Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Age- and Sex-Weighted Swiss PROMIS-29 Reference Values and Exploratory Differential Item Functioning

Matthew Kerry

kerr@zhaw.ch

ZHAW Zurich University of Applied Sciences

Thanh Elsener

ZHAW Zurich University of Applied Sciences

Short Report

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11454940/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Population reference values and evidence of item comparability are needed to interpret PROMIS-29 scores in Switzerland.

Methods: We analyzed Swiss PROMIS-29 v2.1 survey responses from adults aged 18–65 years (N = 1,580). Post-stratification weights aligned six age-by-sex cells with 2025 Swiss permanent-resident population counts. Seven PROMIS domains were expressed as T-scores. Exploratory graded response model (GRM)-based differential item functioning (DIF) screening examined age, sex, education-based socioeconomic strata, chronic conditions, and migration background.

Results: Weighted domain means ranged from 45.79 for fatigue to 53.64 for ability to participate in social roles. Of 140 item–group comparisons, two physical-function items exceeded McFadden's $\Delta R^2 > .02$; neither exceeded $|\Delta b| \geq 0.5$.

Conclusions: These estimates provide preliminary age- and sex-weighted Swiss reference values. The DIF findings require anchor purification and score-impact analysis before decisions about item revision or removal.

Introduction

The PROMIS-29 Profile assesses seven health domains with four items per domain and a single pain-intensity rating. Scores are commonly reported on a T-score metric referenced to United States calibration populations (mean 50, standard deviation 10). National reference distributions can improve interpretation without changing the underlying T-score metric. Population reference values have been established in other settings, including France [1]. A previous Swiss multilingual validation study supported the potential use of PROMIS-29 but identified limitations in selected domains and language comparisons [2]. This secondary analysis aimed to describe age- and sex-weighted Swiss reference values and screen for DIF across relevant population subgroups.

Methods

Design and participants. We used data from a cross-sectional Swiss survey administered January–March 2025 in German, French, and Italian [2]. The present analysis was restricted to respondents aged 18–65 years with recorded binary sex (N = 1,580). This age range and coding reflect the available weighting analysis; estimates should not be generalized to adults older than 65 years.

Scoring and weighting. Four-item domain summed scores were converted to PROMIS-29 v2.1 T-scores using published score-conversion lookups; pain intensity remained on its 0–10 scale. These secondary-analysis scores were not recalculated using the response-pattern IRT scoring server described in the original validation manuscript [2]. All three language groups, including the original Italian fatigue responses, were included in fatigue scoring (n = 1,580). We post-stratified respondents to 2025 Swiss

permanent-resident age-by-sex counts in the cells 18–34, 35–49, and 50–65 years, crossed with male/female sex; weights were normalized to mean one. Weighted means, standard deviations, and empirical percentiles were calculated. The effective sample size was approximately 1,492.

Exploratory DIF screening. We screened the 28 items across five grouping variables: age 18–39 versus 40–65, sex, education-based socioeconomic strata, chronic condition (yes/no), and migration background (yes/no). Pooled marginal GRM calibration was combined with posterior matching based on the other three items in each four-item domain; group-related item effects were assessed using conditional item likelihoods. We flagged McFadden's $\Delta R^2 > .02$ as potentially important DIF, consistent with COSMIN guidance [3, 4], and $|\text{uniform } \Delta b| \geq 0.5$ as a separate exploratory difficulty-shift criterion. Matching anchors were not iteratively purified. Planned impact criteria—mean difference ≥ 0.2–0.3 SD, T-score shift ≥ 2–3 points, and classification change ≥ 5–10%—were not yet evaluated. No item-removal decision was made.

DIF impact analysis. For each grouping variable and domain, we compared original T-scores with exploratory counterfactual scores adjusted for estimated uniform DIF across all four items, including subthreshold effects. We assessed focal-group weighted mean T-score shifts, standardized shifts, proportions with individual changes of ≥ 2 or ≥ 3 T-points, and changes in classification at illustrative T-score cutoffs of 45, 50, 55, and 60. Prespecified impact criteria were mean differences ≥ 0.2–0.3 SD, mean T-score shifts ≥ 2–3 points, or classification changes ≥ 5–10%. These analyses used unpurified pooled GRM calibration and summed-score conversion rather than fully purified multigroup IRT rescoring.

