

Threshold Fragility in MCMI-IV Base Rate Interpretations: a Retrospective Psychometric Analysis

Moein Esghaei

a_esghaei@sbu.ac.ir

Shahid Beheshti University

Fatemeh Ziaeetabar

University of Tehran

Research Article

Keywords: MCMI-IV, Base Rate scores, threshold fragility, psychometrics, psychological assessment, decision thresholds, classification consistency

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10962428/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Threshold Fragility in MCMI-IV Base Rate Interpretations: a Retrospective Psychometric Analysis

Moein Esghaei^{1,3*} and Fatemeh Ziaeetabar^{2,3}

¹Faculty of Computer Science and Engineering, Shahid Beheshti
University, Tehran, Iran.

²Department of Computer Science, School of Mathematics, Statistics
and Computer Science, College of Science, University of Tehran,
Tehran, Iran.

³Cognitive Neuroscience Laboratory, German Primate Center – Leibniz
Institute for Primate Research, Göttingen, Germany.

*Corresponding author(s). E-mail(s): a_esghaei@sbu.ac.ir;

Contributing authors: fziaeetabar@ut.ac.ir;

Abstract

Background: Psychiatric assessment frequently converts continuous measurements into categorical interpretations, yet the robustness of decisions made close to clinical cut-offs is rarely quantified. We examined this problem in the Millon Clinical Multiaxial Inventory–IV (MCMI-IV), asking whether threshold sensitivity at Base Rate (BR) 75 and BR 85 is a uniform property of the instrument or varies systematically across scales.

Methods: We retrospectively analyzed 686 protocol-valid adult MCMI-IV profiles comprising 17,150 decisions across 25 substantive scales. Threshold fragility was quantified using fixed-distance, measurement-error-adjusted, and perturbation-based approaches. We separately characterized fragility around BR 75 and BR 85, evaluated scale-ranking reproducibility across 1,000 random split-half analyses, and used an independently constructed nonlinear profile-space analysis to examine whether fragility reflected broader multivariate profile structure.

Results: Nearly one-third of all decisions fell within ± 3 BR points of an interpretive threshold, but fragility varied approximately eightfold across scales. The scale-specific hierarchy remained broadly consistent after adjustment for estimated measurement precision and across alternative perturbation magnitudes, and showed high split-half reproducibility (median Spearman $\rho = .967$). Fragility was also strongly threshold-dependent: approximately three-quarters of near-threshold decisions occurred around BR 75, although several scales showed substantial BR 85-specific sensitivity. Unadjusted fragility showed marked organization within multivariate profile space, but this association was no longer significant after accounting for overall BR elevation, arguing against an independent “fragile-profile” subtype.

Conclusions: Threshold fragility represents a systematic and reproducible property of MCMI-IV interpretation that is not conveyed by categorical labels alone. Quantifying proximity to BR 75 and BR 85 may provide a clinically useful uncertainty layer alongside conventional interpretation, particularly for scores close to decision boundaries. External replication and prospective clinical validation are needed to determine whether incorporating fragility information improves diagnostic judgment and decision-making.

Keywords: MCMI-IV, Base Rate scores, threshold fragility, psychometrics, psychological assessment, decision thresholds, classification consistency

Background

Clinical assessment repeatedly converts continuous measurements into categorical decisions. A laboratory value triggers treatment, a screening score prompts referral, and a psychological scale crosses from “not elevated” to “elevated.” These boundaries make complex information actionable, but they also create a basic vulnerability: two nearly identical measurements can receive different labels simply because they fall on opposite sides of a cut-off. This issue is especially relevant in mental-health assessment, where measured characteristics are typically continuous and observed scores contain measurement uncertainty. Psychometric theory has long distinguished the reliability of a continuous score from the consistency of the classification produced from that score; classification is least secure when observations lie close to the boundary that determines the label. At the same time, contemporary dimensional

models of psychopathology emphasize continuous variation beneath conventional categories, strengthening the need to characterize what is lost when continuous evidence is reduced to discrete interpretations [1–4].

The Millon Clinical Multiaxial Inventory–IV (MCMI-IV) provides a consequential setting in which to examine this problem. The MCMI-IV is a widely used self-report instrument for characterizing personality patterns and clinical syndromes in clinical populations. It reports Base Rate (BR) scores rather than conventional norm-referenced standard scores and organizes interpretation around established BR boundaries [5, 6]. Scores below BR 75, between BR 75 and 84, and at or above BR 85 are commonly assigned progressively stronger interpretive significance. Previous MCMI research has therefore concentrated on conventional psychometric questions, including internal consistency and test–retest reliability, correspondence with structured diagnostic interviews, convergent validity, and the sensitivity or specificity of particular cut-offs [7–10]. Those studies address whether a scale or threshold performs accurately or consistently, but not how robust an individual threshold-based interpretation is to a small change in the observed BR score. More broadly, the Millon base-rate transformation has prompted discussion about how prevalence, scaling, and cut-score interpretation interact [11].

Although examined here using the MCMI-IV, the underlying methodological issue extends beyond a single instrument. Psychiatric and psychological assessments frequently map continuous measurements onto categorical decision boundaries, yet observations lying close to those boundaries may carry greater classification uncertainty than observations farther away. Quantifying this local decision uncertainty may therefore complement conventional categorical interpretation without replacing the underlying clinical assessment.

Here we address this missing layer of uncertainty by defining threshold fragility as the susceptibility of a test-based interpretive category to change because an observed

BR score lies close to BR 75 or BR 85. We do not test whether the established thresholds should be replaced. Instead, we ask how often, and on which scales, conventional MCMI-IV interpretations occur close enough to those thresholds that a small score change could alter the assigned category. We map this property across all 25 substantive scales in 686 protocol-valid adult profiles using complementary fixed-distance, measurement-error-adjusted, and perturbation-based analyses, and we test the reproducibility of the resulting scale hierarchy across repeated subsamples. A supporting nonlinear profile-space analysis then asks whether fragility is organized within independently derived multivariate MCMI-IV profile structure or merely reflects global score elevation.

The primary objective was therefore to quantify the local robustness of BR 75 and BR 85 interpretations across the complete set of substantive MCMI-IV scales. We prespecified a narrow fixed-distance definition as the primary descriptive measure and used measurement-error-adjusted, perturbation-based, and repeated split-half analyses to examine whether the resulting scale hierarchy depended on a single operational definition or sample partition. A secondary profile-space analysis was used to determine whether profile-level fragility showed nonrandom organization within multivariate MCMI-IV score structure and whether any such organization remained after accounting for overall BR elevation. The study was designed to characterize uncertainty around established thresholds, not to redefine diagnosis, replace BR 75 or BR 85, or infer temporal instability of assessed individuals.

