the other scales the difference might seem marginal, but it was still close to 5 points or half an SD, which is approximately a minimally detectable difference in score (Engel et al., 2018). Finally, we investigated the influence of age on TMDP scale scores with correlational analysis and by comparing four age groups of roughly equal size. Table 5 at the end present the correlations and means and standard deviations for the age groups. There were no differences for pessimism, depression, psychoticism, and paranoid ideation.

There were age effects for Anxiety [$F(3, 2152) = 6.91$; $p = 0.0001$; $\eta^2 = 0.010$; the oldest group has lower scores], Perceived parental care [$F(3, 1172.66) = 22.69$; $p = 0.000$; $\eta^2 = 0.028$; the youngest group has lower scores compared to the other age groups], Antisocial behavior [$F(3, 1187.74) = 37.49$; $p < 0.001$; $\eta^2 = 0.048$; the youngest two groups have higher scores], impulsivity [$F(3, 1181.51) = 20.10$; $p < 0.001$; $\eta^2 = 0.026$], and hypomania [$F(3, 1184.37) = 14.33$; $p = 0.000$; $\eta^2 = 0.018$]. For both the oldest group has lower scores. Thus, where age effects were found, scored tended to decline with increasing age as was also shown by the correlations between age and scale scores. All in all, these results justify only separate norms for some scales and some age groups. We determined separate norms for the oldest group vs. the three younger groups for Anxiety, Impulsivity and Hypomania. Furthermore, separate norms were established for Antisocial behavior for age groups 14-26, 27-36 and 37-79. Finally, for Perceived parental care distinct norms were determined for the youngest age group 14-26 and the other respondents, age 27-79.

Discussion

Our study yielded several key findings regarding the transformation of raw scores on the TMDP scales into interpretable T-scores. First, we observed significant positive skew and kurtosis in the raw scores, particularly for the scales Perceived parental care, Psychoticism, and Pessimism. This characteristic leads to inflated T-scores in the higher range of the scale when a basic linear conversion is employed. To address this issue, we first normalized raw scores. This normalization effectively counteracted excessively high T-scores. We then established curvilinear conversion formulas based on the normalized scores. These formulas generate T-scores that are well-suited for interpretation and closely approximate factor scores derived from Item Response Theory (IRT) analysis. Finally, the combined use of T-scores and PR-scores enhances the clarity of test result interpretation.

Table 2 ■ Pessimism: Crass-Walk Table from Raw Scores to T-scores and PR-scores for the General Population and the Clinical Population

Raw Score	T_{IRT}[1]	PR_n[2]	PR_cl[3]
0	42.8	13	3
1	47.3	38	8
2	51.1	56	12
3	54.2	67	17
4	56.8	74	22
5	58.9	80	26
6	60.7	84	30
7	62.3	87	35
8	63.8	90	42
9	65.4	93	49
10	67.1	95	56
11	69.1	96	64
12	71.5	97	72
13	74.4	99	81
14	78.0*	–	89
15	82.4*	–	97

Note. T_{IRT}: IRT-based T-score; PR_n: Percentile Rank score general population; PR_cl: Percentile Rank score clinical sample.
*: Scores based on the formula, not actually present in the general population sample.
[1]: Formula for normal all (ID:3) for RS->T(IRT): $y = 42.771 + 5.007 \times x - 0.457 \times x^2 + 0.020 \times x^3$; a degree-3 polynom.
[2]: Formula for normal all (ID:5) for RS->PR_n: $y = 13.338 + 30.365 \times x - 5.725 \times x^2 + 0.606 \times x^3 + 0.032 \times x^4 + 0.001 \times x^5$; a degree-5 polynom.
[3]: Formula for clinical all (ID:3) for RS->PR_cl: $y = 3.845 + 3.676 \times x + 0.126 \times x^2 + 0.003 \times x^3$; a degree-3 polynom.

The benefits of normalizing test results before T-score conversion is clearly illustrated with Figures 1 and 2 and the results presented in Table 4 at the end. Linear T-scores appear to be inflated, particularly at the higher end of the scales. This inflation can be substantial, reaching or exceeding 5 T-score points for many scales, which is not trivial and equivalent to half a standard deviation.

