Iran J Psychiatry 2026; 21: 4: 1-15

Original Article

Gender-Related Fairness of Personality Psychopathology Five (PSY-5) Dimensions: Gender Differences in PSY-5 Functioning

Mehdi Rahmani Malek-Abad[1*], Mohammad Reza Abedi[2]

Abstract

Objective: Establishing measurement invariance is essential to ensure that Minnesota Multiphasic Personality Inventory (MMPI-3) scores accurately reflect the intended latent traits and can be interpreted fairly across gender groups. Given the importance of measurement equivalence for valid clinical interpretation, this study evaluated Differential Item Functioning (DIF) and Differential Test Functioning (DTF) in the Personality Psychopathology Five (PSY-5) scales of the Persian version of the MMPI-3 within an Iranian clinical sample.

Method: A total of 871 psychiatric outpatients (59% female; mean age = 28.76 ± 10.58) diagnosed with at least one DSM-5 disorder were recruited using convenience sampling. Invalid response protocols were excluded according to MMPI-3 validity criteria. Dimensionality of the PSY-5 scales was assessed via confirmatory factor analysis and Rasch measurement modeling, leading to the exclusion of the Disconstraint scale due to multidimensionality. Rasch-based analyses were used to evaluate item fit, reliability, local dependence (Q3), and DIF. Overall DTF was quantified using Signed and Unsigned Test Difference statistics, and Fisher's Z tests were used to compare inter-scale correlation patterns across gender.

Results: Significant DTF was observed for the four retained PSY-5 scales (Aggressiveness, Psychoticism, Negative Emotionality, and Introversion) following the exclusion of DISC due to multidimensionality. STDS values ranged from −0.27 to −0.37 and UTDS values ranged from 0.28 to 0.43, indicating consistent gender-related score differences of small-to-moderate magnitude at comparable latent trait levels that may influence clinical interpretation. Inter-scale correlations also differed across gender groups, particularly for PSYC, NEGE, and INTR, suggesting gender-related variation in scale functioning and construct associations.

Conclusion: The findings indicate that several MMPI-3 PSY-5 scales may not demonstrate full measurement equivalence across genders in Iranian clinical populations. These results highlight the importance of local psychometric evaluation and careful interpretation of gender-based score comparisons when applying standardized personality assessment instruments in non-Western clinical settings.

Key words: *Gender Equity; Gender Bias; Psychometrics; Minnesota Multiphasic Personality Inventory(MMPI-3); Patients*

1. Department of Educational Psychology, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran.
2. Department of Counseling, Faculty of Education and Psychology, University of Isfahan, Isfahan, Iran.

***Corresponding Author:**
Address: Department of Educational Psychology, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran, Postal Code: 8418148499.
Tel: 98-9131693034, Fax: 98-031 33660088, Email: m.rahmani_psy@yahoo.com, m_rahmani@iau.ac.ir

Article Information:
Received Date: 2026/02/21, Revised Date: 2026/05/31, Accepted Date: 2026/08/17

The Minnesota Multiphasic Personality Inventory (MMPI) is among the most widely used instruments for assessing psychopathology and personality characteristics. Through successive revisions, including the MMPI-2, MMPI-2-RF, and most recently the MMPI-3, the instrument has been refined to improve its psychometric properties and clinical utility (1). A central component of the MMPI-3 is the Personality Psychopathology Five (PSY-5) model, which assesses five broad dimensions of maladaptive personality functioning: Aggressiveness, Psychoticism, Disconstraint, Negative Emotionality/Neuroticism, and Introversion/Low Positive Emotionality. These dimensions are conceptually linked to contemporary dimensional models of personality and psychopathology (2).

In recent years, PSY-5 scales have been widely applied in clinical, research, and forensic contexts, making the evaluation of measurement equivalence across demographic groups increasingly important for valid score interpretation and evidence-based decision-making (3–7). When measurement invariance is not established, observed score differences may reflect variations in item functioning rather than true differences in the underlying constructs. Despite the widespread use of MMPI-3 PSY-5 scales, evidence regarding measurement invariance and differential item functioning (DIF), particularly across gender within clinical populations, remains limited. Existing studies have primarily relied on CFA- or IRT-based structural analyses conducted in non-clinical samples (8–10), whereas comparatively little attention has been given to how gender- and culture-related factors may influence PSY-5 scale functioning in clinical settings (11-13).

Measurement invariance is commonly evaluated through Differential Item Functioning (DIF) and Differential Test Functioning (DTF) analyses. DIF occurs when individuals from different groups who possess the same level of the underlying trait have different probabilities of endorsing an item. DTF extends this concept to the scale level and evaluates whether the cumulative impact of DIF across items results in biased test scores. Importantly, the presence of DIF does not necessarily imply meaningful bias at the total-score level, whereas DTF evaluates whether the cumulative effects of DIF are sufficiently large to influence overall score interpretation and cross-group comparisons. These approaches provide important evidence regarding measurement fairness and the validity of score comparisons across demographic groups (14).

Examining DIF and DTF at the scale level enables the determination of whether a set of items, as a whole, functions differently between women and men in clinical samples (14, 15). Morris, Shepherd, and Ingram demonstrated sex differences on the MMPI-3 Eating Concerns (EAT) scale, with women scoring higher than men (9). Anestis reported significant racial differences in clinical T-scores across MMPI-3 scales (8), and Shumaker confirmed racial differences on the Anger Proneness scale (10). Using the Rasch model, Gori and Marin identified differential item functioning (DIF) in the MMPI-2 Mt, DEP, and ANX scales (16). Meta-analytic evidence indicated that males show greater direct aggression than females, while indirect aggression differences are minimal (13). Rahmani Malek-Abad *et al*. found substantial item bias in the MMPI-2 depression, aggression, and anxiety scales, highlighting the importance of accounting for DIF in score interpretation (11, 12). Although MMPI-3 research has increasingly examined demographic differences across specific scales, empirical investigations of measurement invariance using DIF and DTF methodologies remain relatively limited. Existing studies have primarily focused on selected clinical or content scales and have generally examined either item-level or group-level differences rather than the cumulative impact of DIF on scale scores. Furthermore, evidence regarding gender-related DIF and DTF in the MMPI-3 PSY-5 scales within clinical populations remains particularly scarce, especially in non-Western settings where cultural factors may influence symptom expression and response patterns.

Despite these investigations, existing studies have primarily focused on mean score differences, racial comparisons, or the structural validity of MMPI scales, and most have been conducted in non-clinical samples (17). To our knowledge, no study has systematically examined both item-level DIF and test-level DTF for the MMPI-3 PSY-5 scales across gender within a clinical population. This gap is particularly important because PSY-5 scales are directly used in clinical formulations and diagnostic decision-making, and any systematic measurement bias may affect the validity of score interpretation (18). Therefore, the present study was designed to examine item-level DIF and test-level DTF in the PSY-5 scales of the MMPI-3 across genders in a diverse clinical sample. Using modern IRT-based methods and DTF indices, this research seeks to determine whether the PSY-5 scales function equivalently across gender groups in clinical contexts. The findings may contribute to a more accurate interpretation of PSY-5 scores and support fairer clinical decision-making.