Methods

This retrospective observational psychometric analysis examined the local robustness of threshold-based MCMI-IV Base Rate interpretations. Reporting was guided by the STROBE recommendations for observational studies [12]. The primary focus was not whether BR 75 and BR 85 are clinically useful cut-offs, but whether decisions made

114 near these cut-offs have the same local robustness as decisions based on scores farther
115 from them. Analyses were organized in three stages. First, threshold fragility was quan-
116 tified at the decision and scale levels using fixed-margin, measurement-error-adjusted,
117 and perturbation-based indices, with the primary fixed-margin analysis decomposed
118 by BR 75 and BR 85. Second, reproducibility of the primary scale ranking was eval-
119 uated using repeated split-half resampling. Third, a supporting whole-profile analysis
120 evaluated whether threshold fragility showed systematic organization within multi-
121 variate MCMI-IV profile space. Throughout, fragility was treated as a property of the
122 test-based threshold decision rather than as temporal instability of the respondent,
123 symptoms, or clinical state.

124 **Data source and analytic sample**

125 The analytic file consisted of previously collected, anonymized assessment-level MCMI-
126 IV profile outputs. The original dataset contained 853 assessment records and included
127 BR scores, raw scores, demographic information, and response-style, validity, and
128 modifying indices. Because the study focused on threshold-based interpretation, the
129 primary analyses used BR scores for the substantive MCMI-IV scales. Before the
130 analytic dataset was provided to the research team, direct personal identifiers were
131 removed. The investigators therefore analyzed only pre-existing anonymized assess-
132 ment records and had no access to directly identifiable personal information. No
133 additional participant contact, intervention, or collection of new participant data
134 occurred as part of the present study.

135 No a priori sample-size calculation was performed because the study was a
136 retrospective analysis of all eligible assessment records available in the source dataset.

137 Profiles were excluded if recorded age was younger than 18 years, if the Invalidity
138 score was positive ($V > 0$), if the Inconsistency score was $W \geq 20$, or if the raw
139 Disclosure score was outside the valid range ($X_{\text{raw}} < 7$ or $X_{\text{raw}} > 114$). The age

140 criterion was met by 46 records and the Invalidity criterion by 47 records, with four
141 records meeting both; these criteria excluded 89 unique records. Across the complete
142 dataset, 85 records met the Inconsistency criterion, including seven already excluded
143 by age and/or Invalidity; 78 additional records were therefore excluded. No record met
144 the Disclosure exclusion criterion. After accounting for overlap, 167 unique records
145 were excluded, yielding 686 adult profiles meeting all eligibility and protocol-validity
146 criteria [5].

147 No missing values were identified across the 25 substantive BR variables, and no
148 records were excluded because of incomplete BR data. A comparison across all avail-
149 able variables identified no exact duplicate records. Because the analytic file did not
150 contain a unique individual identifier, repeated assessments originating from the same
151 individual could not be identified. Each row was therefore treated as one assessment
152 record, and all analyses were conducted at the assessment-record level.

153 The mean age in the final sample was 34.1 years ($SD = 10.2$; range = 18–75),
154 and 476 retained profiles were recorded as female in the source file. The recorded
155 sex variable was used descriptively only; sex-related differences were not an objective
156 of the present analysis. With 686 profiles assessed across 25 substantive scales, the
157 analytic dataset contained 17,150 profile-by-scale threshold decisions.

158 **Measure: MCMI-IV Base Rate profiles**

159 The MCMI-IV reports BR scores for clinical personality patterns, severe personality
160 pathology, clinical syndromes, and severe clinical syndromes. BR scores are interpreted
161 relative to established clinical thresholds rather than as conventional norm-referenced
162 standard scores [5]. The profiles analyzed here were generated using the Persian-
163 language version of the MCMI-IV. The primary analyses included 25 substantive
164 scales: 12 clinical personality pattern scales, three severe personality pathology scales,
165 seven clinical syndrome scales, and three severe clinical syndrome scales. Disclosure

166 (X), Desirability (Y), Debasement (Z), Validity (V), and Inconsistency (W) were
167 excluded as substantive variables because they index response style or protocol valid-
168 ity. Complete scale names, codes, and analytic variable names are provided in Table
169 S1 of Additional file 1.

170 **Base-Rate threshold classification**

171 For profile i and substantive scale j , the observed BR score x_{ij} was assigned to one of
172 three test-based interpretive categories: non-elevated for $x_{ij} < 75$, elevated for $75 \leq x_{ij}$
173 < 85 , and prominent for $x_{ij} \geq 85$. These operational categories describe the MCMI-IV
174 threshold rule used in the analysis; they are not independent clinical diagnoses.

175 **Fixed-margin threshold fragility**

176 The primary analysis quantified proximity to the nearest interpretive threshold. For
177 each profile-by-scale decision, distance to the nearest boundary was $d_{ij} = \min(|x_{ij} -$
178 $75|, |x_{ij} - 85|)$. A decision was defined as fragile under margin m when $d_{ij} \leq m$.
179 The primary margin was ± 3 BR points, selected as a narrow, clinically transparent
180 sensitivity band. Broader ± 5 and ± 7 bands were evaluated as nested sensitivity anal-
181 yses. Scale-level fragility was the proportion of the 686 profiles meeting the specified
182 margin for each scale. Because the margins are nested, the ± 5 and ± 7 estimates are
183 progressively broader sensitivity bands rather than competing definitions.

184 **Threshold-specific decomposition**

185 To determine which boundary contributed to primary fragility, the ± 3 criterion
186 was decomposed separately around BR 75 and BR 85. For threshold $t \in \{75, 85\}$,
187 threshold-specific fragility was $F_{ij,t} = \mathbb{I}(|x_{ij} - t| \leq 3)$, and the scale-level value was
188 the average of this indicator across profiles. The BR 75 and BR 85 ± 3 regions do not
189 overlap, so their counts sum to the primary nearest-threshold count.

190 Measurement-error-adjusted fragility

191 Fixed BR margins do not account for scale-specific measurement precision. A com-
192plementary analysis therefore expressed threshold sensitivity relative to the estimated
193standard error of measurement (SEM). Because item-level responses were unavailable,
194scale-specific Cronbach’s alpha coefficients were obtained from the Persian MCMI-IV
195validation study [9] and combined with the observed BR-score standard deviation s_j
196in the present sample: $SEM_j = s_j \sqrt{1 - \alpha_j}$. The complete reliability inputs and SEM
197values are reported in Table S2 of Additional file 1.

198 Measurement error was modeled as a zero-mean normal variable with standard
199deviation SEM_j . For each observed score, the probability that such error would move
200the score into a different BR category was calculated analytically from the normal
201cumulative distribution function, taking account of the relevant one-sided or two-sided
202category boundary. Scale-level SEM-adjusted fragility was the mean category-change
203probability across profiles. This analysis was treated as a complementary psychometric
204sensitivity analysis rather than as a replacement for the fixed-margin atlas, because
205the reliability coefficients were obtained from an external validation sample.

206 Perturbation-based label volatility

207 A second complementary analysis evaluated category sensitivity under a common per-
208turbation magnitude applied equally to all substantive scales. For each observed BR
209score, $x_{ij}^* = x_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma^2)$. The primary perturbation was $\sigma = 5$ BR
210points. Decision-level volatility V_{ij} was the analytic probability that the perturbed
211score would fall into a category different from that of the observed score. For scores
212below 75 or at/above 85, this is the probability of crossing the nearest category bound-
213ary in the appropriate direction; for scores in the 75–84 interval, it is the sum of the
214probabilities of crossing below 75 or at/above 85. Scale-level volatility V_j was the mean
215of V_{ij} across profiles.