We compared also two approaches to normalization: Rankit transformation and an IRT-based scoring of the TMDP. IRT scoring is generally recommended over simple summed scale scores, provided that certain requirement (e.g., unidimensionality of the scales, monotonicity of the scores, and local independence of item pairs) are met. Moreover, in contrast to Rankit, IRT-based transformations used data from both groups. Moreover, Normalization or raw scores through an IRT model or by means of percentile based Rankit transformation yielded in practice very similar results.

Norms for genders or non-gendered norms

Raw scale scores of men and women differed on most clinical scales of the TMDP, justifying separate norms for the genders. Men scored higher on Social Isolation, Paranoid ideation, Antisocial behavior, and Hypomania; women scored higher on Pessimism, Anxiety, Depression, and Perceived Parental Care. The effect size of most differences was small to medium, but the difference in Antisocial behavior was quite large.

T-scores can be adjusted to account for gender differences, providing more accurate information for men and women. However, this adjustment comes with a trade-off: such Tscores can no longer be used for directly comparing scores across genders. This is because the conversion process ensures that the average T-score for each subgroup (men, women, or specific age groups) will be 50 by definition. Thus, it depends the research question whether non-gendered or gender specific scores are preferred. But there is more: The first argument proposed for nongendered norms for the MMPI-3 is a legal one and concerns personal screening in which it could be illegal to discriminate by race, colour, religion, sex, or national origin in employment practices. There is also a practical consideration: according to the MMPI-3 manual (Ben-Porath & Tellegen, 2020), there was a remarkable absence of substantive differences between gendered and non-gendered norms. Nearly all the mean differences were below 3 T-score points. However, with the TMDP significant differences appeared between male and female respondents on

Table 3 ■ Mean (and SD) and T-test Statistics for the Difference in Score Between Genders.

	Men ($n = 935$)		Women ($n = 1221$)		Significance			
Scale	M	SD	M	SD	$t(2155)$	p	Cohen's d	95% CI
Pessimism	7.13	5.02	7.84	4.67	−3.34	.0008	−0.15	[-0.23, -0.06]
Anxiety	8.09	4.70	9.48	4.55	−6.89	.0000	−0.30	[-0.39, -0.22]
Depression	11.76	6.80	13.14	6.37	−4.79	.0000	−0.21	[-0.30, -0.12]
Social isolation	9.58	5.37	8.62	5.26	4.13	.0000	0.18	[0.09, 0.27]
Perceived parental care	4.37	3.50	5.87	3.89	−9.40	.0000	−0.40	[-0.49, -0.32]
Psychoticism	4.15	3.78	3.96	3.59	1.21	.2266	0.05	[-0.03, 0.14]
Paranoid ideation	5.66	4.11	5.43	3.81	1.37	.1715	0.06	[-0.03, 0.15]
Antisocial behavior	5.56	3.41	3.47	2.82	15.21	.0000	0.68	[0.59, 0.77]
Impulsivity	7.51	4.72	7.58	4.66	−0.31	.7571	−0.01	[-0.10, 0.07]
Hypomania	6.76	3.89	6.57	3.60	1.16	.2463	0.05	[-0.03, 0.14]

Note. CI: Confidence interval.

certain clinical scales and comparison of non-gendered and gendered T- and PR- scores reveals substantial differences. Finally, nongendered norms could be used to accommodate transgender and non-binary individuals whose gender identity doesn't align with the traditional male/female binary. This is a practical solution and fully conceivable for interpretation of scores on the TMDP.

Norms according to age

The results also point out some differences between young and older respondents. The TMDP scoring software used to select adults as 18 and over and had separate norms for teens aged 14 to 17. In an earlier version of the TMDP, separate norms were established for adults (18 and over) and teenagers from 14 to17. This was recently abandoned and the TMDP is now dedicated to adults from 16 and up with adult specific norms. For the scales Pessimism, Depression, Social isolation, Psychoticism, and Paranoid Ideation, there are no differences between age groups. Separate norms may be used for Anxiety for younger respondents (1644) vs. the oldest group (45-79), Antisocial behavior for 16-26, 27-36, and 37-79, Perceived parental care for the youngest group (16-26) vs. the others (27-79), and for Impulsivity and Hypomania for 16-44 compared to the oldest age group (45-79). It possible to adapt scoring of these clinical scales according to different groups of ages and in the supplementary materials cross-walk tables and formulas are provided to do so.