Materials and Methods

Participants

The study included clients attending outpatient psychology and psychiatry clinics over a three-month period who had been diagnosed with at least one psychological disorder according to DSM-5 criteria. Diagnoses were established by board-certified psychiatrists during routine clinical evaluations based on DSM-5 diagnostic criteria and were verified through a review of clinical records. Because psychiatric comorbidity is common in clinical populations,

participants were not excluded if multiple DSM-5 diagnoses were present. For descriptive purposes, each participant was classified according to the principal clinical diagnosis documented by the treating psychiatrist, whereas all eligible participants were retained for psychometric analyses regardless of comorbidity. Participants were recruited from outpatient psychology and psychiatry clinics affiliated with both public and private healthcare centers in Isfahan, Iran. A convenience sampling strategy with consecutive (sequential) recruitment was used over a three-month period. Inclusion criteria were: (a) age 18 years or older, (b) diagnosis of at least one DSM-5 psychiatric disorder, (c) ability to read and complete the questionnaires, and (d) provision of informed consent. Exclusion criteria included invalid MMPI-3 response patterns based on validity indicators and protocols with more than 10% missing responses.

Questionnaires were administered in a standardized individual setting in quiet clinical rooms under the supervision of a trained psychologist. Completion time for the assessment battery ranged from approximately 60 to 90 minutes. Of the initially recruited participants (N = 1034), 163 protocols were excluded due to invalid responding or excessive missing data, resulting in a final sample of 871 participants. Response validity was evaluated following MMPI-3 Technical Manual procedures (17). Protocols were excluded if T-scores were ≥ 80 on TRIN, VRIN, or CRIN; ≥ 100 on the F scale; or had > 10% missing items. After exclusions, 871 valid protocols remained (59% women; mean age = 28.76, SD = 10.58, range 18–66). All participants had completed at least a high school diploma: 40.2% held a diploma, 6.2% an associate degree, 45.7% a bachelor's degree, 6.9% a master's degree, and 1% a doctoral degree. For descriptive purposes, participants were categorized according to their principal clinical diagnosis documented in the medical record. Based on this classification, 8% had schizophrenia, 10% borderline personality disorder, 27% bipolar disorder, 37% depressive disorders, and 18% other psychiatric disorders (Table 1). Sample size adequacy was confirmed using G*Power for a small effect (d = 0.20), α = 0.05, and power = 0.80.

Table 1. Demographic and Clinical Characteristics of the Clinical Sample (N = 871)

Variable	Level	Mean	Std.
Age		28.76	10.58
Gender		n	%
	Men	357	41%
	women	514	59%
Education level			
	diploma	350	40.2%
	associate degree	54	6.2%
	bachelor's degree	398	45.7%
	master's degree	61	6.9%
	doctoral degree	8	1%
Diagnostic			
	Schizophrenia	70	8%
	Borderline Personality Disorder	87	10%
	Bipolar Disorder	235	27%
	Depression	322	37%
	Psychological Disorders	157	18%

Procedure

The study aimed to examine measurement invariance, Differential Item Functioning (DIF), and Differential Test Functioning (DTF) of the MMPI-3 PSY-5 scales in a clinical population. The Rasch model and Mantel–Haenszel method were applied to assess item-level measurement properties and group differences. Data collection was conducted over three months in coordination with clinic directors. Eligible participants were outpatients aged ≥ 18 years with at least a high school diploma. Written informed consent was obtained after participants were provided with detailed explanations regarding the study. Only protocols meeting validity criteria with ≥ 90% completed items were analyzed. All procedures adhered to the Declaration of Helsinki (19), and the APA ethical standards for psychological assessment, including confidentiality, voluntary participation, and safeguards to minimize potential harm. The study protocol was approved by the Ethics Committee of Islamic Azad University, Khomeinishahr Branch (IR.IAU.KHSH.REC.1404.049).

Instrument

Minnesota Multiphasic Personality Inventory-3 (MMPI-3): The primary instrument was the Minnesota Multiphasic Personality Inventory–Third Edition (MMPI-3), originally developed by Hathaway and McKinley (1943) and revised in 1970, 2001, 2003, 2008,

with the final 335-item version published by the University of Minnesota in 2020 (17). The MMPI-3 has a three-tier hierarchical structure, 10 validity indices (CRIN, VRIN, TRIN, F, L, K), and 42 clinical, restructured, content, and special scales. Items use a dichotomous format ["True/Yes" = 1, "False/No" = 0]. Validity and reliability have been supported internationally (1, 18, 20), with Cronbach's alpha ranging from 0.51–0.91 and test-retest coefficients from 0.38–0.92 (21). The Persian version used in the present study was based on the standardized adaptation developed and validated by Rahmani Malek-Abad (22). In the validation study involving 1,987 Iranian adults, the instrument demonstrated satisfactory reliability and validity, with psychometric characteristics broadly comparable to those reported in the original MMPI-3 standardization research, supporting its use in Iranian populations. This study focused on the Personality Psychopathology Five (PSY-5) scales: Aggressiveness (AGGR, 15 items), Psychoticism (PSYC, 20 items), Disconstraint (DISC, 18 items), Negative Emotionality (NEGE, 15 items), and Introversion (INTR, 14 items). Cronbach's alpha coefficients in the Iranian sample were AGGR = 0.63, PSYC = 0.82, DISC = 0.71, NEGE = 0.77, and INTR = 0.63 (22). To protect confidentiality, item content is not disclosed, and scale items were sequentially coded while preserving their original order. Item coding preserved the original MMPI-3 item sequence, with responses recorded in the same order as presented in the official instrument without reordering or restructuring.

Symptom Checklist-90-Revised (SCL-90-R): The SCL-90-R was included to examine convergent validity between PSY-5 dimensions and general psychological symptom severity. The SCL-90-R, developed by Derogatis (1977), is a self-report instrument designed to assess psychological symptoms. The Persian version of the SCL-90-R has been validated in Iranian clinical and non-clinical populations, demonstrating acceptable psychometric properties including internal consistency and construct validity (23). The instrument consists of 90 items and uses a five-point Likert-type scale, namely: not at all [0], a little bit [1], moderately [2], quite a bit [3], and extremely [4]. The SCL-90-R includes nine subscales: Somatization, Obsessive-Compulsive, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, and Psychoticism. An individual's score for each subscale is calculated by summing the responses to the items within that subscale and dividing by the number of items in the subscale. Therefore, the subscale score ranges from 0 to 4, with higher scores indicating greater severity of psychological symptoms (24).

Statistical Analyses

Descriptive statistics, including means and standard deviations, were calculated. In this study, the Rasch model was employed due to its advanced psychometric properties in examining measurement invariance. The Rasch model, by providing a framework based on item response theory, allows for a more precise assessment of item characteristics independent of the sample and a fair comparison of respondent performance (25). Furthermore, Rasch is an appropriate tool for identifying Differential Item Functioning (DIF), as it controls for ability level effects and helps detect true differences in group performance (e.g., gender) (32). For conducting Rasch-based analyses, two key assumptions must be satisfied: unidimensionality and local independence (33). These assumptions were evaluated before conducting DIF and DTF analyses because the violation of either assumption may compromise the validity of Rasch parameter estimation. Therefore, any scale failing to satisfy these prerequisites was excluded from subsequent Rasch-based analyses.

