

Development and Psychometric Validation of the Sudan Women's Empowerment Index (SWEI-SD): Integrating Factor Analysis, Item Response Theory, and Measurement Invariance

Sara Mohamed Elhassan Taha^{1,2*}, Khalid Rahamtalla Khider Genawi²

¹University College in Darb, Jazan University, Jazan, Saudi Arabia

²Department of Statistics – College of Science – Sudan University of Science and Technology

*Corresponding author: smtaha@jazanu.edu.sa

Abstract

Background: Women's empowerment is multidimensional and context dependent, yet few psychometrically validated instruments are available for conflict-affected Sudan. This study developed and validated the Sudan Women's Empowerment Index (SWEI-SD) as a culturally grounded measure of women's agency. **Methods:** A cross-sectional psychometric study included 1,100 Sudanese women aged 18 years or older. An initial 42-item pool spanning seven conceptual domains was evaluated using exploratory factor analysis (EFA; $n = 550$), independent confirmatory factor analysis (CFA; $n = 550$), bifactor modelling, graded response item response theory, local-dependence diagnostics, differential item functioning (DIF), and measurement-invariance analyses. Item reduction integrated statistical performance and conceptual coverage. **Results:** EFA supported three factors rather than seven distinct empirical dimensions, while bifactor analysis demonstrated strong general-factor dominance (ECV = 0.830; $\omega = 0.979$; $\omega_{\text{hierarchical}} = 0.959$). All 42 items showed satisfactory discrimination and ordered thresholds. Evidence from redundancy, local dependence, item fit, DIF, invariance, and substantive coverage supported reduction to 28 items, retaining four per conceptual domain. The SWEI-SD-28 achieved EAP reliability of 0.944, mean posterior SE of 0.220, and preserved the 42-item latent ranking almost perfectly (Pearson $r = 0.990$; Spearman $\rho = 0.987$). No practically meaningful DIF remained after recalibration. Among 1,019 eligible respondents, the mean 0–100 score was 54.41 (SD = 21.54). **Conclusion:** The SWEI-SD-28 provides a concise, reliable, and contextually appropriate measure of women's empowerment in Sudan. External replication, longitudinal validation, and full WLSMV invariance testing are warranted before implementation.

Keywords: Empowerment; Item response theory; Measurement invariance; Psychometric validation; Sudan; Women's agency; Women's empowerment

1. INTRODUCTION

Women's empowerment is a central objective of global development and gender-equity agendas because it reflects women's capacity to exercise meaningful choice, influence decisions, access and control resources, and pursue valued goals within households and society. Rather than representing a single attribute, empowerment is widely conceptualized as a multidimensional and context-dependent process. Kabeer's influential framework describes empowerment through the interrelated domains of resources, agency, and achievements, emphasizing that the ability to make strategic life choices depends on access to material and social resources and on the capacity to define goals and act upon them (Kabeer, 1999). Subsequent scholarship has expanded this understanding to intrinsic, instrumental, and collective forms of agency, reinforcing that empowerment operates across individual, relational, community, and institutional levels (Desai et al., 2022; Yount et al., 2019). Valid measurement therefore requires

instruments that capture multiple dimensions while remaining sensitive to the social and cultural contexts in which choices are made.

Despite growing international interest, measuring women's empowerment remains methodologically challenging. Existing instruments vary substantially in theoretical foundations, domains, item selection, scoring procedures, and target populations. Reviews show that commonly assessed dimensions include household decision-making, economic autonomy, mobility, gender norms, reproductive agency, self-efficacy, and access to resources, yet there is little consensus regarding which dimensions constitute empowerment or how they should be combined (Costa et al., 2023; Nahar & Mengo, 2022; Shipanga et al., 2026). Many measures are constructed from variables available in pre-existing surveys rather than from items specifically developed around a locally grounded conceptual model. Others are sector-specific, such as agricultural or reproductive-health measures, potentially limiting their applicability to broader social settings (Alkire et al., 2013; Vizheh et al., 2021). These differences complicate comparisons across populations and can obscure dimensions of agency that are especially salient in settings affected by displacement or conflict.

Several initiatives have attempted to strengthen standardized measurement. The Women's Empowerment in Agriculture Index (WEAI) incorporated production decisions, productive resources, income control, leadership, and time allocation (Alkire et al., 2013), whereas the SWPER index used widely available survey items to assess social independence, decision-making, and attitudes toward violence across African and other low- and middle-income settings (Ewerling et al., 2017, 2020). More recent psychometric work has moved beyond additive indices and principal-component approaches by examining latent structure, item performance, and cross-group comparability. Item response theory (IRT) analyses of the project-level WEAI showed that some empowerment item sets performed adequately while others demonstrated poor fit or limited equivalence across contexts (Yount et al., 2019). Similarly, the Women's Empowerment Metric for National Statistical Systems (WEMNS) used exploratory and confirmatory factor analysis and measurement-invariance testing, illustrating both the promise and difficulty of constructing measures that function equivalently across populations (Yount et al., 2026). These findings support explicit testing of dimensionality, local dependence, differential item functioning, and measurement invariance rather than assuming that an instrument developed elsewhere is valid in a new cultural setting.

This issue is particularly important in Sudan, where women's opportunities, responsibilities, and expressions of agency are shaped by social norms, economic conditions, geographic inequalities, displacement, and political instability. Since large-scale armed conflict began in April 2023, women and girls have faced intensified threats to safety, livelihoods, essential services, and social protection, while women's organizations have also played important roles in humanitarian response and local peacebuilding (UN Women, 2025, 2026). Sudan-focused research has examined specific aspects of agency, including reproductive decision-making and decisions surrounding female genital mutilation/cutting, illustrating how education, gender norms, family relationships, and local context influence women's participation in consequential decisions (Badri et al., 2026; Sabahelzain et al., 2019). However, such domain-specific indicators do not provide a comprehensive, psychometrically validated measure of

A culturally grounded instrument is therefore needed to distinguish genuine variation in empowerment from variation caused by poorly functioning or contextually inappropriate items. Factor-analytic methods are valuable for establishing construct structure, whereas IRT provides complementary information on item discrimination, response thresholds, and measurement precision across the latent empowerment continuum. Differential item functioning and measurement-invariance analyses can further assess whether scores are comparable across demographic and social groups. Integrating these approaches provides a rigorous basis for developing a concise instrument that preserves conceptual breadth while improving precision and interpretability. Evidence from contemporary empowerment-measurement research increasingly supports this combined psychometric approach, particularly when instruments are intended for comparison across heterogeneous populations (Yount et al., 2019, 2026).