216 To assess dependence on perturbation strength, the calculation was repeated for σ
217 $= 2, 3, 5, 7$, and 10 BR points. The 25-scale ranking at each alternative σ was compared
218 with the primary $\sigma = 5$ ranking using Spearman’s rank correlation. The common σ
219 analysis should not be interpreted as scale-specific measurement error or temporal
220 symptom fluctuation; it is a controlled perturbation model used to test whether the
221 fragility hierarchy depends on the exact width of the perturbation.

222 **Split-half reproducibility**

223 Repeated split-half analyses assessed whether the primary ± 3 scale-level fragility
224 ranking was stable across different compositions of the analytic sample. In each of
225 1,000 iterations, the 686 profiles were randomly divided without replacement into two
226 non-overlapping halves of 343 profiles. Scale-level fixed ± 3 fragility was recomputed
227 independently in each half. Agreement between the two complete 25-scale rankings was
228 quantified using Spearman’s rank correlation. Top-five and top-ten set overlap were
229 also recorded, together with exact replication of the top-five set and replication of at
230 least four of the five leading scales. Distributions were summarized by the median and
231 5th and 95th percentiles. The random-number generator was initialized with seed 42.

232 **Profile-space analysis using Diffusion Maps and Mapper**

233 A supporting whole-profile analysis asked whether profiles with greater threshold
234 fragility were spatially organized within multivariate MCMI-IV profile structure
235 derived independently of fragility. Geometry construction used the same 686 profiles
236 and 25 substantive scales. Each scale was standardized across profiles only for mul-
237 tivariate distance calculations: $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$. Threshold fragility continued to
238 be calculated from the original BR values. Profile-level fragility burden was defined
239 as the proportion of the 25 substantive scales whose original BR score lay within ± 3
240 points of BR 75 or BR 85. This quantity was not used to construct the geometry and
241 was introduced only after the profile topology had been defined.

Nonlinear geometry was estimated using Diffusion Maps [13]. Pairwise Euclidean distances were calculated in the standardized 25-dimensional profile space. For the primary analysis, a $k = 20$ nearest-neighbor graph was constructed and symmetrized across profiles. For retained neighborhood edges, Gaussian similarity weights were defined as $W_{ij} = \exp(-d_{ij}^2/\epsilon)$, where ϵ was the median squared Euclidean distance across retained neighborhood edges. Coifman–Lafon density normalization was applied with $\alpha = 1$ before normalization to a Markov transition operator. The stationary first mode was excluded, and the first two nontrivial diffusion coordinates were retained at diffusion time $t = 1$ as the nonlinear lens.

Mapper was applied to the two-dimensional diffusion lens [14]. The lens was covered by 10 intervals along each dimension with 35% overlap. Within each nonempty cover region, local clustering was performed in the complete standardized 25-dimensional profile space using DBSCAN [15], with `min.samples=3`; the DBSCAN neighborhood radius was defined as the 90th percentile of the five-nearest-neighbor distance distribution in the standardized 25-dimensional space. Each local cluster became a Mapper node; nodes were connected when they shared at least one assessment through the overlapping cover. Node size reflected local membership. Only after graph construction was node color assigned from the mean profile-level fixed ± 3 fragility burden.

Graph-level concentration of fragility was quantified using Moran’s I . Let y_r denote mean profile-level fragility for Mapper node r , and let $w_{rs} = 1$ when nodes r and s were connected by an edge and $w_{rs} = 0$ otherwise. With M nodes and $W = \sum_r \sum_s w_{rs}$, Moran’s I was calculated as

$$I = \frac{M}{W} \frac{\sum_r \sum_s w_{rs} (y_r - \bar{y})(y_s - \bar{y})}{\sum_r (y_r - \bar{y})^2}. \quad (1)$$

Binary adjacency weights were used without row standardization. Under this graph-weighting convention, Moran’s I is not constrained to the interval $[-1, 1]$, so values

slightly above 1 are possible. Statistical significance was evaluated using 5,000 assessment-record-level permutations in which profile fragility values were randomly reassigned while graph topology and node memberships remained fixed. The corrected permutation probability was $p = (1 + \#\{I_b \geq I_{obs}\})/(B + 1)$ with $B = 5000$.

Because globally elevated profiles provide more opportunities to lie near BR 75 or BR 85, profile fragility burden was additionally residualized using ordinary least-squares regression with mean BR, the number of scores in the BR 75–84 range, and the number of scores at or above BR 85 as predictors. The regression was used only to remove the linear contribution of these elevation measures; individual regression coefficients were not interpreted as substantive effects. Moran’s I and the same 5,000-permutation procedure were then repeated on the elevation-adjusted residual fragility values while retaining the original topology and node memberships.

Parameter sensitivity of the unadjusted profile-space result was evaluated across 15 Diffusion-Mapper configurations formed by diffusion-neighborhood sizes $k = 10, 15, 20, 30$, and 40 and Mapper resolutions of $8, 10$, or 12 intervals per diffusion dimension. Cover overlap remained 35% , and the local clustering procedure was otherwise unchanged. This analysis was supporting rather than primary; its complete display is reported in Figure S2 of Additional file 1.

Statistical analysis and software

Descriptive results are reported as counts, percentages, means, standard deviations, medians, and percentile intervals as appropriate. Spearman rank correlations quantified agreement between alternative scale-level rankings and between rankings across perturbation magnitudes. Repeated split-half reproducibility was summarized from 1,000 random partitions. Profile-space significance was evaluated with 5,000 assessment-record-level permutations. The primary scale atlas was descriptive; no pairwise inferential comparisons among the 25 scale-specific proportions were performed,

and therefore no multiple-comparison correction was applied to the descriptive scale ranking. All analyses used the assessment record as the analytic unit because unique individual identifiers were unavailable. If repeated administrations from the same individual were present, this working independence assumption could overstate nominal resampling or permutation precision; this limitation is addressed explicitly in the Discussion.

The analyses were conducted in Python 3.13.5 using NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0, NetworkX 3.6.1, and Matplotlib 3.10.8. scikit-learn was used for nearest-neighbor calculations and DBSCAN; NetworkX was used for graph representation and adjacency operations; NumPy, pandas, and SciPy supported numerical and statistical calculations; and Matplotlib was used for visualization. The split-half procedure used random seed 42.

Use of generative AI during manuscript preparation

ChatGPT (OpenAI) was used during manuscript preparation for language refinement, structural drafting, organization of methodological descriptions, and assistance in drafting and debugging analysis code. All analytic decisions, code, numerical outputs, statistical results, interpretations, and references were reviewed and verified by the authors, who retain full responsibility for the work. The tool was not listed as an author.

Table 1 Sample characteristics and analytic dataset summary

Characteristic	Value
Original dataset	853 profiles
Final analytic sample	686 profiles
Age, years	$M = 34.1$, $SD = 10.2$ (range 18–75)
Female	476
Analyzed substantive MCMI-IV BR scales	25
Profile-by-scale decisions	17 150
Eligibility criteria	$\text{Age} \geq 18$, $V = 0$, $W < 20$, $7 \leq X_{raw} \leq 114$

Results

Analytic sample and Base-Rate decision distribution

Application of the age and protocol-validity criteria yielded 686 eligible adult profiles from the original 853 assessment records. The mean age was 34.1 years ($SD = 10.2$; range = 18–75), and 476 profiles were recorded as female. No missing values were present across the 25 substantive BR scales. The analytic dataset therefore comprised 17,150 profile-by-scale threshold decisions (Table 1).

Using BR 75 and BR 85 as the decision boundaries, 10,752 decisions (62.7%) were non-elevated, 4,012 (23.4%) were elevated, and 2,386 (13.9%) were prominent. Thus, 6,398 decisions (37.3%) were at or above BR 75, including 2,386 at or above BR 85. These frequencies describe test-based interpretive categories and should not be interpreted as counts of independent clinical diagnoses.

Fixed-margin fragility atlas

Near-threshold decisions were common. Under the primary ± 3 BR-point criterion, 5,062 of 17,150 decisions (29.5%) fell within three points of either BR 75 or BR 85. This increased to 7,238 decisions (42.2%) under the ± 5 criterion and 8,364 (48.8%) under the ± 7 criterion (Figure 1A). Because the margins were nested, the increase reflects progressively broader sensitivity bands.

Scale-level fragility differed markedly across the 25 substantive scales (Figure 1B; Table 2). Under the primary ± 3 criterion, the highest values were observed for Melancholic (53.2%), Somatic Symptom (47.8%), Borderline (44.9%), Sadistic (43.9%), and Negativistic (43.6%). The lowest values were observed for Drug Use (6.3%), Turbulent (9.3%), Compulsive (10.1%), Alcohol Use (11.2%), and Histrionic (12.1%). The ratio between the highest and lowest values was approximately 8.4, indicating more than an eightfold span in the frequency of near-threshold decisions across scales.

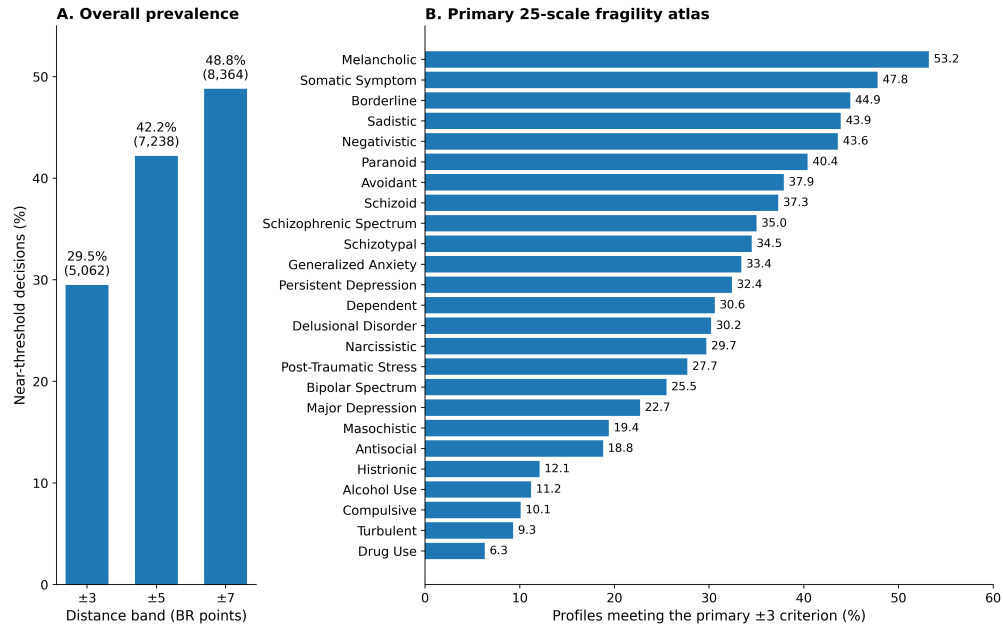


Fig. 1 The MCMI-IV fragility atlas. (A) Overall proportion of the 17,150 profile-by-scale decisions lying within ± 3 , ± 5 , and ± 7 BR points of BR 75 or BR 85. The margins are nested sensitivity bands. (B) Primary ± 3 scale-level fragility across the 25 substantive scales, ordered from highest to lowest. Higher percentages indicate that a scale more frequently produced threshold-sensitive interpretations; they do not indicate poorer scale validity or a greater probability of clinical diagnosis. BR = Base Rate.

Threshold-specific and SEM-adjusted characterization

Across all 17,150 decisions, 3,781 (22.0%) fell within ± 3 BR points of BR 75 and 1,281 (7.5%) fell within ± 3 BR points of BR 85. These non-overlapping sets together accounted for all 5,062 primary fragile decisions; 74.7% of primary near-threshold decisions were therefore associated with BR 75 and 25.3% with BR 85.

The relative contribution of the boundaries differed across scales (Figure 2A). BR 75 fragility was highest for Melancholic (43.0%), Sadistic (37.5%), Borderline (35.6%), Paranoid (34.8%), and Schizoid (34.5%). BR 85 fragility was highest for Generalized Anxiety (24.1%), Somatic Symptom (18.7%), Persistent Depression (14.1%), Major Depression (13.7%), and Post-Traumatic Stress (13.6%). Generalized Anxiety and Major Depression showed greater fragility around BR 85 than BR 75, while Post-Traumatic Stress was approximately balanced. Similar overall fragility could therefore

Table 2 Scale-level fixed-margin threshold fragility across the 25 substantive MCMI-IV scales

Scale	Profiles near BR 75 or BR 85 (%)		
	± 3	± 5	± 7
Melancholic	53.2	69.2	72.6
Somatic Symptom	47.8	60.8	63.8
Borderline	44.9	63.4	69.5
Sadistic	43.9	65.3	72.9
Negativistic	43.6	71.1	75.4
Paranoid	40.4	59.2	66.9
Avoidant	37.9	51.9	55.8
Schizoid	37.3	56.6	64.4
Schizophrenic Spectrum	35.0	49.6	59.2
Schizotypal	34.5	47.5	56.6
Generalized Anxiety	33.4	44.5	51.9
Persistent Depression	32.4	45.2	50.7
Dependent	30.6	42.7	47.5
Delusional Disorder	30.2	48.3	57.1
Narcissistic	29.7	42.4	49.7
Post-Traumatic Stress	27.7	40.8	45.3
Bipolar Spectrum	25.5	36.0	45.8
Major Depression	22.7	32.2	39.9
Masochistic	19.4	31.3	44.0
Antisocial	18.8	29.0	38.2
Histrionic	12.1	16.6	21.3
Alcohol Use	11.2	13.0	18.1
Compulsive	10.1	15.2	20.4
Turbulent	9.3	14.7	21.0
Drug Use	6.3	8.6	11.1

Note. Entries are percentages of the 686 profiles whose BR score fell within the indicated distance of either BR 75 or BR 85. Scales are ordered by the primary ± 3 criterion. The margins are cumulative sensitivity bands.