Unidimensionality of the MMPI-3 PSY-5 scales was examined using Confirmatory Factor Analysis (CFA) and Rasch-residual-based Principal Components Analysis (PCAR). CFA fit was assessed using CFI, GFI, AGFI, RMSEA, and CMIN/DF, with CFI, GFI, and AGFI values > 0.90 and RMSEA ≤ 0.05 indicating good fit, while CMIN/DF < 5 was considered acceptable (34-37). In PCAR, residuals from the primary dimension are analyzed to detect secondary dimensions or conflicts, while item and person reliability indices reflect reproducibility of item difficulty and respondent rankings, respectively. Explained variance ≥ 20% and unexplained variance in the first contrast < 2 eigenvalues and < 10% indicate unidimensionality (25- 26).

Local item independence was evaluated using Yen's Q3 statistic. Q3 values < 0.20 indicate local independence; positive values > 0.20 suggest local dependence, whereas negative values are generally non-problematic (27, 28). Item fit to the Rasch model was assessed using Infit and Outfit Mean Square statistics, and item difficulty was indicated by the Measure index. Differential Item Functioning (DIF) was identified via the Rasch model, providing higher precision than traditional methods (11, 12, 29). Items with a significance level of $P < 0.05$ were considered indicative of statistically significant DIF. Differential Item Functioning in the Rasch model was assessed using the DIF Contrast index with Welch probability tests, while in the Mantel–Haenszel method, the size index and its corresponding probability value were used. According to Linacre, size values < 0.43 indicate negligible DIF (A), 0.43–0.64 moderate DIF (B), and > 0.64 large DIF (C), with large DIF representing item bias favoring one gender. Differential Test Functioning (DTF) was evaluated using the Signed Test Difference (STDS) and Unsigned Test Difference (UTDS) indices. For DTF, values exceeding ± 1.0 for STDS or UTDS were interpreted as indicating meaningful test-level bias (30). SPSS version 27 was employed for descriptive statistics and Pearson correlations, CFA was conducted in AMOS version 26, and Rasch analyses, principal component analyses of residuals (PCAR), Yen's Q3 statistics, item

fit analyses, DIF analyses (Rasch and Mantel–Haenszel methods), and DTF indices (STDS and UTDS) were computed using Winsteps version 4.1.0.

Results

Initially, descriptive findings for the Personality Psychopathology Five (PSY-5) scales were examined. The Mean ± Standard deviation values for the AGGR, PSYC, DISC, NEGE, and INTR scales were 7.72 ± 2.65, 6.68 ± 4.16, 6.17 ± 3.10, 8.50 ± 3.33, and 5.72 ± 2.87, respectively. Moreover, all skewness coefficients (-0.27 to 0.42) and kurtosis coefficients (-0.64 to -0.34) fell within the range of -2 to +2, indicating a normal distribution of the data (31).

Dimensionality

The Rasch model assumes unidimensionality of item responses. To evaluate this, Confirmatory Factor Analysis (CFA) and Principal Components Analysis of Residuals (PCAR) were conducted. RMSEA values for PSYC, NEGE, and INTR indicated good fit; AGGR and DISC were acceptable. Other fit indices were generally satisfactory, though DISC showed comparatively lower fit.

PCAR was applied to assess residual-based unidimensionality (26, 38). Results showed that all PSY-5 scales met the explained variance criteria. Unexplained variance in the first contrast was < 2 eigenvalues and < 10% for AGGR, PSYC, NEGE, and INTR, confirming unidimensionality. The DISC scale exceeded these thresholds (2.37 eigenvalues, 10.4%), indicating multidimensionality, Since the DISC scale failed to satisfy the Rasch assumption of unidimensionality, it was excluded from all subsequent Rasch-based DIF and DTF analyses. Having established the dimensional structure of the retained PSY-5 scales, subsequent Rasch analyses focused on evaluating measurement precision, local independence, and item-level model fit before proceeding to DIF and DTF analyses.

In Table 2, Item Reliability, which indicates the reproducibility of the item difficulty hierarchy across similar samples, and Person Reliability, which reflects the reproducibility of the ordering of individuals based on the measured ability or latent trait, are also presented. The range of these coefficients, similar to conventional reliability indices, is between 0 and 1, and the closer the value is to 1, the more desirable it is. A value of 1 in Item Reliability means that if the same items are administered to another sample with similar characteristics, the difficulty ordering of the items will remain virtually the same. Similarly, a value of 1 in Person Reliability indicates that the ranking of respondents according to their ability is stably reproducible (26, 39). The range of Person Reliability coefficients showed that the test was able to distinguish respondents into an adequate number of levels (two or three strata) for analytic purposes. The high Item Reliability indicates that the sample size was sufficiently large to precisely locate the items along the latent variable continuum (26).

Table 2. Reliability, Unidimensionality, and Model-Data Fit of the MMPI-3 PSY-5 Scales Based on CFA and PCAR

Scale	CFA					PCAR		Reliability	
	RMSEA	CFI	AGFI	GFI	CMIN/DF	Variance explained %	Unexplained variance 1st contrast	Unexplained variance 1st contrast %	Item	Person
AGGR	0.06	0.95	0.91	0.95	4.64	23.1%	1.88	9.6%	0.99	0.65
PSYC	0.04	0.94	0.92	0.95	3.02	23.3%	1.84	7.1%	0.98	0.72
DISC	0.07	0.92	0.91	0.94	4.75	21.4%	2.37	10.4%	0.99	0.62
NEGE	0.05	0.95	0.93	0.96	4.06	26.5%	1.78	8.7%	0.99	0.70
INTR	0.04	0.96	0.95	0.97	2.71	23.6%	1.62	8.9%	0.98	0.66

Note. AGGR = Aggressiveness, PSYC = Psychoticism, DISC = Disconstraint, NEGE = Negative Emotionality, INTR = Introversion, CFA = Confirmatory Factor Analysis, PCAR = Principal Component Analysis of Residuals, MMPI = Minnesota Multiphasic Personality Inventory, PSY-5 = Personality Psychopathology Five

Local Independence and Model–Data Fit

Local independence, a key assumption of the Rasch model, was assessed using Yen's Q3 statistic, which measures correlations between item residuals (27, 28). Results showed acceptable local independence for AGGR, INTR, and PSYC items (Q3 < 0.20). One NEGE item (Item 2) exhibited Q3 = 0.30, indicating local dependence, and was removed. Rasch model fit was evaluated using Infit and Outfit Mean Square (MNSQ) statistics. Values 0.5–1.5 indicate acceptable fit, 1.5 – 2.0 are tolerable, < 0.5 indicates overfit, and > 2.0 indicates poor fit. In this clinical sample, acceptable MNSQ was set at 0.5 – 1.7 (25). ZSTD values were expected to fall within ± 2, though they are sensitive to sample size (26). The Measure index, reflecting item difficulty, was also reported. All retained items met these model–data fit criteria (Table. 3).