Accordingly, this study aimed to develop and psychometrically validate the Sudan Women's Empowerment Index (SWEI-SD), beginning with a 42-item culturally informed pool and deriving a more parsimonious final instrument. Exploratory and confirmatory factor analysis, bifactor modelling, reliability assessment, graded-response IRT, local-dependence diagnostics, differential item functioning, and measurement-invariance testing were used to evaluate dimensional structure, item performance, precision, and cross-group comparability. The resulting SWEI-SD-28 is intended to provide a theoretically grounded and empirically robust measure of multidimensional women's empowerment in Sudan and to support future research, program evaluation, and context-sensitive monitoring.

2. METHODS

2.1 Study Design, Population, and Sampling

This study used a quantitative, cross-sectional, sequential psychometric design to develop and validate the Sudan Women's Empowerment Index (SWEI-SD). The target population comprised Sudanese women aged 18 years or older. The individual survey followed a multistage cluster field design organized across state, locality, community, and household levels. To ensure adequate representation for prespecified subgroup analyses, particularly differential item functioning (DIF) and measurement-invariance analyses, selected residence and displacement groups were purposively oversampled. Accordingly, the sampling strategy was designed primarily for psychometric validation and subgroup comparability rather than for the estimation of nationally weighted prevalence.

The achieved sample comprised 1,100 women. Sample-size planning considered both survey and psychometric requirements. Cochran's formula initially yielded a requirement of approximately 385 participants; after adjustment for clustering, the estimated requirement increased to approximately 970, with an upper planning benchmark of approximately 1,141 after allowance for anticipated nonresponse. An operational target of approximately 1,100 was therefore adopted to balance field feasibility with the requirements of factor analysis, IRT calibration, DIF, and subgroup-equivalence testing.

For independent structural validation, participants were reproducibly divided into an exploratory factor analysis (EFA) development sample ($n = 550$) and an independent confirmatory factor analysis (CFA) validation sample ($n = 550$). The split was stratified

jointly by residence and displacement status using a fixed random seed (20260925). The validation sample was not used to determine the exploratory factor count or initial EFA item-review decisions. Subsequent IRT calibration, DIF analysis, item reduction, recalibration, and final scoring used the full sample after completion of the independent structural analyses.

2.2 Instrument Development and Cultural Adaptation

The initial SWEI-SD comprised 42 candidate items organized a priori into seven conceptual domains, with six items per domain: intrinsic agency and self-efficacy; household and strategic decision-making; economic, financial, and asset control; mobility and access autonomy; information, digital, and learning agency; collective, community, and civic agency; and health, bodily, reproductive, and time autonomy. The domains were informed by agency-based and multidimensional conceptualizations of women's empowerment, particularly Kabeer's framework of strategic life choices (Kabeer, 1999). The seven-domain framework was intended to ensure content breadth and was not assumed to represent the final empirical factor structure.

Items were formulated to capture women's perceived capacity to decide, act, obtain resources, refuse unwanted actions, participate, communicate, and seek assistance rather than relying exclusively on conventional indicators of who makes household decisions. Each item used four ordered response categories scored from 0 ("not true at all") to 3 ("completely true"), with higher values representing greater empowerment. "Not applicable" and unavailable responses were treated as missing and were never recoded as zero.

The questionnaire was initially developed in English and translated into Sudanese colloquial Arabic. Bilingual reconciliation was undertaken to resolve wording differences and preserve conceptual rather than literal equivalence, followed by back-translation to identify potential shifts in meaning (Beaton et al., 2000). Expert review assessed clarity, cultural relevance, comprehensibility, and conceptual coverage. Cognitive interviewing was subsequently used to evaluate respondents' interpretation of the items, response categories, and key empowerment concepts, followed by pilot testing of wording, administration flow, response-option functionality, and field feasibility. Because numerical expert-rating data were not retained, item- and scale-level content-validity indices were not calculated retrospectively.

2.3 Data Management and Missing Item Responses

Individual responses were captured electronically using the structured study questionnaire under standardized field procedures designed to accommodate variation in literacy and protect privacy for sensitive content. A reproducible data-cleaning workflow preserved the raw dataset while standardizing variable names, recoding categorical responses, checking allowable ranges and logical consistency, and explicitly coding missing values.

No single-value imputation was performed for SWEI-SD item responses. Pairwise information was retained where permitted by the relevant psychometric analysis, while IRT scoring accommodated partial response patterns subject to prespecified score-eligibility criteria. "Not applicable," unavailable, blank, and otherwise invalid responses remained missing rather than being interpreted as low empowerment.

2.4 Exploratory and Confirmatory Factor Analyses

Because SWEI-SD items were ordinal, EFA was based on a polychoric correlation matrix. Factorability was assessed using the Kaiser-Meyer-Olkin measure of sampling adequacy, item-level measures of sampling adequacy, Bartlett's test of sphericity, and

inspection of the correlation structure. Factor retention was determined using 95th-percentile parallel analysis with 5,000 random reference datasets rather than the eigenvalue-greater-than-one criterion alone (Garrido et al., 2013; Hayton et al., 2004). Principal-axis common-factor extraction with oblique oblimin rotation was used because empowerment dimensions were expected to correlate. Item review considered primary loadings, cross-loadings, communalities, substantive interpretability, and conceptual coverage; items were not removed solely because they crossed a single exploratory loading threshold.

CFA was conducted exclusively in the reserved validation sample. Competing structures included a single general factor, a two-factor model, the EFA-derived three-factor structure, the original seven correlated conceptual domains, a second-order seven-factor structure, a bifactor model, and an EFA-informed three-factor model. Because indicators were ordinal, the CFA used a polychoric/diagonally weighted least-squares (DWLS)-style correlation-structure implementation. Model evaluation considered χ^2 , comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR), together with loading patterns, residuals, model complexity, and substantive interpretability. The completed implementation was not equivalent to full WLSMV; therefore, global fit indices were interpreted as approximate and in conjunction with convergent psychometric evidence. Bifactor modelling was used to assess whether a dominant general empowerment factor coexisted with residual domain-specific variance. Interpretation included general-factor loadings, explained common variance (ECV), omega total (ω), and omega hierarchical (ω_H) (Reise, 2012; Rodriguez et al., 2016). Cronbach's alpha was retained as a descriptive internal-consistency statistic, whereas omega-based indices and subsequent IRT reliability were emphasized when evaluating score reliability.

2.5 Item Response Theory and Item-Level Diagnostics

Following structural evaluation, the 42 candidate items were calibrated in the full sample using Samejima's graded response model (GRM), appropriate for ordered polytomous responses. For each item, the discrimination parameter (a_j) quantified the strength with which the item differentiated respondents along the latent empowerment continuum, while three category-threshold parameters (b_{j1} , b_{j2} , and b_{j3}) represented transitions between successive response categories. Ordered thresholds ($b_1 < b_2 < b_3$) were required for coherent category functioning.