T- and PR-scores to stimulate measurement-based care

Using tests, such as the TMDP, to evaluate the mental health of our clients at the start of treatment and to monitor progress over time is an essential element of evidence-based care (Lewis et al., 2019). Consequently, monitoring outcomes is included in treatment guidelines (Prevolnik Rupel et al., 2021). Measurement is also recommended by researchers warning for the unreliability of unstructured clinical judgment (Meehl, 1954; Kahneman et al., 2021). Despite the established benefits (Delgadillo et al., 2017; Lambert & Harmon, 2018), standardized assessment remains underutilized (Boswell et al., 2013; Jensen-Doss et al., 2018). One of the reasons may be difficulty with the interpretation of test results. We hope that the plea to use of standard measurement scales, such as T- and PR-scores (de Beurs, Boehnke, & Fried, 2022) will stimulate standardized assessment in the therapy office.

Strengths and limitations

A strength of this research is sufficiently sizable samples to fit an IRT model and to establish norms for separate norm groups. There was some overrepresentation of women in the population and clinical samples. In combination with the gender effect found for most scales, this may lead to slightly biased T- and PR-scores when established for the entire group, irrespective of gender. This can be overcome by giving men and women different weights in the data analysis. However, as the effect size of gender differences was generally small and the overrepresentation of women was also small, we have decided against weighting the data differently for the genders. Some unresearched issues remain, such as, investigation of other background or demographic variables that may affect the test score, e.g., socioeconomic status or highest attained education level. There may be limitations of the sample in regard to diversity factors and we note that study findings may not generalize to the broader population.

Conclusion

By utilizing common metrics, such as T-scores and PR-scores, mental health professionals can create a more collaborative understanding of test results with their clients. This approach bridges the gap between the psychometrics

expertise of the professional and the everyday experience of the client. It facilitates discussions about the implications of the test results for tailoring therapeutic interventions and determining treatment completion when therapeutic goals have been achieved.

Authors' note

R code for analyses made available at www.psycharchives.org/en/item/86e598e9-4828-4127-86ae-5f0d18e9586a.

References

Ben-Porath, Y. S., & Tellegen, A. (2020). *Mmpi-3. manual for administration, scoring, and interpretation*. University of Minnesota Press.

Bliss, C. I. (1967). *Statistical tests of skewness and kurtosis statistical tests of skewness and kurtosis* (Vol. 1). McGraw-Hill.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197. doi: 10.1007/BF02291262.

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2013). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research*, *25*(1), 6–19. doi: 10 . 1080 / 10503307.2013.817696.

Bowman, M. L. (2002). The perfidy of percentiles. *Archives of Clinical Neuropsychology*, *17*(3), 295303. doi: 10 . 1016/S0887-6177(01)00116-0.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota multiphasic personality inventory-2 (mmpi-2): Manual for administration and scoring*. University of Minnesota Press.

Cavanagh, A., Caputi, P., Wilson, C. J., & Kavanagh, D. J. (2016). Gender differences in self-reported depression and co-occurring anxiety and stress in a vulnerable community population. *Australian Psychologist*, *51*(6), 411–421. doi: 10.1111/ap.12184.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of statistical software*, *48*(6), 1–29. doi: 10 . 18637 / jss . v048.i06.

Cloutier, J. (1997). Test multidimensionnel de personnalité (tmdp): Validité des échelles. *Revue québécoise de psychologie*, *22*, 1–99.

Cloutier, J., & Renaud, S. (2012). L'évaluation des troubles de la personnalité au mmpi-2 et au tmdp. *Revue québécoise de psychologie*, *32*, 1–99.

Colligan, R. C., Osborne, D., & Offord, K. P. (1984). Normalized transformations and the interpretation of mmpi t scores: A reply to hsu. *Journal of Consulting and Clinical Psychology*, *52*, 824–826. doi: 10.1037/0022-006X.52.5.824.

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. doi: 10.1080/13854040801968450.