Table 3. Rasch Model Assumptions and Item Fit Statistics for the MMPI-3 PSY-5 Scales

AGGR					NEGE				
Item Number	Measure	Infit MNSQ	Outfit MNSQ	Q3	Item Number	Measure	Infit MNSQ	Outfit MNSQ	Q3
8	0.31	1.18	1.27	-0.19	2	1.01	1.35	1.79	0.30
1	-0.71	1.16	1.23	-0.14	14	0.24	1.23	1.70	-0.18
3	-1.42	1.04	1.16	-0.14	7	1.45	1.13	1.14	-0.16
6	0.21	1.04	1.04	-0.14	8	-0.41	1.05	1.11	-0.04
13	1.04	1.01	1.03	-0.15	1	-0.75	1.05	0.99	-0.15
5	-0.94	1.00	1.02	0.17	12	-1.02	1.04	1.00	-0.26
2	-0.17	0.97	0.99	-0.14	3	0.76	1.02	0.98	-0.14
10	-0.98	0.99	0.97	-0.18	10	0.53	1.01	0.95	-0.15
4	0.52	0.97	0.92	-0.15	5	-0.59	0.92	0.81	-0.15
15	0.84	0.97	0.94	-0.22	11	0.27	0.88	0.84	-0.14
12	-0.58	0.96	0.92	-0.14	13	-0.42	0.88	0.80	-0.19
11	0.44	0.95	0.93	-0.14	4	-0.38	0.87	0.81	0.14
14	0.63	0.94	0.93	-0.17	6	-.09	0.87	0.79	-0.14
7	0.04	0.92	0.89	-0.16	9	0.01	0.86	0.78	-0.14
9	0.80	0.88	0.85	-0.19	15	-0.62	0.86	0.75	-0.16
PSYC					**INTR**				
Item Number	**Measure**	**Infit MNSQ**	**Outfit MNSQ**	**Q3**	**Item Number**	**Measure**	**Infit MNSQ**	**Outfit MNSQ**	**Q3**
1	-1.45	1.28	1.51	-0.17	7	-0.66	1.17	1.25	-0.16
3	-0.34	1.29	1.49	-0.13	1	-0.69	1.19	1.23	-0.20
10	-0.60	1.18	1.20	-0.16	14	-0.62	1.14	1.19	-0.18
19	-0.57	1.14	1.14	-0.14	2	0.34	1.08	1.13	-0.12
18	-0.04	1.04	1.08	-0.09	5	1.13	1.05	1.07	-0.15
4	0.48	1.00	0.93	-0.10	11	-0.50	1.02	1.05	-0.13
7	-0.40	1.00	0.97	-0.08	13	-0.26	1.02	1.04	-0.14
2	-0.15	0.99	0.94	0.13	9	0.25	0.99	0.95	-0.14
12	-0.56	0.99	0.95	-0.20	6	-0.29	0.95	0.95	-0.18
20	-0.29	0.99	0.95	-0.07	10	0.52	0.92	0.87	-0.13
8	0.00	0.98	0.92	-0.16	3	-0.02	0.91	0.85	-0.18
11	-0.29	0.97	0.94	-0.12	4	-0.39	0.90	0.85	-0.15
9	-0.05	0.96	0.91	-0.05	8	0.78	0.87	0.74	-0.20
5	1.34	0.93	0.68	0.14	12	0.42	0.80	0.70	-0.12
6	1.03	0.93	0.87	-0.2	**Mean**				
15	0.41	0.89	0.76	0.13	**Scale**	**Measure**	**Infit MNSQ**	**Outfit MNSQ**	
13	0.57	0.88	0.74	-0.13	AGGR	0.0	1.00	1.01	
14	0.43	0.87	0.71	-0.14	PSYC	0.0	1.00	0.95	
17	0.62	0.87	0.69	-0.13	NEGE	0.0	1.00	0.99	
16	0.58	0.81	0.62	-0.14	INTR	0.0	1.00	0.99	

Note. Bolded items indicate borderline misfit but retained due to acceptable fit criteria, MMPI = Minnesota Multiphasic Personality Inventory, PSY-5 = Personality Psychopathology Five.

Rasch model diagnostics indicated that the AGGR, PSYC, NEGE, and INTR scales demonstrated acceptable psychometric performance. Item-fit statistics generally fell within recommended ranges, and no substantial violations of local independence were observed. Item difficulty estimates covered a broad range of the latent traits, supporting adequate measurement precision across different levels of psychopathology. Detailed item-level fit statistics are presented in Table 3. Item difficulty was expressed in Logit units (Measure index) (26). In the AGGR scale, Item 13 was the most difficult (1.04) and Item 3 was the easiest (−1.42). For the PSYC, NEGE, and INTR scales, the most difficult items were 5 (1.34), 7 (1.45), and 5 (1.13), respectively, while the easiest items were 1 (−1.45), 12 (−1.02), and 1 (−0.69). These findings indicate that, after excluding the few problematic items, the retained item sets provided a stable basis for subsequent DIF and DTF analyses, thereby improving the interpretability of gender comparisons at both the item and scale levels.

Infit and Outfit MNSQ statistics were examined to evaluate item–model fit. For AGGR, Infit ranged from 0.88 to 1.18 and Outfit from 0.85 to 1.27; for PSYC, Infit ranged from 0.81 to 1.28 and Outfit from 0.62 to 1.51; for INTR, Infit 0.80 – 1.17 and Outfit 0.70 – 1.25—all within the acceptable range of 0.50 – 1.70. In the NEGE scale, Outfit MNSQ values for Items 2 (1.79) and 14 (1.70) exceeded the predefined acceptable threshold, indicating insufficient conformity with the Rasch model expectations. Because retaining misfitting items may distort item parameter estimation and subsequently affect DIF and DTF analyses, these two items were excluded from subsequent Rasch-based analyses. The remaining items demonstrated a satisfactory model fit. Overall, the remaining items showed a satisfactory fit to the Rasch model, consistent with previous findings on MMPI-2-RF scales (17, 40).

DIF and DTF

After confirming the Rasch model fit, differential item functioning (DIF) across gender was examined for the MMPI-3 PSY-5 scales. Both the Rasch model and the Mantel–Haenszel method were employed to identify potential item bias. Items with a probability < 0.05 were considered significant. Table 4 summarizes items with significant DIF, including the direction of bias: items favoring males are marked “M” and those favoring females are marked “F.”

Table 4. Differential Item Functioning of the Personality Psychopathology Five Scales Using the Rasch Model and Mantel–Haenszel Method Across Gender

Scale	Item Number	Rasch Model			Mantel Mantel Metod			Direction of DIF
		DIF Contrast	Welch Prob.	ETS DIF Category	Size	Prob.	ETS DIF Category	
AGGR (15 item)	1	-0.45	0.01	B	-0.41	0.02	A	F
	3	0.41	0.02	A	0.38	0.05	A	--
	4	0.59	0.01	B	0.59	0.01	B	M
	5	0.33	0.04	A	0.37	0.04	A	--
	6	-0.42	0.01	A	-0.39	0.02	A	--
	7	-0.65	0.01	C	-069	0.01	C	F
	9	-0.66	0.01	C	-0.77	0.01	C	F
	11	-0.71	0.01	C	-0.69	0.01	C	F
	13	-036	0.03	A	-0.34	0.05	A	--
	14	-0.46	0.01	B	-0.50	0.01	B	F
	15	-0.74	0.01	C	-0.76	0.01	C	F
PSYC (20 items)	2	-0.54	0.01	B	-0.57	0.01	B	F
	7	-0.43	0.02	B	-0.44	0.02	B	F
	12	-0.36	0.03	A	-0.40	0.02	A	--
	13	-0.59	0.01	B	-0.67	0.01	C	F
	14	-0.48	0.01	B	-0.52	0.01	B	F
	15	-0.43	0.02	B	-0.46	0.02	B	F
	16	-0.38	0.04	A	-0.49	0.03	B	--
	17	-0.54	0.01	B	-0.55	0.01	B	F
	20	-0.51	0.01	B	-0.52	0.01	B	F

Scale	Item Number	Rasch Model			Mantel Mantel Metod			Direction of DIF
		DIF Contrast	Welch Prob.	ETS DIF Category	Size	Prob.	ETS DIF Category	
NEEG (15 items)	3	-0.64	0.01	C	-0.82	0.01	C	F
	5	-0.60	0.01	B	-0.85	0.01	C	F
	6	-1.12	0.01	C	-0.78	0.01	C	F
	7	-0.42	0.02	A	-0.33	0.06	A	--
	9	-0.41	0.02	A	-0.35	0.05	A	--
	12	-0.48	0.01	B	-0.74	0.01	C	F
INTR (14 Items)	2	-1.36	0.01	C	-1.28	0.01	C	F
	4	-0.46	0.01	B	-0.58	0.01	B	F
	5	-0.49	0.02	B	-0.47	0.03	B	F
	8	-0.66	0.01	C	-0.92	0.01	C	F
	10	-0.45	0.01	B	-0.59	0.01	B	F
	12	-0.65	0.01	C	-1.05	0.01	C	F
	14	-0.48	0.01	B	-0.39	0.02	A	F

Note. A = ETS negligible, B = ETS moderate, C = ETS large, M = Male, F = Female, -- = Not DIF.

Differential item functioning analyses identified multiple items exhibiting significant gender-related DIF across all four PSY-5 scales (Table 4). The AGGR scale showed the largest number of DIF items, followed by PSYC, INTR, and NEGE. Most DIF effects favored female respondents, although a limited number of items favored males or showed negligible DIF. The consistency of findings across both Rasch and Mantel–Haenszel procedures supports the robustness of the observed item-level non-invariance.

DIF was examined across genders for the PSY-5 scales. In AGGR, 4 items showed large DIF, 3 showed moderate DIF, and 4 exhibited small DIF. In PSYC, 7 items showed moderate and 2 showed small DIF. In NEGE, 2 items showed large, 2 showed moderate DIF, and 2 showed small DIF. In INTR, 3 items showed large DIF and 4 exhibited moderate DIF (Table 4 and Figure 1). A comparison between the Rasch model and the Mantel–Haenszel method using the McNemar–Bowker test ($\chi^2 = 3.33$, df = 2, $P > 0.05$) indicated no significant differences, suggesting the consistent detection of DIF items across methods (26). Item-level bias was observed: in AGGR, six items favored men and one favored women; in PSYC, seven favored women; in NEGE, four favored women; and in INTR, seven favored women.

Although several items demonstrated significant gender-related DIF, item-level differential functioning does not necessarily translate into bias at the scale level (15), because opposing DIF effects may offset one another (14, 41). Therefore, DTF analyses were conducted to determine whether the cumulative impact of DIF produced meaningful differences in total scale scores (42, 43).

The DTF analysis revealed statistically significant gender-related differential functioning across all four PSY-5 scales, with Z values ranging from 12.33 to 13.87 (all $P < 0.01$). The direction of DTF was consistently in favor of women across all scales, as indicated by negative signed test difference scores (STDS). Specifically, the AGGR scale yielded an STDS of −0.34 and an unsigned test difference score (UTDS) of 0.43 ($Z = 13.79$, $P < 0.01$). Comparable patterns were observed for PSYC (STDS = −0.27, UTDS = 0.28, $Z = 12.33$, $P < 0.01$), NEGE (STDS = −0.37, UTDS = 0.37, $Z = 13.17$, $P < 0.01$), and INTR (STDS = −0.35, UTDS = 0.36, $Z = 13.87$, $P < 0.01$). Thus, all four PSY-5 scales showed statistically significant DTF, with the magnitude of the unsigned differences remaining relatively small across scales. The consistently negative STDS values indicate that women tended to have slightly higher expected scale scores than men at comparable levels of the underlying latent traits. Importantly, these findings indicate differential test functioning rather than definitive measurement bias, because observed gender-related differences may reflect a combination of measurement-related factors and genuine differences in the expression of personality psychopathology across genders.

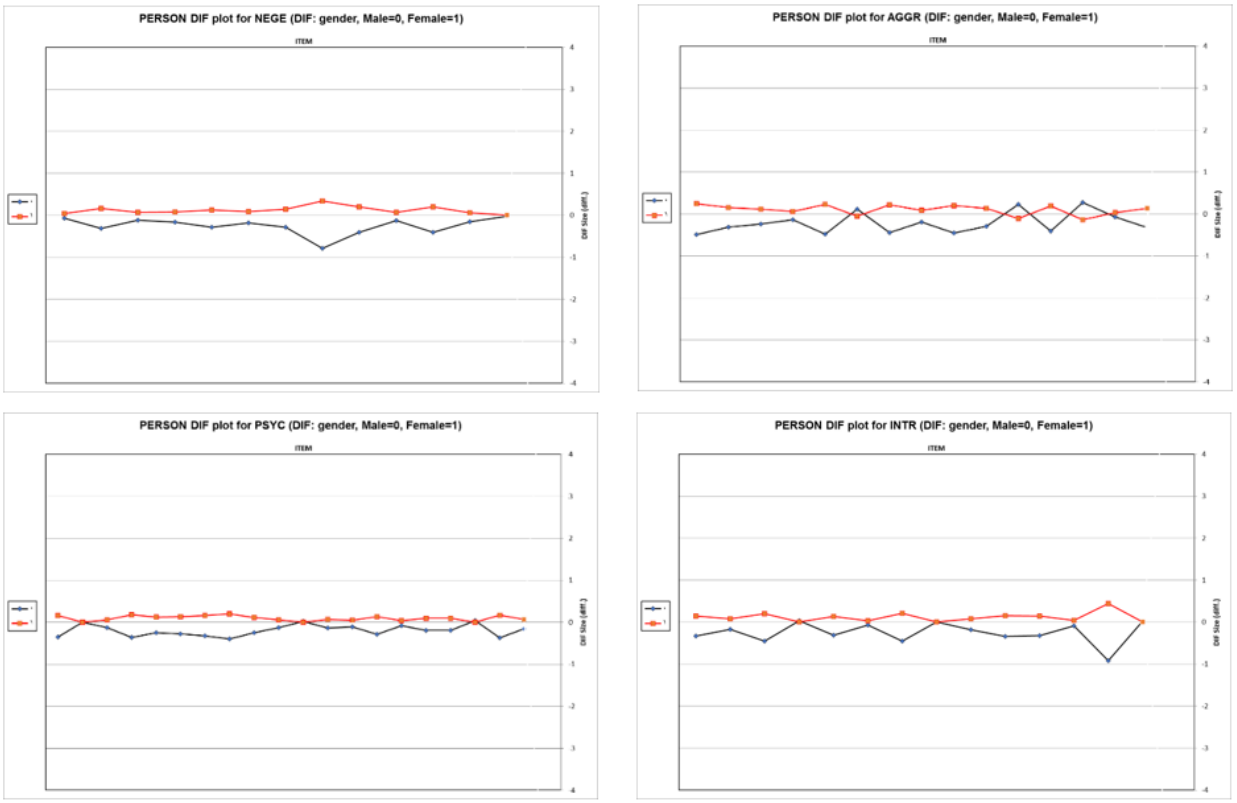

Figure 1. Differential Item Functioning Plot of the Personality Psychopathology Five Scales Across Gender (0 = Male, 1 = Female)

Convergent Validity and Cross-Gender Structural Difference

The investigation of item and test differential functioning aims to enhance the precision and fairness of test interpretation and has only a limited impact on test validity. Therefore, to ensure the validity and interpretability of the PSY-5 results, the convergent validity of these scales with the dimensions of the Symptom Checklist-90 (SCL-90) was examined. Pearson's correlation coefficient (r) was used to assess validity, with r = 0.10 considered small, r = 0.30 moderate, and r = 0.50 large (30, 44).

According to the results presented in Table 5., AGGR, PSYC, DISC, NEGE, and INTR demonstrated significant positive correlations with all dimensions of the SCL-90 at $P < 0.01$. All PSY-5 scales showed significant positive correlations with SCL-90 dimensions at $P < 0.01$ (Table 5). Following the confirmation of differential test functioning (DTF) among the scales, the study further examined gender-based differences in the correlations between the PSY-5 scales of the MMPI-3 and the corresponding dimensions of the SCL-90. For this purpose, correlation coefficients were first computed separately for men and women, and differences between the two groups' correlations were assessed using Fisher's Z transformation (45). The correlation between AGGR and Hostility showed no significant difference between men and women ($Z = 0.01$, $P = 0.87$), indicating comparable relational structures ($r = 0.46$ in both groups). However, significant gender differences were observed in several other correlations. The correlation between PSYC and Psychoticism was significantly stronger among men ($r = 0.83$) than women ($r = 0.59$), $Z = 3.08$, $P = 0.01$. Similarly, the association between NEGE and Phobic Anxiety was significantly higher in women ($r = 0.55$) compared to men ($r = 0.23$), $Z = -2.31$, $P = 0.01$. The INTR–Interpersonal Sensitivity correlation also differed, being stronger in women ($r = 0.59$) than men ($r = 0.32$), $Z = -2.07$, $P = 0.01$. These findings suggest gender-specific patterns in the associations between PSY-5 domains and relevant clinical constructs, which may reflect differences in symptom manifestation or underlying personality–psychopathology structures across genders.

Table 5. Means, Standard Deviations, and Correlations Between the Personality Psychopathology Five Scales and Symptom Checklist-90 Dimensions

	Dimensions	Mean	Std.	Correlation				
				AGGR	PSYC	DISC	NEGE	INTR
MMPI-3	AGGR	7.72	2.65	1.00				
	PSYC	6.68	4.16	0.46**	1.00			
	DISC	6.17	3.10	0.55**	0.68**	1.00		
	NEGE	8.50	3.33	0.38**	0.36**	0.48**	1.00	
	INTR	5.72	2.87	0/19	0.23**	0.33**	0.38**	1.00
SCL-90	Somatization	1.15	0.72	0.37**	0.63**	0.67**	0.61**	0.48**
	Obsessive-Compulsive	1.16	0.96	0.52**	0.61**	0.61**	0.64**	0.41**
	Interpersonal Sensitivity	1.19	0.94	0.50**	0.55**	0.62**	0.70**	0.50**
	Depression	1.20	0.94	0.37**	0.52**	0.54**	0.69**	0.46**
	Anxiety	1.07	0.94	0.50**	0.63**	0.67**	0.64**	0.44**
	Hostility	0.96	0.94	0.52**	0.69**	0.78**	0.56**	0.46**
	Phobic Anxiety	0.76	0.84	0.44**	0.59**	0.65**	0.48**	0.53**
	Paranoid Ideation	1.37	1.02	0.62**	0.59**	0.68**	0.61**	0.44**
	Psychoticism	0.81	0.75	0.55**	0.69**	0.74**	0.57**	0.45**

Note. **P < 0.01

Discussion

This study investigated the gender-based measurement invariance of the PSY-5 scales of the MMPI-3 in an Iranian clinical sample using Rasch-based DIF and DTF analyses. Overall, the findings provide evidence that although the PSY-5 scales generally retain their psychometric utility across genders, several items and scale scores function differently for men and women, suggesting that gender-related sources of measurement non-invariance should be considered when interpreting MMPI-3 PSY-5 scores in clinical settings.

The dimensionality analyses supported the use of Rasch modeling for four PSY-5 scales (AGGR, PSYC, NEGE, and INTR), whereas the DISC scale did not satisfy unidimensionality requirements and was therefore excluded from subsequent DIF and DTF analyses. This finding is consistent with previous psychometric investigations indicating that Disconstraint may represent a relatively heterogeneous construct encompassing multiple behavioral and personality components (3, 5). Also, the relatively low proportion of explained variance across scales (< 30%) is consistent with previous Rasch-based studies in heterogeneous clinical samples and may reflect the multidimensional nature of psychopathological constructs. Consequently, the interpretation of DISC scores may require additional caution when examined within culturally diverse clinical populations.

The most important finding of the present study was the presence of significant gender-related DIF and DTF across the AGGR, PSYC, NEGE, and INTR scales. Although item-level DIF does not necessarily compromise the validity of total scores, the observed DTF results indicate that the cumulative impact of DIF may influence scale-level interpretations. These findings are consistent with previous research demonstrating that demographic characteristics, including gender, can affect response patterns on MMPI-based personality and psychopathology measures (9, 11–15). Similarly, studies examining MMPI-3 and MMPI-2 family instruments have reported differential functioning across demographic groups, highlighting the importance of evaluating measurement invariance when psychological assessments are used for clinical decision-making (8–10, 16).

The observed gender differences in inter-scale correlations may further reflect differential expression patterns of internalizing and externalizing psychopathology across gender groups. Specifically, the observed gender differences in inter-scale correlations may reflect differences in the relationships between PSY-5 dimensions and associated clinical constructs across men and women. For example, the stronger association between PSYC and global symptom indices in men, and the stronger association between NEGE and phobic anxiety in women, suggest that the clinical correlates of these personality dimensions may vary across genders within this sample. However, these findings should be interpreted cautiously, as they may also reflect sample-specific clinical characteristics, differences in diagnostic composition, or cultural influences on symptom expression rather than stable gender-related patterns. This interpretation is consistent with contemporary psychometric theory, which emphasizes that observed score differences may arise from both substantive trait variation and measurement-

related factors (11-13). Beyond simple agreement or disagreement, these findings may reflect differences in sample clinical composition and cultural modulation of trait expression, which can partially account for inconsistencies across prior MMPI-3 and PSY-5 studies.

The presence of DIF items suggests that some item responses may be influenced by factors beyond the intended latent construct. However, such findings should not automatically be interpreted as evidence of measurement bias, as they may also reflect meaningful gender-related differences in the manifestation or expression of personality psychopathology. Therefore, the observed DIF and DTF effects are best interpreted as indicators of potential non-equivalence that warrant careful consideration in clinical assessment.

The cultural context of Iran may also help explain the observed pattern of gender-related DIF and DTF. Gender roles within Iranian society continue to influence the social acceptability of emotional expression, interpersonal behavior, and help-seeking practices. Emotional vulnerability, anxiety-related symptoms, and interpersonal sensitivity are generally more socially acceptable for women to express, whereas men may experience stronger expectations regarding emotional control, independence, and behavioral restraint. Such cultural norms may shape how respondents interpret and endorse specific PSY-5 items, particularly those assessing negative emotionality, introversion, and psychotic-like experiences. Accordingly, the observed differential functioning may reflect not only measurement-related factors but also culturally influenced patterns of symptom expression within this clinical population.

The findings also contribute to the growing literature on cross-cultural assessment using MMPI instruments. Ben-Porath and Sellbom emphasized that MMPI-3 development procedures incorporated extensive psychometric evaluation across gender groups to minimize systematic bias (18). The present findings generally support the robustness of the MMPI-3 framework while simultaneously demonstrating that gender-related measurement differences may still emerge when the instrument is applied in specific cultural and clinical contexts. Therefore, evidence obtained from North American standardization samples should not automatically be assumed to generalize fully to non-Western populations without empirical verification.

An additional contribution of this study is the integration of DTF analyses with structural comparisons across gender groups. The combination of significant DTF findings and gender differences in several PSY-5–SCL-90 associations suggests that both psychometric and substantive factors may contribute to the observed score differences. This finding highlights the importance of simultaneously evaluating measurement equivalence and construct relationships when examining gender differences in personality assessment. Failure to consider these issues may result in inaccurate comparisons and potentially biased clinical interpretations.

The present findings may also be considered alongside observations from the Federal Aviation Administration (FAA), which reported that the MMPI-2-RF, and by extension the MMPI-3, may demonstrate lower sensitivity than the MMPI-2 for identifying certain psychological conditions in Air Traffic Control Specialist applicants. Although those investigations were conducted in a different population and did not evaluate gender-related measurement invariance, they underscore the broader importance of examining psychometric performance across applied settings and demographic groups. In this context, the current findings extend existing evidence by demonstrating that gender-related DIF and DTF may also contribute to score interpretation within a clinical sample, emphasizing the need for ongoing evaluation of measurement equivalence when MMPI-3 scales are applied across diverse populations.

From a clinical perspective, the findings underscore the importance of considering measurement invariance when interpreting MMPI-3 PSY-5 scores across gender groups. Although the observed differences do not necessarily indicate true disparities in psychopathology, they suggest that gender-related response patterns may influence score interpretation. Consequently, clinicians and researchers should consider the potential impact of differential functioning when making diagnostic inferences, comparing group scores, or evaluating treatment outcomes. More broadly, the present findings highlight the value of incorporating DIF and DTF analyses into cross-cultural validation studies to better understand how psychological constructs and response processes may vary across demographic and cultural groups.

Limitation

This study has several limitations. First, the use of a convenience sampling method restricts the generalizability of the findings to the broader Iranian clinical population. Although the sample was recruited from multiple clinical centers and included diagnostically heterogeneous patients, future studies using probability-based sampling designs are needed to confirm the robustness of these findings. Additionally, the observed gender imbalance (59% women vs. 41% men) may have influenced the stability of DIF and DTF estimates, as unequal group sizes can affect the precision of parameter estimation and the detection of measurement non-invariance. Second, the reliance on self-report measures may introduce response biases such as social desirability or subjective reporting error. Third, the analyses did not consider DIF and DTF across diagnostic categories (e.g., depressive, bipolar, or psychotic spectrum disorders), which limits the ability to determine whether measurement non-invariance is consistent across clinical subgroups or is partly diagnosis-dependent. Fourth, the focus on only the five

PSY-5 scales limits the ability to draw conclusions regarding other dimensions of the MMPI-3. Fifth, the exclusion of the DISC scale due to multidimensionality, while methodologically necessary, also reflects a potential substantive limitation of the PSY-5 model in this cultural and clinical context, suggesting that certain constructs (e.g., disconstraint) may not function as a unidimensional latent trait in Iranian clinical samples. Finally, the cross-sectional design precludes causal inferences regarding the observed gender differences, and longitudinal studies are warranted to examine potential temporal variations in measurement functioning.

Conclusion

The present study provides evidence of systematic gender-related measurement non-equivalence in four MMPI-3 PSY-5 scales (AGGR, PSYC, NEGE, and INTR) within an Iranian clinical sample. The observed DTF indices (STDS = −0.27 to −0.37; UTDS = 0.28 to 0.43) indicate that the cumulative impact of item-level DIF may influence scale-score interpretation across genders. These findings underscore the importance of evaluating measurement invariance and applying gender-sensitive interpretation when using the MMPI-3 PSY-5 scales in clinical practice. The results further highlight the need for culturally validated psychometric assessment to support accurate and fair clinical decision-making. Future studies should replicate these findings using larger and more gender-balanced clinical samples, examine diagnostic subgroup differences, and extend DIF/DTF analyses to other MMPI-3 scales.

Acknowledgment

The authors thank the participating patients and clinical staff for their cooperation in data collection.

Funding

None.

Conflict of Interest

None.