Results

Reference values. The age- and sex-weighted sample included 1,580 respondents (755 male, 825 female). Mean scores were above the US reference mean for physical function and social roles, and below 50 for depression, fatigue, sleep disturbance, and pain interference (Table 1). Physical function showed a marked upper-end concentration (weighted median and 95th percentile both 57.0). The fatigue estimates include all three survey languages. Age- and sex-specific distributions are available in the accompanying analysis dataset.

Table 1
Age- and sex-weighted Swiss PROMIS-29 reference values

Domain	n	Mean (SD)	P5	Median	P95
Physical function	1580	53.60 (6.37)	39.2	57.0	57.0
Anxiety	1580	49.73 (8.31)	40.3	51.2	63.4
Depression	1580	48.38 (7.97)	41.0	49.0	62.2
Fatigue	1580	45.79 (9.43)	33.7	46.0	62.7
Sleep disturbance	1580	47.80 (7.87)	32.0	48.4	61.7
Social roles	1580	53.64 (8.28)	40.5	53.7	64.2
Pain interference	1580	48.82 (7.85)	41.6	49.6	62.5
Pain intensity†	1570	3.10 (2.07)	1.0	3.0	7.0

† Pain intensity is 0–10; other domains are PROMIS T-scores. Weights correct age and sex distributions.

Known-groups differences. Chronic-condition status was associated with worse scores across all seven domains, with weighted between-group differences ranging from 2.13 T-score points for anxiety to 5.75 for pain interference. Migration-background differences were smaller, ranging in absolute magnitude from 0.69 to 2.28 points (Table 2). These comparisons support descriptive known-groups interpretation rather than estimates of minimal important change.

Table 2
Weighted known-groups differences (Yes minus No)

Domain	Chronic ΔT	d	Migration ΔT	d
Physical function	-5.43	-0.79	-0.94	-0.14
Anxiety	+ 2.13	+ 0.25	+ 1.79	+ 0.21
Depression	+ 2.95	+ 0.36	+ 0.69	+ 0.09
Fatigue	+ 4.69	+ 0.50	+ 1.21	+ 0.13
Sleep disturbance	+ 3.25	+ 0.42	+ 2.28	+ 0.29
Social roles	-3.77	-0.45	-1.58	-0.19
Pain interference	+ 5.75	+ 0.73	+ 0.75	+ 0.09

ΔT = difference in weighted group means (PROMIS T-score points); d = weighted pooled-SD standardized difference. Fatigue comparisons were recalculated using all three language groups.

DIF. Across 140 item–group comparisons, two comparisons exceeded McFadden's $\Delta R^2 > .02$: PhysFun2R by age and PhysFun1R by chronic-condition status (Table 3). Neither exceeded the

prespecified |Δb| ≥ 0.5 threshold. No other comparison exceeded $\Delta R^2 > .02$. Because anchor purification and score-impact evaluation remain incomplete, these signals do not establish consequential DIF.

Cumulative DIF impact. Across 35 domain-by-group focal comparisons, the largest absolute weighted mean shift was 1.04 T-points (0.122 SD), observed for physical function among respondents with chronic disease; no comparison met the prespecified mean-impact criteria. At an illustrative T-score cutoff of 50, five comparisons had ≥ 5% classification changes: physical function/chronic disease (10.7%), sleep disturbance/chronic disease (8.9%), social roles/female sex (7.6%), social roles/migration background (6.6%), and social roles/chronic disease (5.4%). For the age-related physical-function comparison, the cumulative mean shift was 0.45 T-points (0.063 SD) and classification change was 3.7%. Classification findings were cutoff-dependent and do not establish clinical misclassification.