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist*, *23*(7), 1173–1195. doi: 10.1080/13854040902795018.

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

de Beurs, E., Oudejans, S., & Terluin, B. (2022). A common measurement scale for scores from selfreport instruments in mental health care: T scores with a normal distribution. *European Journal of Psychological Assessment*, *38*, 1–99. doi: 10.1027/1015-5759/a000740.

Delgadillo, J., Overend, K., Lucock, M., Groom, M., Kirby, N., McMillan, D., Gilbody, S., Lutz, W., Rubel, J. A., & de Jong, K. (2017). Improving the efficiency of psychological treatment using outcome feedback technology. *Behaviour Research and Therapy*, *45*(2), 170–184. doi: 10.1016/j.brat.2017.09.011.

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

Engel, L., Beaton, D. E., & Touma, Z. (2018). Minimal clinically important difference: A review of outcome measure score interpretation. *Rheumatic Disease Clinics*, *44*(2), 177–188. doi: 10.1016/j.rdc.2018.01.011.

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*(1), 85–99. doi: 10.1080/00273171.2018.1491381.

Gass, C. S., & Wald, H. S. (1997). Mmpi-2 interpretation and closed-head trauma: Cross-validation of a correction factor. *Archives of Clinical Neuropsychology*, *12*(3), 199–205. doi: 10.1037/0022-006X.52.5.824.

Jensen-Doss, A., Haimes, E. M. B., Smith, A. M., Lyon, A. R., Lewis, C. C., Stanick, C. F., & Hawley, K. M. (2018). Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. *Administration and Policy in Mental Health and Mental Health Services Research*, *45*(1), 48–61. doi: 10.1007/s10488-016-0763-0.

Kahneman, D., Sibony, O., & Sunstein, C. (2021). *Noise: A flaw in human judgment*. HarperCollins.

Kurtz, A. K., & Mayo, S. T. (1979). Percentiles and percentile ranks. In *Statistical methods in education and psychol-*

ogy (pp. 145–163). Springer. doi: 10.1007/978-1-4612-61292_6.

Lambert, M. J., & Harmon, K. L. (2018). The merits of implementing routine outcome monitoring in clinical practice. *Clinical Psychology: Science and Practice*, *25*, e12268–9. doi: 10.1111/cpsp.12268.

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*(3), 324–335. doi: 10 . 1001 / jamapsychiatry . 2018 . 3329.

McCall, W. A. (1922). *How to measure in education*. MacMillan Co. doi: 10.1037/13551-000.

McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S., & Lin, X. (2019). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, *76*, 1262–1272. doi: 10.1111/biom.13214.

McLean, C. P., & Anderson, E. R. (2009). Brave men and timid women? a review of the gender differences in fear and anxiety. *Clinical Psychology Review*, *29*(6), 496–505. doi: 10.1016/j.cpr.2009.05.003.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.

Pesce, L., van Veen, T., Carlier, I., van Noorden, M. S., van der Wee, N. J., van Hemert, A. M., & Giltay, E. J. (2016). Gender differences in outpatients with anxiety disorders: The leiden routine outcome monitoring study. *Epidemiol Psychiatr Sci*, *25*(3), 278–287. doi: 10.1017/s2045796015000414.

Prevolnik Rupel, V., Jagger, B., Fialho, L. S., Chadderton, L.-M., Gintner, T., Arntz, A., Baltzersen, Å.-L., Blazdell, J., van Busschbach, J., Cencelli, M., Chanen, A., Delvaux, C., van Gorp, F., Langford, L., McKenna, B., Moran, P., Pacheco, K., Sharp, C., Wang, W., et al. (2021). Standard set of patient-reported outcomes for personality disorder. *Quality of Life Research*, *30*(12), 3485–3500. doi: 10.1007/s11136-021-02870-w.

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*(3), 717–746. doi: 10.1007/s11336-021-09776-z.

Sigmon, S. T., Pells, J. J., Boulard, N. E., Whitcomb-Smith, S., Edenfield, T. M., Hermann, B. A., LaMattina, S. M., Schartel, J. G., & Kubik, E. (2005). Gender differences in self-reports of depression: The response bias hypothesis revisited. *Sex Roles*, *53*, 401–411. doi: 10.1007/s11199-005-6762-3.