Person-level latent empowerment estimates were obtained using expected-a-posteriori (EAP) scoring. Item and test information functions and associated conditional standard errors were examined to characterize measurement precision across the latent empowerment continuum. Item fit was assessed by comparing observed and model-predicted category proportions across deciles of EAP ability. Root mean square deviation (RMSD) >0.08 was treated as a review flag rather than an automatic exclusion criterion. Local dependence was evaluated using adjusted Yen's Q3, with values >0.20 similarly flagged for substantive review. These diagnostics were interpreted jointly with discrimination, threshold ordering, information, redundancy, and conceptual content.

2.6 Differential Item Functioning and Measurement Invariance

DIF was evaluated across four prespecified comparisons: urban versus rural residence, displacement-affected versus never-displaced status, currently married versus not currently married, and postsecondary versus secondary-or-less education. The EAP latent-trait estimate served as the matching variable. Three nested proportional-odds ordinal logistic models were fitted for each item and comparison. Uniform DIF was

evaluated by adding group membership to the latent-trait model, whereas non-uniform DIF was assessed through the latent-trait-by-group interaction.

Statistical significance was controlled within each subgroup family using the Benjamini–Hochberg false discovery rate procedure (Benjamini & Hochberg, 1995). A change in McFadden pseudo- $R^2 \geq 0.01$ was prespecified as evidence of practically non-negligible DIF. Statistically flagged effects underwent leave-one-item-out purification, whereby the matching latent score was recalculated without the item under evaluation.

Multi-group measurement invariance was evaluated across the same principal groups through sequential configural, metric, and ordinal threshold/scalar comparisons. Changes in CFI, RMSEA, and SRMR were considered jointly rather than relying solely on χ^2 difference testing. When complete equality was not supported, partial invariance was examined by freeing a limited number of parameters showing evidence of non-equivalence (Chen, 2007; Cheung & Rensvold, 2002; Putnick & Bornstein, 2016).

2.7 Evidence-Based Item Reduction and Final SWEI-SD-28 Scoring

Item reduction was based on convergent psychometric and substantive evidence rather than a single statistical criterion. Evidence considered included EFA loadings and cross-loadings, CFA and bifactor behaviour, reliability contribution, GRM discrimination and threshold ordering, item information, RMSD, adjusted-Q3 local dependence, DIF, measurement invariance, redundancy, and conceptual coverage. A design constraint of four retained items per conceptual domain was applied to maintain breadth, yielding a final 28-item instrument across the seven original conceptual domains.

The retained SWEI-SD-28 items were recalibrated in the full sample using a unidimensional GRM, consistent with the dominant general factor identified during structural evaluation. The final model was reassessed for discrimination, threshold ordering, test information, EAP reliability, posterior standard error, local dependence, and DIF. Short-form performance was compared with the 42-item bank using Pearson and Spearman correlations between EAP estimates and the mean absolute difference between latent scores.

For respondent (i), the final 0–100 score was derived from model-implied expected item scores:

$$\text{SWEI-SD}_i = \frac{100}{3J} \sum_{j=1}^J E(X_{ij} | \hat{\theta}_i), \quad J = 28.$$

Higher scores indicate greater empowerment. A score was generated only when at least 21 of 28 retained items were observed and at least six of the seven conceptual domains were represented; a domain was considered represented when at least three of its four items had valid responses. Missing and “not applicable” responses were not scored as zero. No low, moderate, or high normative categories were imposed because no external criterion or nationally representative reference distribution was available to justify such thresholds. (1)

2.8 Statistical Software and Reproducibility

All analyses were implemented using reproducible scripted workflows in Python 3.13.5. The analytical environment included pandas 2.2.3, NumPy 2.3.5, SciPy 1.17.0, statsmodels 0.14.6, scikit-learn 1.8.0, patsy 1.0.2, and Matplotlib 3.10.8. Intermediate outputs—including the EFA/CFA split, polychoric matrices, IRT parameters, EAP scores, DIF and invariance results, item-retention decisions, and final scoring outputs—were archived. Reproducibility safeguards included preservation of raw data, fixed random seeds, explicit cleaning rules, separate development and validation samples,

2.9 Ethical Considerations

Ethical safeguards were implemented in accordance with the approved study protocol. Participation was voluntary, and data-collection procedures emphasized privacy, confidentiality, non-coercive administration, secure handling of responses, and participants' ability to decline sensitive questions. Particular safeguards were applied to items concerning bodily autonomy, consent, displacement, and safety, and de-identified analytical records were used wherever possible.

3. RESULTS

3.1 Participant characteristics

The psychometric analysis included 1,100 Sudanese women. Valid age data were available for 996 participants; mean age was 35.96 years (SD = 13.54), with a median of 34 years (IQR, 24–44). Urban residents accounted for 57.4% of the sample and rural residents for 23.7%, while 10.5% were living in temporary camps or settlements. Approximately one-third of participants were currently married (34.9%), and 32.3% had never married. Educational attainment was heterogeneous: 14.4% reported no formal education, 17.8% had completed secondary education, 23.7% held a bachelor's degree, and 7.8% had postgraduate education. Conflict-related displacement was also well represented, with 26.3% currently internally displaced, 26.5% having returned following displacement, and 27.2% reporting no displacement related to the current or recent conflict (Table 1).

Table 1. Selected characteristics of the psychometric sample (N = 1,100)

Characteristic	Category	n (%)
Age, years	Mean ± SD	35.96 ± 13.54
	Median (IQR)	34 (24– 44)
Residence	Urban	631 (57.4)
	Rural	261 (23.7)
	Temporary camp/settlement	116 (10.5)
	Other	92 (8.4)
Marital status	Never married	355 (32.3)
	Married	384 (34.9)
	Divorced	148 (13.5)
	Separated	97 (8.8)

	Widowed	116 (10.5)
Education	No formal education	158 (14.4)
	Primary	119 (10.8)
	Intermediate	114 (10.4)
	Secondary	196 (17.8)
	Diploma	112 (10.2)
	Bachelor's degree	261 (23.7)
	Postgraduate	86 (7.8)
	Other	54 (4.9)
Displacement	Never displaced	299 (27.2)
	Currently internally displaced	289 (26.3)
	Returned after displacement	291 (26.5)
	Previously displaced, living elsewhere	111 (10.1)
	Other	110 (10.0)

Note. Age statistics are based on 996 valid observations.

3.2 Latent structure and dimensionality

The development subsample ($n = 550$) showed strong suitability for factor analysis. The Kaiser-Meyer-Olkin statistic was 0.932, and Bartlett's test of sphericity was significant, $\chi^2(861) = 15,006.92$, $p < 0.001$. Parallel analysis retained three factors: the first three observed eigenvalues (20.033, 1.817, and 1.557) exceeded their corresponding 95th-percentile random eigenvalues, whereas the fourth did not (Figure 1). The three-factor principal-axis solution accounted for 52.47% of the variance. Importantly, the seven domains specified during instrument development did not emerge as seven statistically distinct factors; instead, substantial cross-domain clustering was observed.

In the independent validation sample ($n = 550$), multidimensional models reproduced the observed correlation structure better than the one-factor model, although none of the strict simple-structure models achieved conventionally strong fit across all approximate global indices. The one-factor model showed CFI = 0.743 and

RMSEA = 0.099, whereas the strict EFA-derived three-factor model improved fit to CFI = 0.798, RMSEA = 0.087, and SRMR = 0.053. The lowest residual discrepancy was observed for the EFA-informed three-factor model allowing prespecified secondary loadings (CFI = 0.813, RMSEA = 0.085, SRMR = 0.049) (Table 2).

Despite residual multidimensionality, the three latent factors were strongly correlated ($r = 0.832$ – 0.885). Bifactor analysis further demonstrated pronounced general-factor dominance: explained common variance was 0.830, omega total was 0.979, and omega hierarchical was 0.959. General-factor loadings ranged from 0.556 to 0.819 (median, 0.686), indicating that most reliable variance in the candidate item bank was attributable to a common empowerment dimension.

Table 2. Structural and dimensionality evidence for the 42-item SWEI-SD

Analysis/model	χ^2	df	CFI	TLI	RMSEA	SRMR
One-factor CFA	4626.3	819	0.743	0.73	0.099	0.062
Two-factor CFA	4272.7	818	0.767	0.754	0.094	0.058
EFA-derived three-factor CFA	3799	816	0.798	0.787	0.087	0.053
Original seven-correlated-factor	4012.5	798	0.783	0.766	0.092	0.054
Second-order seven-factor	4239.2	812	0.768	0.754	0.094	0.058
Bifactor: general + 7 specific factors	3662.8	777	0.805	0.784	0.088	0.055
EFA-informed three-factor + cross-loadings	3571	804	0.813	0.8	0.085	0.049
Bifactor diagnostic	Value					
Explained common variance	0.830					
Omega total	0.979					
Omega hierarchical	0.959					
Median general-factor loading	0.686					
General-factor loading range	0.556–0.819					

Note. χ^2 , CFI, TLI, and RMSEA are approximate statistics from the polychoric/DWLS-style implementation.

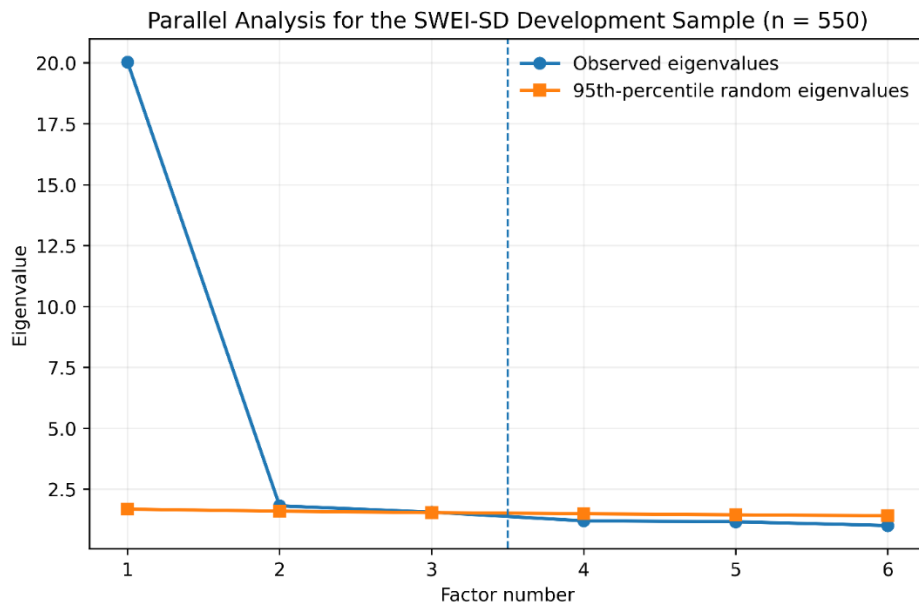


Figure 1. Parallel analysis and scree plot for the 42-item SWEI-SD in the development sample (n = 550). The first three observed eigenvalues exceeded the corresponding 95th-percentile random eigenvalues, supporting retention of three exploratory factors. The fourth and subsequent observed eigenvalues fell below their random reference values.

3.3 Item-response characteristics and measurement precision

The 42-item bank was subsequently calibrated using the graded response model in the full sample. All items demonstrated satisfactory discrimination, with parameters ranging from 1.218 to 2.345 (median, 1.793), and all response thresholds were correctly ordered. The 42-item model achieved EAP reliability of 0.963 and a mean posterior standard error of 0.170. Test information peaked at $\theta = -0.15$ (information = 39.79), and conditional SEM remained ≤ 0.30 between approximately $\theta = -2.48$ and 2.20.

Item-level misfit was limited. Only D4_3 and D7_3 exceeded the prespecified category-fit RMSD threshold of 0.08. Residual dependence represented the more important source of inefficiency: 25 item pairs had adjusted Q3 > 0.20 , largely involving conceptually adjacent or redundant items. Thus, the principal rationale for shortening the instrument was reduction of local redundancy rather than elimination of poorly discriminating items.

DIF analyses showed substantial item-level stability across the four prespecified comparisons. No FDR-significant DIF was detected between urban and rural participants. Displacement status produced two statistically significant items, but only D2_1 demonstrated reproducible practically non-negligible DIF. Statistically detectable effects in the marital-status and education comparisons remained below the $\Delta\text{McFadden } R^2 = 0.01$ practical threshold. Loading invariance was supported for displacement, marital status, and education. Strict urban–rural metric invariance was mixed; however, after six loadings were freed, partial metric invariance was supported ($\Delta\text{CFI} = -0.008$, $\Delta\text{RMSEA} \approx 0$, $\Delta\text{SRMR} = 0.009$) (Table 3).

Table 3. IRT performance, DIF, and measurement-invariance findings

Psychometric property	Finding
42-item discrimination,	1.218–2.345; median 1.793

a	
Threshold ordering	Ordered for all 42 items
EAP reliability	0.963
Mean posterior SE	0.17
Maximum test information	39.79 at $\theta = -0.15$
θ range with SEM ≤ 0.30	-2.48 to 2.20
RMSD > 0.08	D4_3, D7_3
Adjusted Q3 > 0.20	25 item pairs
Urban vs rural DIF	0 FDR-significant; 0 practical DIF
Displacement DIF	2 FDR-significant; 1 practical DIF (D2_1)
Marital-status DIF	1 FDR-significant; 0 practical DIF
Education DIF	2 FDR-significant; 0 practical DIF
Urban-rural invariance	Partial metric invariance supported
Displacement invariance	Metric supported; item/threshold partial
Marital-status invariance	Supported in practical terms
Education invariance	Supported in practical terms

3.4 Development and performance of the SWEI-SD-28

Integration of EFA, CFA/bifactor findings, IRT discrimination and information, category fit, local dependence, DIF, measurement invariance, redundancy, and substantive coverage resulted in removal of 14 items. Four items were retained from each conceptual domain, yielding the 28-item SWEI-SD-28. D2_1 was removed because it was the only item with reproducible practically important DIF; other removals

Recalibration of the 28 retained items showed satisfactory discrimination throughout, ranging from 1.303 to 2.282, with ordered thresholds for all items (Table 4). EAP reliability remained high at 0.944. Maximum information was 26.77 at approximately $\theta = -0.20$, and SEM remained ≤ 0.30 from $\theta = -2.20$ to 1.90 (Figure 2). The reduction substantially decreased residual dependence, from 25 adjusted-Q3 pairs >0.20 in the 42-item bank to eight in the final form. Repeated DIF analysis identified two FDR-significant findings, but neither met the practical $\Delta R^2 \geq 0.01$ criterion.

The shortened form closely reproduced scores from the original bank. Pearson correlation between 28-item and 42-item EAP estimates was 0.990, Spearman correlation was 0.987, and the mean absolute latent-score difference was 0.109 θ units. Thus, removal of one-third of the original items produced limited loss of information while materially reducing respondent burden and local redundancy. Of the 1,100 participants, 1,019 (92.6%) met the final scoring eligibility rule. Among eligible women, the mean 0–100 SWEI-SD-28 score was 54.41 (SD = 21.54), with a median of 56.09 and IQR of 38.20–67.37 (Figure 3).

Table 4. Final SWEI-SD-28 items and graded-response model parameters

Domain	Item	Brief content	a	b1	b2	b3
D1	D1_1	Important decisions about own life	1.865	−1.058	0.13	1.252
D1	D1_2	Express opinion when disagreeing	1.488	−1.828	−0.512	0.96
D1	D1_5	Deal with unexpected difficulties	1.821	−1.157	−0.158	0.929
D1	D1_6	Refuse seriously harmful actions	1.88	−1.217	−0.200	0.866
D2	D2_3	Say in own work/income activities	1.369	−1.472	−0.267	1.075
D2	D2_4	Say in dependents' education/welfare	1.676	−1.350	−0.257	0.829
D2	D2_5	Say in household residence/relocation	1.59	−1.213	−0.043	1.02
D2	D2_6	Decisions about visiting relatives/friends	1.927	−1.163	−0.150	0.92
D3	D3_1	Obtain money for essential needs	1.678	−0.991	0.294	1.377
D3	D3_2	Say in use of money earned/received	1.823	−1.481	−0.390	0.73
D3	D3_5	Decisions about owned/joint assets	1.736	−1.166	−0.129	1.014
D3	D3_6	Decide to start/join income activity	1.987	−0.993	−0.123	0.817

D4	D4_1	Go to healthcare when needed	1.944	−1.011	0.112	1.171
D4	D4_2	Go to market/bank/services	1.64	−1.511	−0.393	0.747
D4	D4_5	Travel for education/training/work	1.619	−1.098	−0.103	1.057
D4	D4_6	Move self/dependents to safety	1.882	−1.029	−0.174	0.817
D5	D5_2	Use communication device when needed	1.93	−1.610	−0.503	0.69
D5	D5_3	Communicate privately digitally	1.581	−1.448	−0.436	0.728
D5	D5_5	Decide on education/training opportunities	1.95	−1.112	−0.214	0.782
D5	D5_6	Identify information/assistance sources	1.922	−1.143	−0.143	0.96
D6	D6_1	Join women's/professional/community group	2.161	−0.681	0.25	1.153
D6	D6_2	Express views in group discussions	1.674	−1.731	−0.561	0.748
D6	D6_4	Approach leaders/organizations	1.303	−1.399	−0.231	1.034
D6	D6_6	Accept leadership responsibility	1.71	−1.208	−0.082	0.943
D7	D7_1	Say in own healthcare decisions	2.181	−0.957	0.058	1.07
D7	D7_2	Bodily consent respected	1.596	−1.936	−0.746	0.513
D7	D7_4	Say in pregnancy timing/family planning	1.781	−1.192	−0.203	0.825
D7	D7_6	Allocate time across personal activities	2.282	−0.905	−0.086	0.83

Overall SWEI-SD-28 performance: EAP reliability = 0.944; mean posterior SE = 0.220; maximum information = 26.766; correlation with 42-item EAP = 0.990; adjusted-Q3 pairs >0.20 = 8; practically important DIF findings = 0; scoring eligibility = 92.6%.

Test Information: 42-item Bank vs SWEI-SD-28

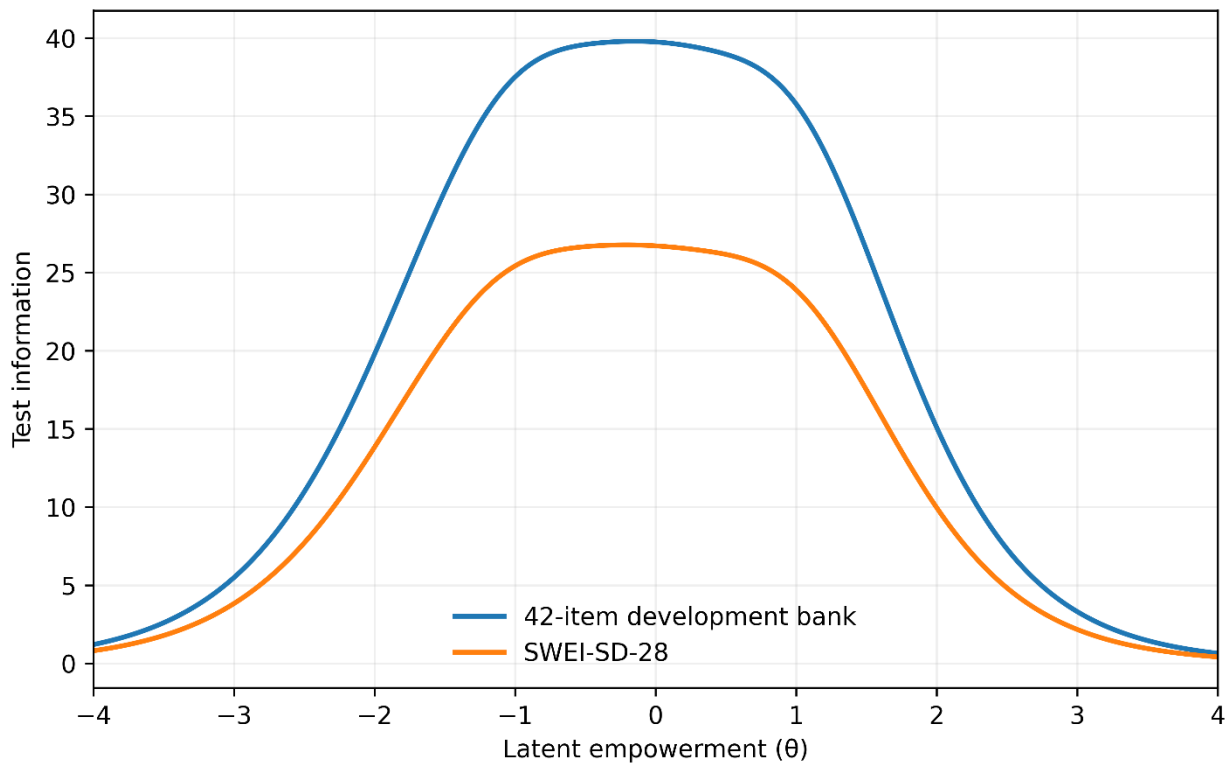


Figure 2. Test information functions for the original 42-item SWEI-SD and the final SWEI-SD-28. The 28-item form retained high information across the central range of the latent empowerment continuum despite removal of 14 items. Conditional SEM remained ≤ 0.30 from approximately $\theta = -2.20$ to 1.90.

Figure 4.26. Distribution of final SWEI-SD-28 scores

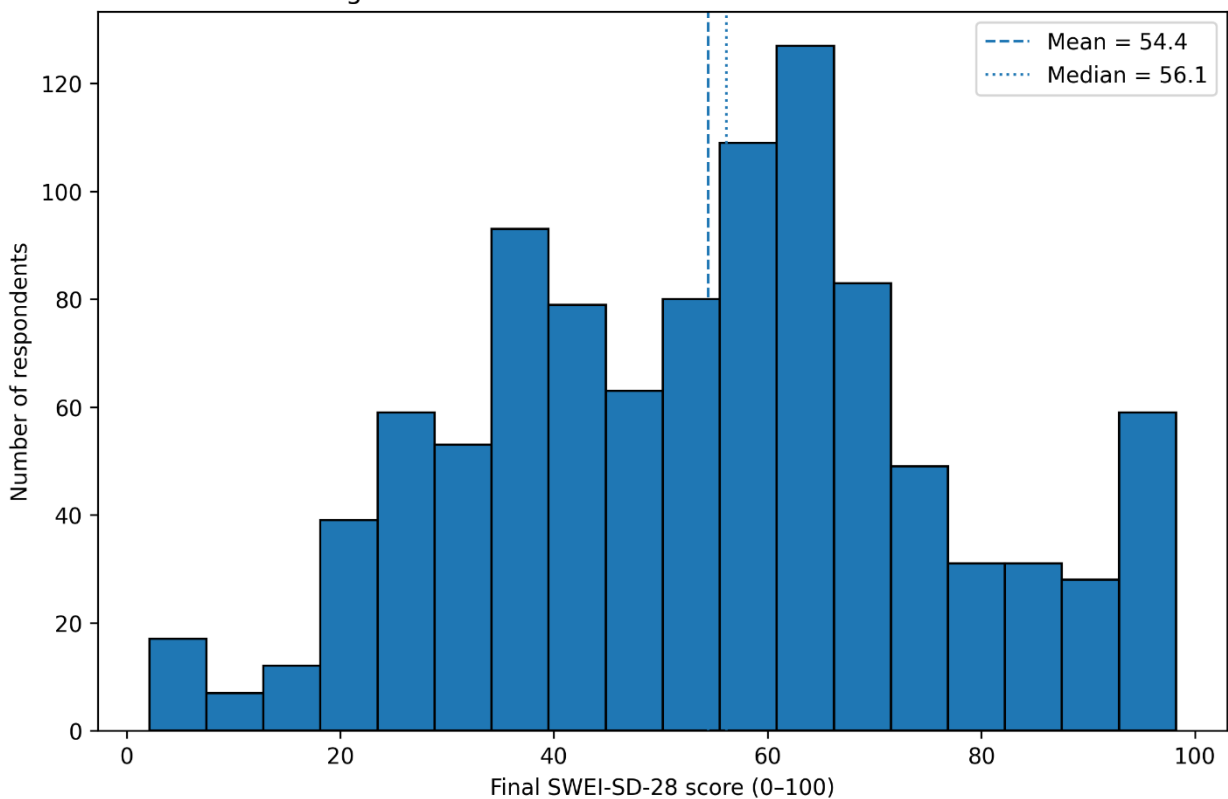


Figure 3. Distribution of SWEI-SD-28 scores among respondents meeting the primary scoring criteria ($n = 1,019$). SWEI-SD-28 scores are presented on the 0-100 expected-score metric. Mean score was 54.41 ($SD = 21.54$), median 56.09, and interquartile range

4. DISCUSSION

This study developed and psychometrically evaluated the Sudan Women's Empowerment Index and produced a 28-item final form, the SWEI-SD-28, designed to measure a broad latent construct of women's empowerment while preserving substantive coverage across seven conceptually important domains. The principal finding is that empowerment in this sample was empirically multidimensional but strongly organized around a dominant general factor. Exploratory analysis favored three factors rather than the seven domains originally specified for content development, while the three confirmatory factors were highly correlated ($r = 0.832\text{--}0.885$). The bifactor results were especially informative: the general factor accounted for 83.0% of common variance, with $\omega_{\text{total}} = 0.979$ and $\omega_{\text{hierarchical}} = 0.959$, indicating that most reliable variance in the candidate item bank was attributable to a broad empowerment dimension rather than to statistically independent domain-specific constructs. This pattern is consistent with contemporary bifactor guidance, which emphasizes that multidimensional content can coexist with sufficient general-factor dominance to justify an overall score (Reise, 2012; Rodriguez et al., 2016). Accordingly, the seven SWEI-SD domains are best interpreted as conceptually meaningful areas of agency within an overarching empowerment construct rather than as seven independent reflective subscales.

This distinction is theoretically important. Kabeer (1999) conceptualized empowerment as the expansion of the ability to make strategic life choices, embedded within interrelated processes of resources, agency, and achievements. The SWEI-SD was deliberately centered on women's capacity to decide, act, refuse, participate, communicate, move, and seek assistance rather than treating resources or socioeconomic circumstances themselves as empowerment. This helps address a recurrent challenge identified in the literature: measures labeled as women's empowerment often combine agency, resources, outcomes, and demographic status within the same index, creating ambiguity about what the resulting score represents. Costa et al. (2023), in a review of individual-level empowerment measures, documented marked heterogeneity in domains, definitions, and analytical approaches and highlighted the tension between context-specific validity and cross-population comparability. The SWEI-SD contributes to this literature by explicitly separating the latent empowerment construct from contextual resources while retaining content relevant to Sudanese women's lived circumstances.

The factor-analytic findings also caution against forcing theoretically defined domains into a rigid simple structure. None of the strict confirmatory models achieved conventionally strong global fit across all approximate indices, and the EFA-informed three-factor model with prespecified cross-loadings provided the closest residual reproduction of the data ($\text{CFI} = 0.813$, $\text{RMSEA} = 0.085$, $\text{SRMR} = 0.049$). Rather than indicating instrument failure, the combined evidence suggests that empowerment-related behaviors overlap substantially across life domains. Financial agency, mobility, bodily autonomy, access to information, and civic participation are unlikely to operate as fully isolated capacities in daily life. Similar measurement complexity has been observed in other empowerment instruments. Psychometric evaluation of the project-level Women's Empowerment in Agriculture Index (pro-WEAI) showed that some item sets required refinement because theoretical domains did not always satisfy assumptions

of one-dimensionality or local independence (Yount et al., 2019). This supports an approach in which theoretical content coverage and empirical structure are integrated rather than requiring one-to-one correspondence between conceptual domains and latent factors.

The item-response findings provide complementary evidence that the original item bank functioned well despite this structural complexity. All 42 candidate items demonstrated satisfactory discrimination and ordered response thresholds, while the full bank achieved an EAP reliability of 0.963 and high conditional precision across much of the latent continuum. The main psychometric inefficiency was localized residual dependence among conceptually similar items rather than poor discrimination or dysfunctional categories. This distinction is important because modern scale refinement should not rely solely on internal consistency or item-total correlation. Item response theory allows item discrimination, threshold placement, information, and local dependence to be considered simultaneously, thereby identifying items that may be statistically adequate but redundant (Stover et al., 2019). In the present study, this broader evidence base allowed shortening decisions to focus on unique contribution, precision, subgroup stability, and conceptual coverage.

The transition from 42 to 28 items is therefore one of the strongest findings. Removal of 14 items reduced the number of adjusted-Q3 residual-dependence pairs above 0.20 from 25 to 8 while preserving excellent agreement with the original bank: Pearson $r = 0.990$, Spearman $\rho = 0.987$, and the mean absolute latent-score difference was only 0.109 theta units. The final form retained high EAP reliability (0.944) and useful measurement precision from approximately $\theta = -2.20$ to 1.90. Thus, a one-third reduction in instrument length produced little loss of respondent ordering or latent-score information while materially reducing redundancy and expected administration burden. This result is consistent with contemporary psychometric scale-development principles, in which shorter forms are preferred when they maintain information and content coverage without sacrificing validity (Prinsen et al., 2018; Stover et al., 2019). In conflict-affected and resource-constrained field settings, this efficiency may be particularly valuable because respondent fatigue, privacy concerns, and time constraints can directly affect data quality.

Subgroup-equivalence findings further strengthen the final instrument. In the 42-item bank, only one item, D2_1, showed reproducible differential item functioning of practical magnitude, specifically by displacement status, and that item was removed during refinement. No urban-rural DIF was detected, whereas marital-status and educational differences that reached statistical significance remained practically negligible. Strict urban-rural metric invariance was mixed, but partial metric invariance was supported after a limited number of loadings were freed. Following recalibration of the SWEI-SD-28, two FDR-significant DIF effects remained, but neither exceeded the prespecified practical threshold. These results support broad subgroup comparability while appropriately stopping short of claiming complete invariance. Measurement invariance is especially important when group comparisons are central to substantive interpretation, because high reliability alone does not demonstrate that respondents from different groups interpret or use items equivalently (Putnick & Bornstein, 2016). The combined use of multi-group invariance and item-level DIF therefore represents an important strength of the present validation strategy.

The SWEI-SD-28 should also be understood in relation to existing empowerment indices rather than as a replacement for them. SWPER, for example, was developed

from Demographic and Health Survey items and offers substantial advantages for standardized cross-national monitoring, particularly in settings where existing survey data are the principal resource (Ewerling et al., 2017, 2020). By contrast, the SWEI-SD was designed as a context-grounded measurement instrument for adult Sudanese women and includes dimensions that are particularly salient in contemporary Sudan, including mobility under insecurity, digital and information agency, collective and civic participation, bodily autonomy, and control over time. Its contribution lies in providing a psychometrically refined measure that can capture these broader expressions of agency while remaining sufficiently concise for field application. The strong general factor supports an overall score as the primary inferential outcome, whereas the seven conceptual domains retain value as descriptive profiles of where relative constraints or strengths may be concentrated.

Finally, the observed mean SWEI-SD-28 score of 54.41 should be interpreted as a descriptive property of this validation sample rather than as a national benchmark. The sampling strategy intentionally emphasized psychometric coverage and subgroup representation, not nationally weighted prevalence estimation. For the same reason, the present data do not justify normative low, moderate, or high cut-points. Future thresholds, if desired for programmatic interpretation, should be anchored to external criteria, meaningful outcomes, or representative reference distributions rather than selected statistically from the present score distribution. This conservative interpretation is consistent with measurement standards that distinguish reliable score production from the separate task of establishing clinical, policy, or normative decision thresholds (Prinsen et al., 2018).

4.1 Limitations

Several limitations should guide interpretation. First, the study was cross-sectional; therefore, the analyses establish internal structure, precision, and subgroup comparability but do not demonstrate responsiveness to change or longitudinal stability. Second, the sampling strategy prioritized psychometric coverage and subgroup representation rather than nationally weighted population estimation, so the observed mean SWEI-SD-28 score should not be interpreted as a national prevalence estimate of empowerment. Third, CFA and invariance testing used an approximate polychoric/DWLS-style implementation rather than full WLSMV; consequently, global fit statistics require cautious interpretation and independent replication. Fourth, external criterion validity and predictive validity were not established in this paper. Fifth, although translation, expert review, cognitive interviewing, and pilot testing were undertaken, retained quantitative expert ratings were unavailable and formal content-validity indices could not be calculated. Finally, partial urban-rural metric invariance indicates that some parameter heterogeneity remains, supporting continued evaluation before high-stakes subgroup comparisons or normative classification.

4.2 Implications

The SWEI-SD-28 provides a practical framework for measuring women's empowerment in research and program evaluation in Sudan while maintaining a clear distinction between empowerment itself and its contextual determinants. Its strong general factor supports use of a single overall score as the primary outcome, whereas domain profiles can be used descriptively to identify relative areas of constraint or strength. The reduction from 42 to 28 items should improve feasibility for field surveys, humanitarian assessments, and intervention studies without materially sacrificing information. Future research should prioritize external validation in independent Sudanese samples, full

WLSMV-based invariance testing, test-retest reliability, responsiveness to intervention-related change, and associations with theoretically relevant external outcomes. Longitudinal studies should determine whether changes in SWEI-SD-28 scores correspond to meaningful changes in women's agency. Cross-language and administration-mode validation will also be necessary before extending the instrument beyond the Sudanese Arabic version or using it routinely in digital self-administration.

5. CONCLUSION

The SWEI-SD-28 offers a contextually grounded and psychometrically rigorous measure of women's empowerment among Sudanese women. Starting from 42 candidate items across seven conceptual domains, integrated factor analysis, bifactor modelling, graded-response item response theory, local-dependence assessment, differential item functioning, and measurement-invariance testing supported a shorter 28-item instrument with a dominant general empowerment factor. The final form retained high reliability and measurement precision, substantially reduced residual redundancy, demonstrated no practically meaningful DIF after recalibration, and reproduced the latent ordering of the original 42-item bank with very high concordance. These findings support use of the overall SWEI-SD-28 score as the primary inferential measure, with domain-level results interpreted as secondary descriptive profiles rather than independent subscales. The instrument provides a strong foundation for future epidemiological, programmatic, and longitudinal research on women's agency in Sudan, while external replication and longitudinal validation remain necessary before normative cut-points or broad population-level interpretations are established.

Declarations

Consent for publication

Not applicable.

Availability of data and material

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Generative AI statement

The authors used generative AI solely for language refinement. All scientific interpretation, analysis, and conclusions were independently verified by the authors.

Funding

NA

Acknowledgements

NA

REFERENCES

1. Alkire, S., Meinzen-Dick, R., Peterman, A., Quisumbing, A., Seymour, G., & Vaz, A. (2013). The Women's Empowerment in Agriculture Index. *World Development*, 52, 71-91. <https://doi.org/10.1016/j.worlddev.2013.06.007>
2. Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24), 3186-3191. <https://doi.org/10.1097/00007632-200012150-00014>

3. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
4. Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504. <https://doi.org/10.1080/10705510701301834>
5. Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233-255. https://doi.org/10.1207/s15328007sem0902_5
6. Costa, J. C., Saad, G. E., Hellwig, F., Maia, M. F. S., & Barros, A. J. D. (2023). Measures of women's empowerment based on individual-level data: A literature review with a focus on the methodological approaches. *Frontiers in Sociology*, 8, 1231790. <https://doi.org/10.3389/fsoc.2023.1231790>
7. Ewerling, F., Lynch, J. W., Victora, C. G., van Eerdewijk, A., Tyszler, M., & Barros, A. J. D. (2017). The SWPER index for women's empowerment in Africa: Development and validation of an index based on survey data. *The Lancet Global Health*, 5(9), e916-e923. [https://doi.org/10.1016/s2214-109x\(17\)30292-9](https://doi.org/10.1016/s2214-109x(17)30292-9)
8. Ewerling, F., Raj, A., Victora, C. G., Hellwig, F., Coll, C. V. N., & Barros, A. J. D. (2020). SWPER Global: A survey-based women's empowerment index expanded from Africa to all low- and middle-income countries. *Journal of Global Health*, 10(2), 020434. <https://doi.org/10.7189/jogh.10.020434>
9. Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, 18(4), 454-474. <https://doi.org/10.1037/a0030005>
10. Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7(2), 191-205. <https://doi.org/10.1177/1094428104263675>
11. Kabeer, N. (1999). Resources, agency, achievements: Reflections on the measurement of women's empowerment. *Development and Change*, 30(3), 435-464. <https://doi.org/10.1111/1467-7660.00125>
12. Nahar, S., & Mengo, C. W. (2022). Measuring women's empowerment in developing countries: A systematic review. *Journal of International Development*, 34(2), 322-333. <https://doi.org/10.1002/jid.3594>
13. Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90. <https://doi.org/10.1016/j.dr.2016.06.004>
14. Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696. <https://doi.org/10.1080/00273171.2012.715555>
15. Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(2), 137-150. <https://doi.org/10.1037/met0000045>
16. Sabahelzain, M. M., Gamal Eldin, A., Babiker, S., Kabiru, C. W., & Eltayeb, M. (2019). Decision-making in the practice of female genital mutilation or cutting in Sudan: A cross-sectional study. *Global Health Research and Policy*, 4, 5. <https://doi.org/10.1186/s41256-019-0096-0>
17. Stover, A. M., McLeod, L. D., Langer, M. M., Chen, W.-H., & Reeve, B. B. (2019). State of the psychometric methods: Patient-reported outcome measure development and

refinement using item response theory. *Journal of Patient-Reported Outcomes*, 3, 50. <https://doi.org/10.1186/s41687-019-0130-5>

18. Vizheh, M., Muhidin, S., Behboodi Moghadam, Z., & Zareiyan, A. (2021). Women empowerment in reproductive health: A systematic review of measurement properties. *BMC Women's Health*, 21(1), 424. <https://doi.org/10.1186/s12905-021-01566-0>
19. Yount, K. M., Cheong, Y. F., Maxwell, L., Heckert, J., Martinez, E. M., & Seymour, G. (2019). Measurement properties of the project-level Women's Empowerment in Agriculture Index. *World Development*, 124, 104639. <https://doi.org/10.1016/j.worlddev.2019.104639>
20. Yount, K. M., Quisumbing, A., Meinzen-Dick, R., Malapit, H., Hassan, M. Z., Kanyanda, S. S. E., Sharma, S., Hassan, M. I., Heckert, J., Paz, F., Pokhrel, P., Vundru, W. D., Doss, C., Seymour, G., Moylan, H., Kilic, T., Myers, E., Faas, S., Cheong, Y. F., & Sinharoy, S. S. (2026). