British Journal of Multidisciplinary Studies
ISSN: 2978-8110
DOI: 10.32996/bjmss
Journal Homepage: www.al-kindipublisher.com/index.php/jbmss

BJMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Educational Assessment and Statistical Robustness in University Athletes

Daniela Burgueño Theurel[1], Leslie Sue Hérnandez Mendoza[2] ✉, Nestor Otilio Hernández Flores[1] and José Juan Muñoz León[3]

[1] *Dirección de Actividades Deportivas, Universidad Veracruzana, Veracruz, Mexico*
[2] *Facultad de Matemáticas, Universidad Veracruzana, Veracruz, Mexico*
[3] *Facultad de Estadística e Informática, Universidad Veracruzana, Veracruz, Mexico*
Corresponding author: Leslie Sue Hérnandez Mendoza, **Email**: leshernandez@uv.mx

| ABSTRACT

Background: Psychological and physical assessments commonly combine item responses and field tests into categorical summaries, yet their stability depends on scoring keys and decision thresholds. Objective: To describe an auditable workflow for tracing and assessing classification sensitivity using an existing, previously published university-athlete dataset, without re-presenting its original well-being/performance association as new evidence. Methods: We linked 40 athletes (20 per team) across the supplied consolidated, response and field-test workbooks. We checked 39 six-point well-being responses per participant, applied the documented 17-item reverse key, evaluated strict pairwise rank changes and quartile-class transitions, exhaustively varied two unverified source responses over their admissible values, and contrasted recorded with printed physical-index thresholds. Conditional fixed-margin Monte Carlo comparisons were interpreted as exploratory sensitivity analyses. Results: Of 1,560 expected original responses, 1,558 could be cross-checked. The corrected-key total mean was 167.825 versus 148.325 for the direct sum. Among 780 pairs, 314 reversed their strict score order. Twenty-three of 40 quartile categories changed in the consolidated scenario; across 36 admissible missing-response configurations this number was 23–25. Seven athletes with physical-index total 10 were assigned a stored class inconsistent with the earlier article's printed thresholds. Conclusions: The study documents substantial classification sensitivity and provides an applied, traceable audit example. It neither establishes new psychological associations nor determines whether the original journal must issue a correction; that issue requires separate provenance review.

| KEYWORDS

Measurement sensitivity; psychological well-being; classification stability; reverse scoring; secondary data analysis; university athletes.

| ARTICLE INFORMATION

ACCEPTED: 15 October 2026 **PUBLISHED:** 16 November 2026 **DOI:** 10.32996/bjmss.2026.4.4.1

1. Introduction

Psychological and physical assessments often convert a set of responses or test measurements into a single index and subsequently into a small number of categories. Such reduction can simplify communication, but it creates an additional analytical layer: the interpretation of any categorical analysis depends on how individual questions were oriented, how component scores were added, and where category boundaries were defined. Consequently, the stability of derived classifications is a measurement question in its own right, separate from the substantive association the original study set out to test.

Ryff's multidimensional framework motivated the six-domain psychological measure (Ryff, 1989). The Spanish adaptation of Ryff's Psychological Well-Being Scales contains positively and negatively worded items whose scoring orientations must be respected when constructing dimensions (Díaz et al., 2006). In a distinct published investigation, Burgueño Theurel et al. (2025) analyzed well-being and athletic-performance indicators in 40 university athletes using categorical association measures and multiple correspondence analysis. That publication established the empirical context and reported the primary association question; the present analysis does not introduce a second participant sample or independently replicate the field study.

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

Instead, this manuscript asks a narrower methodological question: how stable are a participant's score, relative ordering, assigned class and inferential summary under traceable transformations of the existing data? The emphasis is on consequences of documented scoring and threshold rules, rather than on seeking new significant relationships. Sensitivity across defensible analysis choices is broadly consistent with the rationale of specification-curve and multiverse approaches, although no claim is made that the present case study develops those frameworks (Steegen et al., 2016; Simonsohn et al., 2020).

The contribution is an applied workflow, rather than a proposed new statistical estimator: (i) preserve provenance and distinguish original records from consolidated scores; (ii) take the instrument's documented scoring key as the reference and classify unkeyed sums solely as historical audit scenarios; (iii) report both cut-off-free pairwise reorderings and category changes; (iv) bound uncertainty from unavailable source responses without treating imputed values as observed; and (v) compare performance threshold versions transparently. This applied workflow is intended to make the consequences of measurement choices intelligible to researchers in sport, educational and psychological assessment.

2. Materials and methods

2.1. Data provenance, sampling, and unit of analysis

The empirical material consists solely of the consolidated workbook "DB BASE FINAL.xlsx", the workbook of original Ryff responses separated by team, the six-sheet workbook of field tests, the underlying thesis, and the article already published in Emergentes (Burgueño Theurel et al., 2025). The same 40 university football athletes—20 in the women's team and 20 in the men's team—contribute to both the previous publication and this secondary analysis. No new measurement or participant recruitment was performed. The combination (team, within-team subject identifier) was used to join records, since identifiers 1–20 recur in both teams.

The design is observational and cross-sectional. Team, recorded sex, testing location and competitive preparation phase cannot be disentangled within these data; consequently, team-specific differences are reported only as descriptive sensitivity analyses. Historical competition results were not treated as new participant-level observations.

2.2. Response traceability and treatment of uncertainty

The instrument has 39 items scored from 1 to 6, giving 40 × 39 = 1,560 expected responses. For each response, the numeric consolidated entry was checked against the response workbook. Two original cells were empty although the consolidated workbook held the value 2: women's team, local subject 14, item 36, and men's team, local subject 8, item 29. Since the source files cannot independently verify those two entries, they were retained only as the consolidated-file scenario and additionally varied over all six permissible responses each. This creates exactly $6^2 = 36$ finite scenarios; the scenario enumeration does not fabricate recovered observations or claim to identify their true responses.

2.3. Documented reverse-scoring rule

Let $x_{ij} \in \{1,2,3,4,5,6\}$ denote the observed response of participant i to item j. The reverse-key set identified in the 39-item adaptation of Díaz et al. (2006) is $I = \{2,4,5,8,9,13,15,20,22,25,26,27,29,30,33,34,36\}$. The keyed response is $y_{ij} = 7 - x_{ij}$ for $j \in I$ and $y_{ij} = x_{ij}$ otherwise. The unkeyed and keyed global totals for individual i are $S_i = \Sigma_j x_{ij}$ and $T_i = \Sigma_j y_{ij}$, respectively. It follows exactly that:

$$T_i = S_i + 7|I| - 2\,\Sigma_{(j\in I)}\, x_{ij} \quad (|I| = 17).$$

The same operation was applied within each of the six dimensions using their published item assignments. Cronbach's alpha was recomputed as a descriptive internal-consistency index, not as an independent validation of the six-factor structure. We refer to the recorded totals as "unkeyed": they are arithmetically reproducible yet omit the documented reverse transformations. This comparison does not by itself establish the exact processing history of the already published article.

2.4. Rank and category stability

For any two distinct participants a and b, the pair is concordant if $\text{sign}(S_a - S_b) = \text{sign}(T_a - T_b) \neq 0$, discordant if the nonzero signs are opposite, or tied under one or both scoring rules. We report all pair types and Kendall's tau-b, which accommodates ties. Because the pairs share participants, 780 pair comparisons are descriptive combinatorial units and must not be treated as 780 statistically independent observations.

For classification, an explicitly stated sample-relative rule was used separately for S and T: lower category if the total is at or below the empirical 25th percentile, upper category if strictly above the 75th percentile, and middle category otherwise. Percentiles were evaluated with linear interpolation. Categories therefore express relative sample location, not validated clinical thresholds. The two sets of cut-offs were recomputed independently. The number of changed categories and Cohen's unweighted kappa characterize agreement; no significance test was used to privilege either categorization.

2.5. Separating item uncertainty, thresholds, and sampling variability

We kept three uncertainty mechanisms conceptually distinct. First, the scoring comparison used the historically stored unkeyed total and the documented keyed total; since the key is determined by the instrument, these are not equally valid competing scoring models. Second, the two unverifiable original responses were enumerated over all 36 finite combinations; this is bounded source-record uncertainty, not bootstrap uncertainty. Third, using 4,000 resamples of size 40 with replacement from the keyed total distribution (fixed random seed 20261022), the two empirical quartiles were recomputed in each resample. Each originally observed keyed score was then reclassified using the resampled cut-offs, so this exercise measures sensitivity of cut-offs to sample composition while holding the evaluated subject scores fixed. The fraction of resamples in which each subject retained their reference category was recorded. This procedure should not be read as an individual diagnostic confidence interval.

2.6. Physical-performance index and published thresholds

The consolidated total RD is the sum of three component level scores (force, endurance and speed), and this arithmetic identity was checked for all observations. In the earlier article, the stated class boundaries were low 3–9, medium 10–12 and high 13–14 (Burgueño Theurel et al., 2025, p. 487). The observed consolidated labels instead group totals 6–10, 11–12 and 13–14, respectively. We thus evaluate both the historical stored grouping and the grouping obtained by applying the published numerical cut-offs literally. Neither alternative is silently substituted for the historical record.

The source field-test workbook labels the speed protocol "3/4 Court Sprint", whereas the thesis describes a 50 m sprint in other passages. Some female sprint-time entries are serialized Excel time values and the corresponding male entries are decimals. Therefore, no unified continuous sprint-time variable was asserted here without a verified timing convention. Distances for the Course Navette and ten-jump tests have clearer units, but the focus of this manuscript is on documented index construction rather than physiological calibration.

2.7. Association summaries and finite-margin simulation

To show the inferential consequences of classification choices without repeating the original research question as a new discovery, we formed 3 × 3 contingency tables for three psychological categorization definitions (stored class, recomputed unkeyed quartiles, and keyed quartiles) crossed with two performance definitions (stored class and literal published thresholds). Pearson's χ^2 and Cramer's V are descriptive comparisons of the resulting tables. The expected cell counts are too small for unqualified reliance on χ^2 asymptotics. Accordingly, conditional Monte Carlo tail areas for Pearson's χ^2 were calculated by drawing 100,000 random tables with fixed observed margins per specification using the multivariate hypergeometric/random-table distribution (seed 20261022); the estimate is (b+1)/(B+1), where b is the number of simulated statistics at least as large as the observed statistic and B=100,000. These six analyses are dependent sensitivity specifications, not a family of confirmatory independent hypothesis tests.

3. Results

3.1. Source verification

All 40 records can be paired across files by team and local subject identifier. Of the 1,560 expected individual responses, 1,558 match the consolidated numerical coding; two source cells remain unverifiable. Each of the six recorded dimension sums, the overall recorded well-being sum and the three-component RD total agrees with its arithmetic inputs for all 40 subjects. Arithmetic agreement does not resolve the orientation of negatively worded psychological items or the interpretation of all field-test units.

3.2. Consequences of keyed psychological scoring

The mean recorded unkeyed psychological total was 148.325 (sample standard deviation 10.923), whereas the documented keyed total had a mean of 167.825 (standard deviation 23.287). The corresponding totals were only weakly associated in rank (Spearman ρ = 0.2156; Kendall τ_b = 0.1559). This change is not a constant offset and may modify which subjects occupy extreme sample-relative positions.

Dimension	Unkeyed alpha	Documented-key alpha
Self-acceptance	−0.010	0.762
Environmental mastery	0.371	0.683
Positive relations	0.416	0.794
Personal growth	0.379	0.667
Autonomy	0.247	0.669

Dimension	Unkeyed alpha	Documented-key alpha
Purpose in life	0.620	0.761

Table 1. Sample Cronbach's alpha values (n = 40). The six keyed coefficients vary from 0.667 to 0.794. Their sample-specific magnitudes do not by themselves validate the factor structure, and a data-dependent increase in alpha is not a substitute for the documentary scoring rule.

3.3. Pairwise ordering and classification agreement

Of the 780 unordered subject pairs, 433 had a concordant strict order, 314 had a discordant strict order, 23 were tied only in the unkeyed total, 10 were tied only in the keyed total, and none were tied in both. The strict rank reversals represent 314/780 = 40.3% of all pairs; the fraction among 747 pairs having strict order on both scales is 42.0%.

Unkeyed category	Keyed low	Keyed middle	Keyed high	Row total
Low	3	7	0	10
Middle	5	10	5	20
High	3	3	4	10
Total	11	20	9	40

Table 2. Cross-tabulation of sample-relative quartile categories before and after application of the documented scoring key. Twenty-three of 40 subjects (57.5%) change category. Unweighted Cohen's kappa = 0.080; categorical disagreement should not be interpreted as actual within-person psychological change.

The empirical Q1 and Q3 were 139.75 and 155.25 for the unkeyed totals, and 153 and 184 for the keyed totals, respectively. In the cut-off resampling exercise, 26 of the 40 fixed keyed subject scores retained their reference categories in at least 90% of the 4,000 resampled thresholds. This is an internal measure of relative cut-off stability rather than external validation of low/medium/high well-being.

3.4. Exact uncertainty from two unverifiable original answers

The full 36-scenario enumeration yielded 23, 24 or 25 changed categories in the unkeyed-to-keyed comparison, and between 313 and 323 strict pairwise rank reversals. Crucially, the keyed category of each of the 40 subjects remained unchanged across all 36 scenarios when thresholds were recomputed within each scenario. Hence the major scoring-sensitivity finding cannot be explained solely by the two unavailable source answers. The responses remain unknown; this robustness statement is not evidence that the two consolidated values are authentic.

3.5. Sensitivity of outcome categories and contingency summaries

Seven participants had an RD total of 10. Their stored outcome class places them in the low group, while the printed cut-off in the prior article places them in the medium group. Thus the stored outcome distribution (low/middle/high: 13/13/14) becomes 6/20/14 under the published rule. Table 3 reports the full analysis multiverse, which has been defined for audit purposes, not for retrospective selection of a favorable p value.

Psychological classification	Performance classification	Pearson χ^2	Cramer V	Conditional MC p
Stored class	Stored class	8.8544	0.3327	0.0652
Stored class	Printed cut-offs	4.7023	0.2424	0.3305
Unkeyed quartiles	Stored class	9.6264	0.3469	0.0434
Unkeyed quartiles	Printed cut-offs	6.5571	0.2863	0.1703
Documented-key quartiles	Stored class	0.2368	0.0544	1.0000
Documented-key quartiles	Printed cut-offs	0.8941	0.1057	0.9385

Table 3. Dependence of nominal association summaries on scoring and threshold specifications, n = 40. Conditional p values use 100,000 fixed-margin Monte Carlo draws per table; the zero-presentation of a p value of 1.0000 denotes an estimated tail probability of one, not universal proof of independence. The unkeyed condition is a historical comparator, not an equally defensible alternative to the documented item key.

The apparent exception with conditional p approximately 0.0434 corresponds to recomputed quartiles of the unkeyed response sum and the stored performance classes. Reporting it as evidence of a discovered psychological relationship would be inappropriate, because its underlying score omits the specified item reversals and the specification was considered among

dependent post hoc comparisons. For the documented-key score, the two tested outcome groupings produce small Cramer's V estimates and large tail areas. These comparisons should not be taken as either causal evidence or a definitive null effect.

3.6. Descriptive associations with the numerical index

Using Spearman's coefficient with the stored numeric RD total, the overall unkeyed total yielded $\rho = 0.0002$ (approximately $p = 0.9991$), whereas the documented-key total yielded $\rho = -0.0375$ (approximately $p = 0.8181$). The within-team keyed estimates were −0.0342 for the women's team ($n = 20$) and 0.1643 for the men's team ($n = 20$); both are imprecise. No causal sex-specific interpretation is supported by the sample structure.

4. A formal framework for classification stability

Although the transformation identity is elementary, it gives a transparent algebraic account of why a scoring key can change the relative ordering of respondents. For participants a and b, subtracting their exact keyed-score identities yields:

$(T_a - T_b) = (S_a - S_b) - 2\,\Sigma_{(j \in I)}(x_{aj} - x_{bj})$.

If $S_a > S_b$, their order is strictly reversed if and only if twice the difference in the sum of reversed-item responses exceeds $S_a - S_b$. Thus stability of score order depends on the composition of the item responses rather than only on the number of reverse-keyed items.

A second elementary inequality clarifies how fixed or moving cut-offs can produce unstable classes. Suppose a baseline scalar score s is classified by thresholds $c1 < c2$, and a sensitivity analysis changes the score to $s^* = s + \delta$ and thresholds to $cj^* = cj + \gamma j$, with $|\delta| \le \varepsilon$ and $|\gamma j| \le \eta j$. Whenever $|s - cj| > \varepsilon + \eta j$ for $j = 1,2$, the participant remains on the same side of each threshold, and hence in the same three-level class. The proof is immediate from $|(s^* - cj^*) - (s - cj)| = |\delta - \gamma j| \le \varepsilon + \eta j$. This is a sufficient rather than necessary condition, and is useful to identify participants for whom classification is protected against prespecified bounded perturbations.

Neither identity is claimed as a new theorem of statistics. Their contribution here is organizational: an audit should separately quantify the scoring perturbation δ, movement in thresholds γj, and uncertainty in the source observations, rather than conflating them into one reported p value.

4.1. A sufficient margin certificate when score and cut-offs both move

Proposition 1 (joint perturbation certificate). Consider a three-class rule with score s and thresholds $c1 < c2$; a score revision produces s^* and revised thresholds $c1^* < c2^*$. Put $e = |s^* - s|$ and $hj = |cj^* - cj|$. If $|s - cj| > e + hj$ for both $j = 1,2$, then the class of s under (c1,c2) equals the class of s^* under (c1*,c2*), irrespective of the magnitudes or signs of the perturbations. The inequalities are sufficient, not necessary; boundary cases depend on whether equality belongs to the lower, middle or upper class.

$|(s^* - cj^*) - (s - cj)| \le |s^* - s| + |cj^* - cj| = e + hj$.

Proof. Under each strict margin condition, the difference between the score and that threshold cannot change sign, by the reverse triangle inequality. Since the class is uniquely determined by which side of each ordered threshold the score occupies, the class is unchanged. QED.

Corollary 1 (a checkable margin screen). Given a family of permissible source or processing configurations ω, define $e_i = \max_{\omega} |s_i(\omega) - s_i(\omega 0)|$ and $h_j = \max_{\omega} |cj(\omega) - cj(\omega 0)|$. Every participant satisfying $\min_j \{|s_i(\omega 0) - cj(\omega 0)| - e_i - h_j\} > 0$ has a certified invariant category throughout that family. Participants failing this screen are merely uncertified, not necessarily unstable. In the present study the complete 36-scenario enumeration gives an exact classification stability result and is therefore preferable to a conservative bound for the two uncertain responses.