Author's Contributions

Mehdi Rahmani Malek-Abad conceptualized and designed the study, conducted the statistical analyses, interpreted the data, drafted the manuscript, and participated in the critical revision of the manuscript. Mohammad Reza Abedi contributed to the conceptual development of the study, critically reviewed the manuscript, and participated in revising the manuscript for important intellectual content. Both authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

References

1. Whitman MR, Elias LS, Cappo BM, Ben-Porath YS. Criterion validity of MMPI-3 scores in preemployment evaluations of public safety candidates. Psychol Assess. 2021;33(12):1169-80.
2. Floyd AE, Gupta V. Minnesota Multiphasic Personality Inventory. [Updated 2023 Apr 24]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2026 Jan-. Available from:https://www.ncbi.nlm.nih.gov/books/NBK557525/
3. Reeves CK, Brown TA, Sellbom M. An examination of the MMPI-3 validity scales in detecting overreporting of psychological problems. Psychol Assess. 2022;34(6):517-27.
4. Haugh S, Ramirez D, Priebe K, Marek RJ, Anderson JL. Evaluating the MMPI-3's utility in the assessment of pathological personality traits. Curr Psychol. 2024;43(23):20521-31.
5. Brown JR. An Examination of the Construct Validity of the MMPI-3 PSY-5 Scales: Kent State University; 2025.
6. Delawalla CN, Lee TTC, Keen MA. The use of MMPI-3 scales to assess personality-based vulnerabilities for alcohol use and problems. Psychol Assess. 2023;35(8):633-45.
7. Sellbom M, Brown JR. Development and validation of an MMPI-3 Antagonism scale. Pers Disord. 2025;16(4):332–49.
8. Anestis JC, Rodriguez TR, O'Dell C, Preston OC, Harrop TM, Charles NE. Psychometric Properties of the MMPI-3 in a Sample of Black and White American Undergraduate Students: Examining Group Differences and Convergent/Discriminant Validity with the ASEBA Adult Self-Report. J Pers Assess. 2024;106(1):1-16.
9. Morris CS, Shepherd EA, Ingram PB. Investigating the Validity of the MMPI-3 Eating Concerns (EAT) scale across Men and Women in a University Sample. J Pers Assess. 2024;106(1):17-26.
10. Shumaker N, Long T, Torres A, Mercado A, Marek RJ, Anderson JL. Exploring Potential Ethnic Bias Among MMPI-3 Scales in Assessing Personality Psychopathology. Assessment. 2024;32(4):475–87.
11. Rahmani Malek Abad M, Falsafinejad MR, Minaei A, Farrokhi NA. Differential Item Functioning of Minnesota Multiphasic Personality Inventory-2 (Depression) Items in among Recruitment Candidates: A Comparison Between the Rasch Model and Mantel-Haenszel. Journal of Applied Psychological Research. 2024.
12. Rahmani Malek Abad M, Falsafinejad MR, Minaei A, Farrokhi NA. Differential item functioning of Psychasthenia and Psychopathic Deviate items in entrance employment tests: a comparison between the Mantel-Haenszel and Rasch model. Knowledge & Research in Applied Psychology. 2024;25(3):33-45.

13. Card NA, Stucky BD, Sawalani GM, Little TD. Direct and indirect aggression during childhood and adolescence: a meta-analytic review of gender differences, intercorrelations, and relations to maladjustment. Child Dev. 2008;79(5):1185-229.
14. Stark S, Chernyshenko OS, Drasgow F. Examining the effects of differential item (functioning and differential) test functioning on selection decisions: when are statistically significant effects practically important? J Appl Psychol. 2004;89(3):497-508.
15. Yavuz Temel G. A simulation and empirical study of differential test functioning (DTF). Psych. 2023;5(2):478-96.
16. Gori E. Rasch Analysis of some MMPI-2 scales in a sample of university freshmen. International Journal of Arts & Sciences. 2015;8(3):107.
17. Ben-Porath YS, Sellbom M. Interpreting the MMPI-3. Minneapolis (MN): University of Minnesota Press; 2023. 656 p.
18. Ben-Porath YS, Tellegen A. Minnesota multiphasic personality Inventory-3. PsycTESTS Dataset. 2020.
19. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. Jama. 2013;310(20):2191-4.
20. Ben-Porath YS, Tellegen A. Minnesota multiphasic personality Inventory-3. PsycTESTS Dataset. 2020.
21. Hall JT, Menton WH, Ben-Porath YS. Examining the Psychometric Equivalency of MMPI-3 Scale Scores Derived From the MMPI-3 and the MMPI-2-RF-EX. Assessment. 2022;29(4):842-53.
22. Rahmani Malek-Abad M. Reliability and validation of Minnesota Multiphasic Personality Inventory-3 adult form (MMPI-3). Knowl Res Appl Psychol. 2025;26(1):106-26.
23. Masal E, Koç M, Çolak TS, Takunyaci M. Analyzing the psychological symptoms of students in undergraduate program in elementary mathematics teaching. Educational Research and Reviews. 2013;8(14):1117.
24. Bond TG, Yan Z, Heene M. Applying the Rasch model: fundamental measurement in the human sciences. 4th ed. New York (NY): Routledge; 2020.
25. Linacre JM. A user's guide to Winsteps: Rasch-model computer programs. Program Manual. Version 5.10.2. Beaverton (OR): Winsteps.com; 2025.
26. Yen WM. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl Psychol Meas. 1984;8(2):125-45.
27. Christensen KB, Makransky G, Horton M. Critical Values for Yen's Q(3): Identification of Local Dependence in the Rasch Model Using Residual Correlations. Appl Psychol Meas. 2017;41(3):178-94.
28. Ibrahim A. Differential item functioning (DIF): theory and practice. Issues and practices in tests and measurement. 2024:357-68.
29. Cohen J. Set correlation and contingency tables. Appl Psychol Meas. 1988;12(4):425-34.
30. Ho AD, Yu CC. Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects. Educ Psychol Meas. 2015;75(3):365-88.
31. Hambleton RK, Slater SC. Item response theory models and testing practices: Current international status and future directions. European Journal of Psychological Assessment. 1997;13(1):21-8.
32. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory: Sage; 1991.
33. McNeish D, Wolf MG. Dynamic fit index cutoffs for confirmatory factor analysis models. Psychol Methods. 2023;28(1):61-88.
34. Kline RB. Principles and practice of structural equation modeling. 5th ed. New York (NY): Guilford Press; 2023. 494 p.
35. Pan F, Liu Q. Evaluating fit indices in a multilevel latent growth model with unbalanced design: a Monte Carlo study. Front Psychol. 2024;15:1366850.
36. Marsh HW, Hocevar D. Application of confirmatory factor analysis to the study of self-concept: First-and higher order factor models and their invariance across groups. Psychological bulletin. 1985;97(3):562.
37. Boone WJ, Staver JR. Advances in Rasch analyses in the human sciences: Springer; 2020.
38. McNeish D, Dumas D. Reliability representativeness: How well does coefficient alpha summarize reliability across the score distribution? Behav Res Methods. 2025;57(3):93.
39. Thomas ML, Locke DE. Psychometric properties of the MMPI-2-RF Somatic Complaints (RC1) scale. Psychol Assess. 2010;22(3):492-503.
40. Sijtsma K, Ellis JL, Borsboom D. Recognize the Value of the Sum Score, Psychometrics' Greatest Accomplishment. Psychometrika. 2024;89(1):84-117.
41. Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Appl Psychol Meas. 1995;19(4):353-68.
42. Meade AW. A taxonomy of effect size measures for the differential functioning of items and scales. J Appl Psychol. 2010;95(4):728-43.
43. Bujang MA. An elaboration on sample size determination for correlations based on effect sizes and confidence interval width: a guide for researchers. Restor Dent Endod. 2024;49(2):e21.
44. Meng X-L, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. Psychological bulletin. 1992;111(1):172.