Soloman, S. R., & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods*, *8*(2), 9. doi: 10.22237/jmasm/1257034080.

Tellegen, A., & Ben-Porath, Y. S. (1992). The new uniform t scores for the mmpi-2: Rationale, derivation, and appraisal. *Psychological Assessment*, *4*(2), 145–155. doi: 10.1037/1040-3590.4.2.145.

Tellegen, A., & Ben-Porath, Y. S. (2003). *Mmpi-2-rf (minnesota multiphasic personality inventory-2-restructured form) manual for administration, scoring, and interpretation*.

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The mmpi–2 restructured clinical scales: Development, validation, and interpretation*. University of Minnesota Press.

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on the journal's web site.

Citation

de Beurs, E., & Cloutier, J. (2024). Test multidimensionnel de personnalité (tmdp): Norms in T-scores and percentile rank order scores for the general and clinical population. *The Quantitative Methods for Psychology*, *20*(3), 205–216. doi: 10.20982/tqmp.20.3.p205.

Received: 30/05/2024 ~ Accepted: 24/07/2024

Tables 4 and 5 follow.

Table 4 ■ Raw Scores and Linear T-scores and Normalized T-scores (TIRT or TRankit) for the Various Scales of the TMDP.

Raw	Pessimism		Anxiety		Depression		Social isolation		Perc. paren. care		Psychoticism		Paranoid ideation		Antisoc. behavior		Impulsiveness		Hypomania	
score	T_{linear}	T_{IRT}	T_{linear}	T_{IRT}	T_{linear}	T_{IRT}	T_{linear}	T_{IRT}	T_{linear}	T_{IRT}	T_{linear}	T_{IRT}	T_{linear}	T_{Rankit}	T_{linear}	T_{IRT}	T_{linear}	T_{IRT}	T_{linear}	T_{IRT}
0	41.5	42.8	36.6	37.6	37.8	40.9	36.4	36.5	40.0	43.0	41.0	38.1	40.0	36.6	37.1	37.3	37.1	37.8	35.8	38.6
1	44.5	47.3	39.4	41.6	39.8	42.6	38.6	39.7	43.8	47.3	45.1	46.8	43.8	44.4	40.4	41.5	39.7	41.8	39.6	43.1
2	47.5	51.1	42.2	45.1	41.8	44.4	40.8	42.6	47.7	51.0	49.3	52.9	47.6	49.6	43.6	45.3	42.3	45.3	43.3	47.2
3	50.5	54.2	45.0	48.1	43.8	46.1	42.9	45.1	51.5	54.1	53.4	56.8	51.4	54.0	46.9	48.7	44.9	48.5	47.1	50.9
4	53.5	56.8	47.8	50.8	45.8	47.9	45.1	47.4	55.4	56.8	57.5	59.1	55.2	58.3	50.2	51.7	47.6	51.2	50.9	54.4
5	56.5	58.9	50.6	53.3	47.7	49.6	47.3	49.5	59.2	59.1	61.6	62.0	59.0	62.1	53.4	54.4	50.2	53.7	54.6	57.6
6	59.5	60.7	53.3	55.5	49.7	51.3	49.5	51.3	63.1	61.2	65.7	64.6	62.8	65.3	56.7	56.9	52.8	55.9	58.4	60.5
7	62.6	62.3	56.1	57.6	51.7	53.0	51.6	53.0	66.9	63.0	69.8	66.5	66.6	68.1	60.0	59.1	55.4	58.0	62.1	63.3
8	65.6	63.8	58.9	59.5	53.7	54.7	53.8	54.6	70.8	64.7	74.0	68.7	70.4	71.2	63.2	61.3	58.0	60.0	65.9	66.0
9	68.6	65.4	61.7	61.5	55.7	56.4	56.0	56.0	74.6	66.5	78.1	70.9	74.2	74.6	66.5	63.3	60.7	61.9	69.7	68.6
10	71.6	67.1	64.5	63.5	57.7	58.1	58.2	57.4	78.5	68.2	82.2	73.3	78.0	79.9	69.8	65.3	63.3	63.9	73.4	71.1
11	74.6	69.1	67.3	65.5	59.6	59.8	60.3	58.7	82.3	70.2	86.3	77.7	81.8	NA	73.0	67.3	65.9	65.9	77.2	73.7
12	77.6	71.5	70.1	67.7	61.6	61.5	62.5	60.1	86.2	72.4	90.4	NA	85.6	NA	76.3	69.4	68.5	68.1	80.9	76.3
13	80.6	74.4	72.8	70.2	63.6	63.1	64.7	61.4	90.0	75.0	94.5	NA	89.4	NA	79.6	71.6	71.1	70.4	84.7	79.0
14	83.6	78.0	75.6	72.9	65.6	64.8	66.8	62.8	93.8	78.0	98.6	NA	93.2	NA	82.8	74.0	73.7	73.1	88.5	81.9
15	86.7	82.4	78.4	75.9	67.6	66.4	69.0	64.3			102.8	NA	97.0	NA	86.1	76.7	76.4	76.0	92.2	84.9
16			81.2	79.4	69.6	68.0	71.2	65.9			106.9	NA	100.8	NA	89.4	79.6	79.0	79.3	96.0	88.2
17			84.0	83.3	71.5	69.6	73.4	67.7			111.0	NA			92.6	82.9	81.6	NA	99.7	91.8
18					73.5	71.2	75.5	69.7			115.1	NA					84.2	NA	103.5	95.7
19					75.5	72.8	77.7	71.8			119.2	NA							107.3	NA
20					77.5	74.4	79.9	74.2											111.0	NA
21					79.5	76.0	82.1	76.9												
22					81.5	77.6	84.2	79.9												
23					83.5	79.1														
24					85.4	80.7														