348 arise from distinct boundary-specific patterns: Schizotypal had 34.5% overall fragility
349 driven almost entirely by BR 75 (33.4% vs 1.2%), whereas Generalized Anxiety had
350 33.4% overall fragility but was driven predominantly by BR 85 (9.3% vs 24.1%).

351 Scale-specific SEM values ranged from 6.10 BR points for Melancholic to 22.62
352 for Alcohol Use. Across all decisions, the mean estimated category-change probabil-
353 ity under the measurement-error model was 24.9%. Scale-level SEM-adjusted fragility
354 ranged from 10.7% to 38.2% (Table S3 of Additional file 1). Somatic Symptom
355 (38.2%), Negativistic (35.3%), Dependent (30.1%), Borderline (30.0%), and Schizoid

(29.5%) had the highest values, whereas Turbulent (10.7%), Compulsive (13.6%),
Histrionic (14.6%), Masochistic (16.4%), and Drug Use (16.4%) had the lowest. The
SEM-adjusted ranking retained substantial correspondence with the primary fixed
 ± 3 ranking (Spearman $\rho = .779$), but selected scales shifted materially: Melancholic
moved from first under the fixed-margin criterion to seventh after SEM adjustment,
whereas Alcohol Use moved upward because of its comparatively large estimated SEM
(Figure 2B).

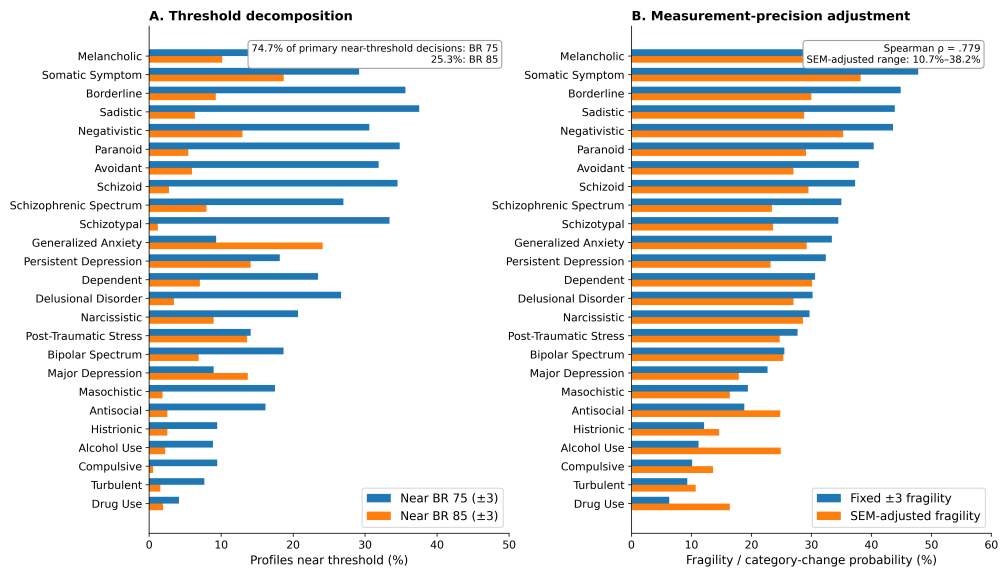


Fig. 2 Characterizing the sources of threshold fragility. (A) Threshold-specific decomposition of the primary ± 3 fragility measure into BR 75- and BR 85-specific components. Across all primary near-threshold decisions, 74.7% were associated with BR 75 and 25.3% with BR 85. (B) Primary fixed ± 3 fragility compared with SEM-adjusted category-change probability in the same scale order. The broad scale pattern persisted after accounting for estimated measurement precision (Spearman $\rho = .779$), although selected ranks shifted. BR = Base Rate; SEM = standard error of measurement.

Perturbation and reproducibility analyses

Under the primary Gaussian perturbation model ($\sigma = 5$ BR points), mean decision-level volatility was 16.1%. Scale-level volatility ranged from 4.0% to 26.6% (Table S4 of Additional file 1). The highest values were observed for Negativistic (26.6%),

367 Melancholic (25.7%), Sadistic (24.4%), Borderline (23.3%), and Somatic Symptom
368 (23.2%); the lowest were Drug Use (4.0%), Turbulent (5.9%), Alcohol Use (6.2%),
369 Compulsive (6.6%), and Histrionic (6.8%). The perturbation-based ranking closely
370 corresponded to the fixed ± 3 ranking (Spearman $\rho = .983$; Figure 3A). This close
371 agreement is a robustness check rather than an independent discovery because both
372 measures depend on distance from the same thresholds.

373 Absolute volatility increased as perturbation magnitude increased: mean decision-
374 level values were 6.9% at $\sigma = 2$, 10.1% at $\sigma = 3$, 16.1% at $\sigma = 5$, 21.4% at $\sigma = 7$, and
375 27.6% at $\sigma = 10$. Despite this change in absolute probability, scale ordering remained
376 highly stable relative to $\sigma = 5$, with Spearman correlations of .958, .987, .995, and
377 .988 for $\sigma = 2, 3, 7$, and 10, respectively (Figure 3B).

378 The repeated split-half analysis further showed strong sample-level reproducibility.
379 Across 1,000 random 343-versus-343 splits, the median Spearman correlation between
380 the two 25-scale rankings was .967, with a 5th–95th percentile interval of .940–.985
381 (Figure 3C; Table 3). Median top-five overlap was 5 of 5 scales (5th–95th percentile
382 4–5), and the exact top-five set was reproduced in 64.8% of splits. At least four of the
383 five leading scales were shared in 99.4% of splits. Median top-ten overlap was 9 of 10
384 scales (5th–95th percentile 8–10).

Table 3 Repeated split-half reproducibility of the primary scale-level fragility ranking

Reproducibility statistic	Result
Spearman rank correlation, median [5th, 95th percentile]	.967 [.940, .985]
Top-five overlap, median [5th, 95th percentile]	5 [4, 5]
Top-ten overlap, median [5th, 95th percentile]	9 [8, 10]
Exact replication of the top-five set	64.8%
At least four of the five leading scales shared	99.4%

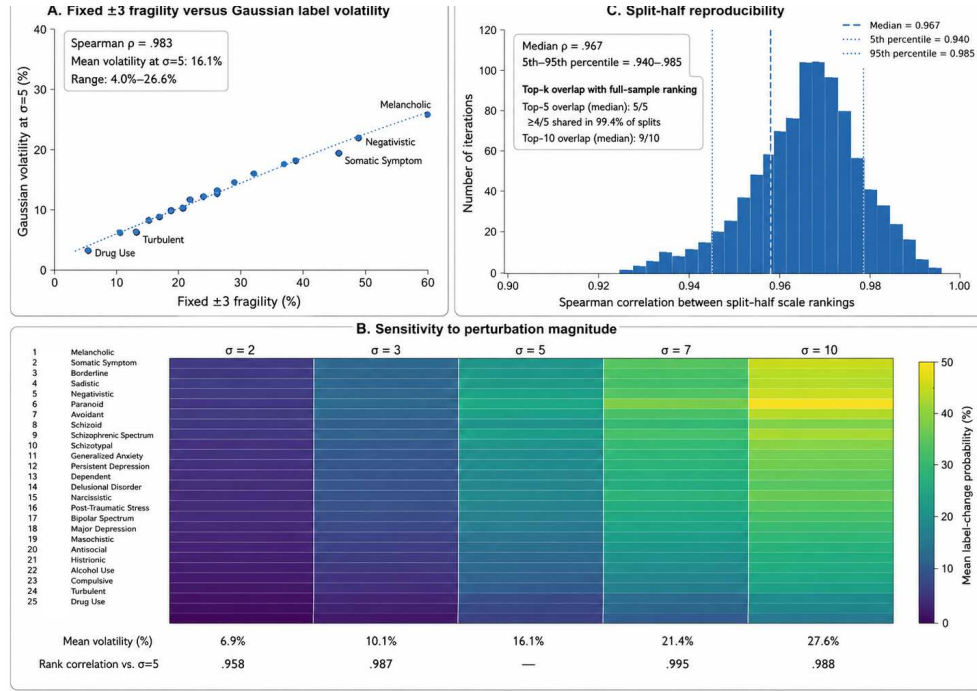


Fig. 3 Robustness and reproducibility of the fragility atlas. (A) Primary fixed ± 3 fragility versus Gaussian label volatility at $\sigma = 5$ BR points ($\rho = .983$). (B) Scale-level Gaussian label volatility across $\sigma = 2, 3, 5, 7$, and 10 BR points. Absolute volatility increased with perturbation magnitude while rank correlations relative to $\sigma = 5$ remained high. (C) Distribution of Spearman correlations from 1,000 random split-half comparisons; median $\rho = .967$ (5th–95th percentile .940–.985).

Supporting profile-space analysis

The independent Diffusion Map and Mapper analysis represented 96.6% of the 686-profile cohort in a primary graph containing 18 nodes and 23 edges (Figure 4A–B). Node mean fixed ± 3 fragility burden ranged from 2.7% to 34.2%. In the unadjusted analysis, fragility was strongly spatially concentrated across the graph (Moran's $I = 1.017$). None of 5,000 assessment-record-level permutations produced a statistic equal to or greater than the observed value, yielding the corrected permutation probability $p = 1/5001 = .00020$ (Figure 4C). The positive unadjusted concentration persisted across all 15 tested Diffusion-Mapper configurations (Moran's I range .650–1.027; median .894; Figure S2 of Additional file 1).

395 However, profile-level fragility was also strongly related to global BR elevation.
396 Mean BR, the number of scores in the BR 75–84 range, and the number of scores
397 at or above BR 85 jointly explained 61.6% of profile-level fragility variance. After
398 removing their linear contribution, Moran’s I decreased to .274; 1,523 of 5,000 permu-
399 tations equaled or exceeded the adjusted statistic, giving $p = .305$ (Figure 4D). The
400 residual spatial concentration was therefore not statistically significant after elevation
401 adjustment.

402 The unadjusted analysis showed that fragility was nonrandomly distributed across
403 whole-profile structure; however, elevation adjustment indicated that this organization
404 was largely attributable to global BR elevation rather than to an independent fragility-
405 related topology.

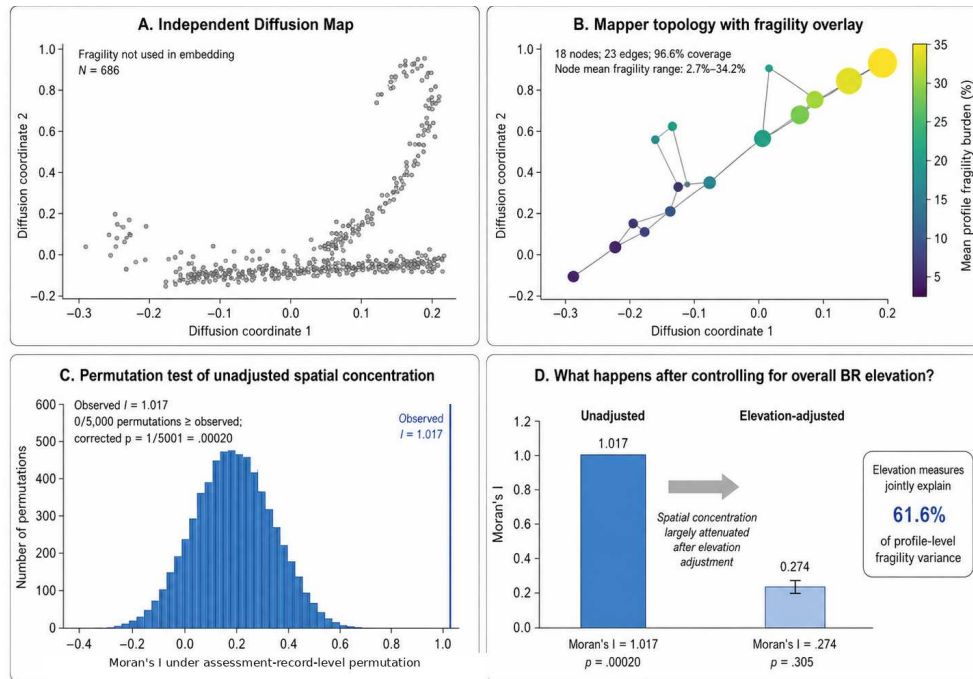


Fig. 4 Diffusion-Mapper profile-space analysis of threshold fragility. (A) Independent Diffusion Map of the 686 complete profiles using 25 standardized substantive BR scales; fragility was not used to construct the embedding. (B) Mapper topology of connected local profile neighborhoods; node size reflects membership and node color represents mean profile-level fixed ± 3 fragility burden. The primary graph contained 18 nodes and 23 edges and covered 96.6% of the cohort. (C) Assessment-record-level permutation distribution for the unadjusted graph. Observed Moran's $I = 1.017$; none of 5,000 permutations equaled or exceeded the observed value (corrected $p = .00020$). (D) After adjustment for mean BR and the numbers of BR 75–84 and BR ≥ 85 scores, Moran's I decreased to .274 ($p = .305$); elevation measures jointly explained 61.6% of profile-level fragility variance. These results do not support an independent fragile-profile or diagnostic subtype.

Discussion

The central finding is that threshold fragility is a systematic, strongly scale-dependent, and highly reproducible property of MCMI-IV interpretation rather than a uniform consequence of using BR 75 and BR 85. The same nominal BR boundaries are applied across scales, yet our findings show that the local robustness of classifications around those boundaries is not equivalent. Near-boundary scores occurred frequently on some scales and rarely on others, producing more than an eightfold span between the most and least fragile scales under the primary definition. The important object is therefore

not only the numerical location of BR 75 or BR 85, but also the local density of observed scores around each boundary. A threshold positioned in a densely populated region of a scale’s score distribution will generate many decisions that could change after a small perturbation, whereas the same nominal threshold in a sparse region will generate comparatively robust decisions. The fragility atlas makes this otherwise invisible property explicit.

This perspective complements rather than competes with conventional psychometric validation. Prior evaluation of the Persian MCMI-IV reported heterogeneity in internal consistency and test–retest reliability across scales [9], while subsequent comparison with diagnostic criteria showed marked variation in diagnostic performance across syndromes [10]. Those studies concern whether scales measure consistently and whether classifications correspond to external clinical criteria. Fragility asks a third question: given the observed score and the established threshold rule, how locally stable is the resulting interpretive category? These quantities need not coincide. A psychometrically reliable scale can produce many fragile classifications if its empirical score distribution is concentrated around a cut-off, while a less precise scale may produce relatively few near-boundary observations if most scores lie far from the cut-off. General classification theory predicts this dependence on both measurement error and cut-score proximity [1, 2, 16]. Our SEM-adjusted analysis supported that distinction: accounting for scale-specific estimated measurement precision changed the relative position of selected scales but retained substantial correspondence with the simple distance-based atlas.

Separating BR 75 from BR 85 further showed why a single aggregate fragility score can conceal meaningful structure. Most near-threshold decisions accumulated around BR 75, but this dominance was not universal. Scales with very similar total fragility could have almost opposite boundary profiles, with one scale dominated by the lower interpretive boundary and another by BR 85. Earlier MCMI-III studies evaluated BR

441 75 and BR 85 chiefly as candidate cut-offs for diagnostic or screening efficiency and
442 reported scale-dependent performance [7, 8, 17]. Grove and Vrieze also questioned
443 aspects of the Millon base-rate transformation, emphasizing that prevalence, scaling,
444 and cut-score interpretation should be considered explicitly [11]. Those MCMI-III find-
445 ings are not direct evidence about the MCMI-IV, which was substantially revised, but
446 they provide an important context: the behavior of a nominal cut-off is not necessarily
447 identical across scales. The present findings extend that issue by showing that, before
448 external diagnostic accuracy is considered, the empirical distributions of MCMI-IV
449 scores expose different proportions of assessments to boundary-sensitive interpretation.

450 The perturbation and resampling analyses strengthen the descriptive atlas without
451 changing its meaning. The near-identity of the fixed-margin and Gaussian rankings
452 shows that the scale hierarchy is not an artifact of converting threshold distance into
453 a binary ± 3 indicator. The stability of the ranking across $\sigma = 2$ –10 BR points like-
454 wise indicates that the relative pattern is preserved under substantially weaker and
455 stronger perturbations even though the absolute probability of label change increases,
456 as expected. Repeated split-half sampling provides a different form of evidence: the
457 same leading scales reappeared across different subsets of the cohort, with a median
458 full-ranking correlation of .967 and at least four of the top five scales shared in 99.4%
459 of iterations. These results support the interpretation of the atlas as a stable property
460 of the observed sample distributions rather than a ranking driven by a small subset
461 of records.

462 The profile-space analysis provides a useful qualification. In the unadjusted
463 Diffusion-Mapper representation, high- and low-fragility profiles were not randomly
464 interspersed; they showed strong graph-level concentration. Because the geometry was
465 built from standardized BR profiles without using fragility, this initial pattern was
466 not a trivial consequence of embedding the outcome into the topology. However, the
467 concentration was no longer statistically significant after accounting for mean BR and

the numbers of elevated and prominent scores, and these elevation variables explained 61.6% of profile-level fragility variance. The appropriate interpretation is therefore not that the data contain a distinct “fragile-profile” subtype. Rather, global elevation creates more opportunities for a profile to approach one or both interpretive thresholds, and this broad elevation structure largely explains the apparent profile-space organization of fragility. This negative adjusted result is informative because it prevents a stronger topological claim than the data support.

The practical implication is modest but potentially useful. A conventional categorical label conveys where a score lies relative to a threshold, but not how far it lies from that threshold. A BR score of 76 and one of 83 are both “elevated” under the same category rule, although their vulnerability to a small downward or upward score change is different. Reporting threshold proximity could therefore serve as an uncertainty annotation rather than as a new diagnostic rule. These findings do not establish that reporting threshold fragility improves clinical outcomes, diagnostic accuracy, or clinical decision-making. Rather, they suggest that proximity to an interpretive boundary may provide an additional uncertainty descriptor that could be evaluated prospectively alongside conventional MCMI-IV interpretation.

Several limitations define the scope of the claim. The exact scale-specific fragility ranking is partly distribution-dependent, because the frequency of observations near BR 75 and BR 85 depends on the BR-score distribution of the population being assessed. Moreover, the present data were obtained using the Persian-language MCMI-IV in a single regional context. Replication in independent clinical populations, other language versions, and different service settings is therefore necessary before assuming that the exact scale ranking is universal. The absence of unique individual identifiers prevented verification that all assessment records represented distinct individuals. If some records reflected repeated assessments from the same person,

the effective number of independent observations would be smaller than the number of assessment records analyzed. The findings should therefore be interpreted at the assessment-record level. Item-level responses were unavailable; consequently, the measurement-error analysis used reliability coefficients from an external Persian validation sample rather than reliabilities estimated directly from the present records [9]. The ± 3 -point definition is an analytic sensitivity band, not a confidence interval, and the Gaussian perturbations are controlled models of score variation rather than estimates of true temporal fluctuation. The present study evaluated the robustness of threshold-based MCMI-IV interpretations rather than diagnostic accuracy or criterion validity. No independent structured clinical diagnosis, longitudinal clinical outcome, or treatment-response criterion was available against which the fragility measures could be validated. These constraints do not undermine the primary descriptive result, which is calculated directly from observed distances to established thresholds, but they restrict its interpretation to the local robustness of test-based decisions.

The next step is therefore not to replace BR 75 and BR 85, but to test whether explicit information about decision fragility improves judgment. Multisite replication can establish whether the relative atlas structure is preserved across languages, referral settings, and diagnostic compositions. Prospective studies with item-level data, repeated administrations, and structured diagnostic interviews could separate measurement error from true short-term change and test whether near-threshold decisions predict clinician disagreement, diagnostic reclassification, or changes in treatment planning. Threshold-aware reports could then be evaluated experimentally by flagging near-boundary scores for contextual review while retaining the conventional BR category. The same general framework may also apply to symptom screeners, neuropsychological classifications, risk scores, and other instruments that convert continuous quantities into discrete clinical labels.

Conclusions

Threshold fragility was a systematic and reproducible feature of MCMI-IV Base Rate interpretations in the present dataset. Nearly one-third of all profile-by-scale decisions fell within ± 3 BR points of either BR 75 or BR 85, but the frequency of such decisions varied markedly across scales. This scale-specific pattern remained broadly consistent after accounting for estimated measurement precision, across different perturbation magnitudes, and across repeated split-half samples. Fragility was also strongly threshold-dependent, with most near-threshold decisions occurring around BR 75, although selected scales showed greater sensitivity around BR 85.

The supporting profile-space analysis further showed that unadjusted fragility was nonrandomly distributed across the multivariate MCMI-IV profile space. However, this spatial concentration was substantially attenuated and no longer statistically significant after adjustment for overall BR elevation. The profile-space findings therefore do not support an independent “fragile-profile” subtype, but instead indicate that much of the observed organization of fragility is related to the broader elevation structure of MCMI-IV profiles.

These findings suggest that reporting proximity to BR 75 and BR 85 may provide an additional uncertainty descriptor alongside conventional categorical interpretation, particularly for scores close to decision boundaries. The present study does not establish diagnostic validity, diagnostic accuracy, or clinical utility of such fragility information. External replication in independent populations and prospective clinical evaluation are therefore required to determine whether incorporating threshold-fragility information improves psychiatric assessment or clinical decision-making.

543 List of abbreviations

544 MCMI-IV, Millon Clinical Multiaxial Inventory–IV; BR, Base Rate; SEM, standard
545 error of measurement; DBSCAN, density-based spatial clustering of applications with
546 noise.

547 Additional files

548 **Additional file 1.** PDF. *Supplementary figures and tables.* Contains Supplementary
549 Figures S1–S2 and Supplementary Tables S1–S4, including the conceptual/method
550 overview, Diffusion/Mapper parameter-sensitivity analysis, MCMI-IV scale list, SEM
551 inputs, SEM-adjusted fragility values, and Gaussian volatility values.

552 Declarations

553 Ethics approval and consent to participate

554 All methods and procedures were performed in accordance with relevant eth-
555 ical guidelines and regulations, including the principles of the Declaration of
556 Helsinki of 1975, as revised in 1983. The study protocol and procedures involv-
557 ing human-participant data were reviewed and approved by the Ethics Committee
558 for Research Involving Human Participants at Shahid Beheshti University (approval
559 code: **IR.SBU.REC.1405.098**). The present work involved a retrospective analy-
560 sis of previously collected, anonymized assessment records. No directly identifiable
561 personal information was available to the researchers, and no additional contact
562 with participants occurred as part of this analysis. All data were handled and ana-
563 lyzed in accordance with the approved ethical framework and applicable institutional
564 requirements for confidentiality and protection of participant information.

565 **Consent for publication**

566 Not applicable. The manuscript contains no directly identifiable individual participant
567 information, images, or case-level details.

568 **Availability of data and materials**

569 The anonymized data supporting the findings of this study are not publicly available.
570 De-identified data may be made available by the corresponding author upon reasonable
571 and methodologically appropriate request, subject to applicable ethical, institutional,
572 and data-sharing restrictions.

573 **Competing interests**

574 The authors declare that they have no competing interests.

575 **Funding**

576 This research received no specific grant from any funding agency in the public,
577 commercial, or not-for-profit sectors.

578 **Authors' contributions**

579 M.E. and F.Z. contributed to the conceptualization and development of the analytic
580 framework and interpretation of the findings. F.Z. performed the formal analyses,
581 prepared the visualizations, and drafted the original manuscript. M.E. contributed
582 to data curation and critically reviewed and revised the manuscript. Both authors
583 reviewed and approved the final manuscript and agree to be accountable for the work.

584 **Acknowledgements**

585 The authors gratefully acknowledge Kamal Barzegar of NeuroCAI for assistance with
586 data recording and preparation.

References

- [1] Emons WHM, Sijtsma K, Meijer RR. On the consistency of individual classification using short scales. *Psychological Methods*. 2007;12(1):105–120. <https://doi.org/10.1037/1082-989X.12.1.105>.
- [2] Cheng Y, Liu C, Behrens J. Standard error of ability estimates and the classification accuracy and consistency of binary decisions. *Psychometrika*. 2015;80(3):645–664. <https://doi.org/10.1007/s11336-014-9407-z>.
- [3] Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*. 2017;126(4):454–477. <https://doi.org/10.1037/abn0000258>.
- [4] Skodol AE, Morey LC, Bender DS, Oldham JM. The Alternative DSM-5 Model for Personality Disorders: a clinical application. *American Journal of Psychiatry*. 2015;172(7):606–613. <https://doi.org/10.1176/appi.ajp.2015.14101220>.
- [5] Millon T, Grossman S, Millon C. *Millon Clinical Multiaxial Inventory–IV (MCMI–IV) Manual*. Bloomington, MN: Pearson Clinical Assessment; 2015.
- [6] Greene RL, Weiner IB. *Millon Clinical Multiaxial Inventory–IV*. In: Weiner IB, Greene RL, editors. *Handbook of Personality Assessment*. 2nd ed. Hoboken, NJ: Wiley; 2017. p. 231–256.
- [7] Rossi G, Hauben C, van den Brande I, Sloore H. Empirical evaluation of the MCMI-III personality disorder scales. *Psychological Reports*. 2003;92(2):627–642. <https://doi.org/10.2466/pr0.2003.92.2.627>.

- [8] Hesse M, Guldager S, Holm Linneberg I. Convergent validity of MCMI-III clinical syndrome scales. *British Journal of Clinical Psychology*. 2012;51(2):172–184. <https://doi.org/10.1111/j.2044-8260.2011.02019.x>.
- [9] Mohammadi MR, Delavar A, Hooshyari Z, Shakiba A, Salmanian M, Ghandi F, et al. Psychometric properties of the Persian version of Millon Clinical Multiaxial Inventory–IV (MCMI-IV). *Iranian Journal of Psychiatry*. 2021;16(1):43–51. <https://doi.org/10.18502/ijps.v16i1.5378>.
- [10] Mohammadi MR, Hooshyari Z, Delavar A, Amanat M, Mohammadi A, Abasi I, et al. Diagnostic validity of Millon Clinical Multiaxial Inventory–IV (MCMI-IV). *Current Psychology*. 2023;42(21):18052–18060. <https://doi.org/10.1007/s12144-022-02972-9>.
- [11] Grove WM, Vrieze SI. An exploration of the base rate scores of the Millon Clinical Multiaxial Inventory-III. *Psychological Assessment*. 2009;21(1):57–67. <https://doi.org/10.1037/a0014471>.
- [12] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Medicine*. 2007;4(10):e296. <https://doi.org/10.1371/journal.pmed.0040296>.
- [13] Coifman RR, Lafon S. Diffusion maps. *Applied and Computational Harmonic Analysis*. 2006;21(1):5–30. <https://doi.org/10.1016/j.acha.2006.04.006>.
- [14] Singh G, Mémoli F, Carlsson G. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: *Eurographics Symposium on Point-Based Graphics*; 2007. p. 91–100.

- 632 [15] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering
633 clusters in large spatial databases with noise. In: Proceedings of the Second
634 International Conference on Knowledge Discovery and Data Mining; 1996. p.
635 226–231.
- 636 [16] Magnano G, Tannoia C, Andrà C. A priori reliability of tests with cut score.
637 Psychometrika. 2015;80(1):44–64. <https://doi.org/10.1007/s11336-013-9371-z>.
- 638 [17] Rossi G, Sloore H. Diagnostic efficiency of the Dutch-language version of the
639 Millon Clinical Multiaxial Inventory-III. In: Halvorsen IV, Olsen SN, editors. New
640 Research on Personality Disorders. New York: Nova Science Publishers; 2008. p.
641 193–211.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementBMCflat.zip](#)