4.2. Rank reversal certificate and the role of ties

Proposition 2 (pairwise reversal criterion). For participants a,b, write $U_{ab} = S_a - S_b$ and $R_{ab} = \Sigma(j \text{ in } I)(x_{aj} - x_{bj})$. Under the common reverse-scoring key, $T_a - T_b = U_{ab} - 2R_{ab}$. If $U_{ab} > 0$, the ordering reverses strictly if and only if $2R_{ab} > U_{ab}$, and becomes tied exactly if $2R_{ab} = U_{ab}$. The analogous inequalities with reversed signs hold when $U_{ab} < 0$. The statement follows by subtraction of the two score identities. In particular, the actual number of strictly discordant pairs is a property of the observed response matrix, not an independent-sample statistical estimate.

This identity clarifies the two denominators reported in Section 3: 314/780 is the proportion of all unordered subject pairs that switch strict order, whereas 314/747 restricts the denominator to pairs untied under both scoring systems. Neither ratio can be interpreted as arising from 780 or 747 independently sampled participants.

4.3. Sensitivity decomposition: score key versus category boundaries

For the stored scenario, the empirical unkeyed quartiles were (139.75, 155.25), and the documented-key quartiles were (153,184). Starting from unkeyed classifications, replacing the scores by keyed totals while artificially holding the old unkeyed cut-offs fixed changes the class of 28 subjects. Updating the thresholds to the keyed quartiles then changes the class of 25 subjects relative to that intermediate classification. The direct beginning-to-end disagreement is 23 subjects. These three figures are not additive because some participants change class twice or return to their original class.

The fixed-cut-off intermediate is a mathematical decomposition only: cut-offs derived from one score distribution do not automatically have a valid substantive meaning on the other distribution. The direct comparison that recalculates quantiles for each score is the pre-specified operational comparator, and all resulting categories remain sample-relative.

4.4. Inference boundaries for fixed-margin Monte Carlo sensitivity

The six contingency comparisons vary BOTH the psychological variable and the outcome grouping. Their p values are dependent and exploratory; they are not adjusted confirmatory evidence for or against a causal effect. The Monte Carlo p estimate is conditional on observed row and column margins and answers whether a Pearson statistic at least as extreme would occur in random tables with those margins under the conditional independence null. It does not correct measurement error, selection bias, confounding of team with sex and competitive phase, or choice of an inappropriate scoring key. No conclusion is based on choosing the smallest of the six p values.

5. Discussion

From an applied perspective, an arithmetically consistent workbook is not necessarily an interpretable measurement instrument. The direct sums were reproducible, but a scoring-orientation check against the cited Ryff instrument substantially changed rank order and quartile assignments. This distinction matters for practitioners who may use high/middle/low categories to summarize teams or identify research subgroups; such categories are relative descriptions, not diagnoses or validated screening thresholds.

The evidence illustrates three noninterchangeable checks. First, the 314 strictly reversed participant pairs measure score-key sensitivity without relying on arbitrary cut-offs. Second, the 23–25 class transitions express the consequences of combined scoring and percentile-class changes, not shifts in athletes' underlying psychological states. Third, the 36 admissible source-response cases evaluate traceability uncertainty, whereas re-estimating percentile boundaries under resampling concerns sampling instability. These checks address different questions and should not be combined into a single significance claim.

This analysis does not establish that one team has higher psychological well-being, that psychological state influences sporting achievement, or that the original reported association is valid or invalid. Nor does an association coefficient near zero demonstrate the absence of every conceivable relationship. Team membership is confounded with sex, setting and competitive phase, while both indices are constructed measurements with limited validation in this sample. The physical performance scale also mixes components and categories, and its original speed records cannot currently support a unit-consistent participant-level continuous analysis.

The earlier article investigated the substantive relationship between psychological well-being and athletic performance. This manuscript evaluates the stability of data processing and derived classifications for precisely the same 40 observations. Manuscripts based on the same data should be transparent about previous publications and offer a substantially distinguishable analysis (International Committee of Medical Journal Editors [ICMJE], 2026). The methodological reanalysis should not be interpreted as new evidence from independently recruited participants.

5.1. Practical reporting recommendations

For psychological-scale users, the questionnaire version and reverse-scoring key should be identified before computing dimensions or totals. A reproducible audit should distinguish copied raw response codes, keyed item values, dimension sums, and final labels; numerical sum checks alone cannot establish conceptual correctness.

For sport-science users, the operational definition of the performance composite, its component directions and the exact inclusion rules at cut-off values should be reported. Whenever original field-test time formats or test distances are ambiguous, analysts should not force a continuous cross-team metric from incomparable entries.

For small observational studies, publish or retain a contingency table supporting each association test; examine expected cell counts, use a justified exact/conditional approach where appropriate, and label alternate cut-offs as sensitivity analyses rather than independent confirmatory tests.

For university decision-makers, sample-relative quartile classes should not be used as individual mental-health diagnoses or as sole criteria for selecting athletes or allocating services. The present findings support data-governance improvements, not claims of individual psychological impairment.

6. Limitations

First, the sample is small (n = 40) and nonprobability-based; sampling uncertainty is consequently substantial. Second, two original response fields cannot be recovered from the available source workbook. Third, the recorded physical-testing protocol and some time formats are inconsistent across sources, preventing an unambiguous analysis of raw sprint seconds. Fourth, empirical quartile categories are sample-relative and do not have independent diagnostic meaning. Fifth, alpha assumes a particular interpretation of score aggregation and cannot establish dimensional validity. Sixth, the set of reported contingency comparisons is a dependent robustness exercise, not a preregistered confirmatory test battery. Finally, this is a secondary reanalysis of an already published cohort; editorial correction of the first article is a separate decision to be made with that journal. In addition, the applied framework is demonstrated in only one dataset and is not evidence of a newly validated scoring protocol or of broad generalizability.

6.1. Scope of the secondary analysis

An amendment to an earlier article and the consideration of a secondary applied-methods manuscript are distinct editorial questions. The former concerns the historical published record; the latter concerns the additional value of an integrated audit of source traceability, rank changes and classification thresholds. This study makes no claim that the earlier journal has examined the present audit, and the reuse of the 40 participants is disclosed in the manuscript and accompanying submission letter.

The scale documentation supports using the reverse key, but the available files do not establish that the published analytical pipeline relied on the particular surviving intermediate workbook inspected here. Likewise, the performance discrepancy compares a printed threshold rule with the present workbook labels; it is not a reconstructed version history of the prior publication. These limitations preclude a definitive claim of error about the published analysis.

7. Conclusions

A small, previously published sports-assessment dataset can provide a concrete example of why source traceability, documented reverse scoring and classification thresholds should be audited before interpreting aggregated categorical outcomes. Here, scoring orientation was associated with 314 strict pairwise reversals and 23 quartile-category changes in the consolidated scenario; two unverified original responses yielded 23–25 changes across their 36 possible value combinations. Seven recorded physical-performance classes differed from the published threshold rule. These are measurement-sensitivity results, not a new test of psychological determinants of performance. Their appropriate use is to improve research reporting and to inform a transparent secondary methodological discussion, while any correction of the original empirical report is handled separately.

Statements and editorial transparency

Relationship to prior publication. This paper is a transparent secondary methodological reanalysis of the same 40 observations reported in Burgueño Theurel et al. (2025), "Relación entre bienestar psicológico y rendimiento deportivo en estudiantes universitarios," Emergentes—Revista Científica, 5(2), 475–497, https://doi.org/10.60112/erc.v5.i2.431. No new participants, data collection or independent cohort are claimed. The prior paper examined psychological well-being and sport performance; the present manuscript focuses on traceability, score-key sensitivity, classification and inferential stability. Any amendment to the earlier publication is a separate editorial matter; no amendment or editorial determination is claimed here.

Ethics and consent. The source thesis reports that the 40 participants signed informed consent for the original observational study and that confidentiality safeguards were applied. The present manuscript is a secondary methodological reanalysis of retained data and reports only aggregated results; no new contact, tests or recruitment occurred. The available materials do not contain the original consent wording concerning secondary analysis or an ethics approval/waiver identifier. Accordingly, this paper does not claim an additional approval or an authorization that the supplied documents do not establish.

Data availability. The participant-level psychological questionnaire data and field-testing records are not publicly available as part of this manuscript because they concern individual research participants. The consolidated analysis and source records are held by the study team. Any request for controlled access is subject to applicable consent, institutional authorization and privacy safeguards; unrestricted sharing is not promised.

Funding: This research received no external or institutional funding. The article processing charge (APC), if applicable, will be paid from personal funds without external or institutional financial support.

Conflicts of Interest: The authors declare no conflicts of interest.

Author Contributions: Daniela Burgueño Theurel, Leslie Sue Hérnandez Mendoza, Nestor Otilio Hernández Flores, and José Juan Muñoz León contributed substantially to every stage of the study, including conceptualization, methodology, data review and validation, formal analysis, interpretation of findings, drafting, critical revision, and preparation of the manuscript. All four authors reviewed and approved the final manuscript and agree to be accountable for the work.

Use of AI-assisted tools: ChatGPT assisted with manuscript organization, academic-language drafting and editing. The four authors are responsible for validating the data interpretations, numerical results, references and final text. No AI system is credited as an author.

ORCID: Daniela Burgueño Theurel 0009-0008-1389-736X

Leslie Sue Hérnandez Mendoza 0009-0003-3108-7778

Nestor Otilio Hernández Flores 0009-0002-5290-2711

José Juan Muñoz León 0000-0003-3557-8251.

References

Burgueño Theurel, D., Muñoz León, J. J., Hernández Suárez, J., Montero Mora, J. G., & Argüello Ortiz, Á. F. (2025). Relación entre bienestar psicológico y rendimiento deportivo en estudiantes universitarios. Emergentes—Revista Científica, 5(2), 475–497. https://doi.org/10.60112/erc.v5.i2.431

Díaz, D., Rodríguez-Carvajal, R., Blanco, A., Moreno-Jiménez, B., Gallardo, I., Valle, C., & Van Dierendonck, D. (2006). Adaptación española de las escalas de bienestar psicológico de Ryff. Psicothema, 18(3), 572–577. https://www.psicothema.com/pdf/3255.pdf

International Committee of Medical Journal Editors. (2026). Recommendations: Overlapping publications; manuscripts based on the same database. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/overlapping-publications.html

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. Journal of Personality and Social Psychology, 57(6), 1069–1081. https://doi.org/10.1037/0022-3514.57.6.1069

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. Nature Human Behaviour, 4, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5), 702–712. https://doi.org/10.1177/1745691616658637

Appendix A. Reproducible analysis specification

Input unit: individual participant. Linkage key: (team, local subject number). Well-being items: 39 integer responses scored 1–6. Reverse-key set: {2, 4, 5, 8, 9, 13, 15, 20, 22, 25, 26, 27, 29, 30, 33, 34, 36}. Aggregate totals as recorded and as keyed; verify all component sums independently. Cut-offs: 25th percentile, 75th percentile with linear interpolation; low ≤ Q1, high > Q3, middle otherwise. Sensitivity to missing-source entries: six choices for each of two specific cells, all 36 combinations. Rank-pair ledger: concordant, discordant, unkeyed-only tie, keyed-only tie, both tie. Conditional comparison: Pearson statistic for each of six fixed-margin 3×3 tables; 100,000 draws, seed 20261022, with plus-one tail estimate. Bootstrap threshold audit: 4,000 resamples of 40 keyed totals; classify fixed subject scores against resampled cut-offs. Do not report pairs as independent cases or threshold-bootstrap frequencies as diagnostic probabilities.

Editorial transparency: this manuscript is a methodological reanalysis of the same 40 participants reported in Burgueño Theurel et al. (2025); any potential correction to that publication must be addressed separately.