Note. Perc. paren. care: Perceived parental care; Antisoc. behavior: Antisocial behavior; T_{Linear}: linear T-score; T_{IRT}: IRT-based T-score

Table 5 ■ Means for Age Groups

	All; $N = 2156$		Correlation	16-26; $n = 557$		27-36; $n = 620$		37-44; $n = 498$		45-79; $n = 481$		Significance			
Scale	M	SD	r	M	SD	M	SD	M	SD	M	SD	F	p	η^2	Pairwise comp.
Pessimism	7.53	4.83	−.03	7.48	4.84	7.56	4.87	7.72	4.82	7.34	4.80	0.53	.660	0.001	
Anxiety	8.88	4.66	−.11∗	9.36	4.67	9.08	4.72	8.84	4.55	8.10	4.63	6.91	<.001	0.010	$1 > 4, 2 > 4$
Depression	12.54	6.59	.01	12.13	6.62	12.47	6.63	13.01	6.56	12.63	6.53	1.61	.186	0.002	
Social isolation	9.04	5.33	.02	8.87	5.44	8.92	5.38	9.24	5.50	9.17	4.96	0.63	.593	0.001	
Perceived parental care	5.22	3.8	.08∗	4.18	3.46	5.61	3.87	5.84	3.92	5.28	3.74	22.69	<.001	0.028	$1 < 2, 1 < 3, 1 < 4$
Psychoticism	4.04	3.67	−.06∗	4.12	3.74	4.06	3.63	4.15	3.77	3.80	3.55	0.92	.430	0.001	
Paranoid ideation	5.53	3.94	.03	5.55	3.97	5.81	4.09	5.66	3.90	5.01	3.71	4.25	.005	0.006	$2 > 4, 3 > 4$
Antisocial behavior	4.38	3.26	−.21∗	5.27	3.44	4.74	3.41	3.85	2.93	3.43	2.79	37.49	<.001	0.048	$\begin{cases} 1 > 2, 1 > 3, 1 > 4, \\ 2 > 3, 2 > 4 \end{cases}$
Impulsivity	7.55	4.68	−.16∗	8.13	4.83	8.11	4.80	7.45	4.52	6.26	4.26	20.10	<.001	0.025	$1 > 4, 2 > 4, 3 > 4$
Hypomania	6.65	3.73	−.14∗	7.09	3.84	7.00	3.90	6.55	3.61	5.80	3.33	14.33	<.001	0.018	$1 > 4, 2 > 4, 3 > 4$

Note. *: p < .001