Table 3
Preliminary GRM-based DIF signals exceeding McFadden's $\Delta R^2 > .02$

Domain / item	Focal group	ΔR^2	Uniform Δb
Physical function / PhysFun2R	Age 40–65	0.0256	+ 0.323
Physical function / PhysFun1R	Chronic condition: Yes	0.0297	+ 0.325

ΔR^2 is McFadden's total pseudo-R^2 increment; Δb is the estimated uniform GRM difficulty shift. Provisional pooled calibration and unpurified three-item anchors were used.

Exploratory anchor-based analyses. Self-rated health (SRH) was moderately correlated with all seven PROMIS-29 domains (Spearman's |ρ| = 0.32–0.48), supporting its potential suitability as an external anchor (Supplementary Table S1). Weighted PROMIS T-scores showed consistent gradients across SRH categories, with poorer SRH associated with worse scores in all domains. Differences between adjacent SRH categories were generally smaller at the healthier end of the distribution, particularly between very good and good health, than between good and fair health. The former contrast was therefore considered the most promising candidate for further investigation of minimally important differences (MIDs). These cross-sectional estimates are exploratory and should not be interpreted as established MIDs for longitudinal change.

Discussion

These age- and sex-weighted estimates offer a starting point for Swiss interpretation of PROMIS-29 v2.1. Differences from the US-centered T-score reference should not be interpreted as evidence of superior or inferior national health without consideration of sampling and measurement comparability. In particular, post-stratification by age and sex does not correct disproportionate recruitment across Switzerland's language regions; "nationally representative norms" would therefore overstate the evidence. All three language groups were included in the recalculated fatigue norms. Short-form ceiling effects may restrict discrimination among people reporting high physical functioning.

Exploratory DIF screening identified only two comparisons above the COSMIN-associated McFadden threshold [3, 4]. However, a four-item domain is especially vulnerable to contaminated matching anchors. Before drawing conclusions about substantive DIF, future analyses should purify anchors, evaluate model fit and group linking, examine content interpretation, and quantify DIF-adjusted T-score shifts and classification changes. Items should not be removed solely because a screening threshold is exceeded; decisions must account for construct coverage and practical consequences. These results support cautious use of provisional reference distributions and further psychometric validation.

The cumulative sensitivity analysis likewise suggested limited impact on average scores, although several classification percentages crossed the exploratory 5% threshold at T = 50. Because this is not an established clinical cutoff, these findings should not motivate item deletion without content review, purified multigroup GRM calibration, and validation against clinically meaningful thresholds. We therefore recommend retaining the items provisionally and treating the reported impact estimates as exploratory.

Declarations

Author Contribution

MK conceived of follow-up study. TE conducted initial analyses and scoring. MK conducted initial write-up. TE verified data accuracy and reviewed.

Data Availability

All data is provided to interested researchers upon written request.

References

1. Coste J, Rouquette A, Valderas JM, Rose M, Leplège A (2018) The French PROMIS-29: Psychometric validation and population reference values. Rev Épidémiol Santé Publique 66(5):317–324. 10.1016/j.respe.2018.05.563
2. Elsener T, Kerry M, Christensen KB, Biller-Andorno N Measurement Properties of the PROMIS-29 v2.1 Profile in German, French, and Italian: A Cross-Sectional Survey in the Swiss Population. Source manuscript supplied by study authors; publication status and citation details to be confirmed
3. Prinsen CAC, Mokkink LB, Bouter LM et al (2018) COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 27:1147–1157. 10.1007/s11136-018-1798-3
4. Mokkink LB, Elsman EBM, Terwee CB (2024) COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. Qual Life Res. 10.1007/s11136-024-03761-6
5. Gagnier JJ, de Arruda GT, Terwee CB, Mokkink LB (2025) COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures: version 2.0. Qual Life Res

34:1901–1911. 10.1007/s11136-025-03950-x

Supplementary Files

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

- PROMIS29NormsDIFSupplementaryTableS1SRH.docx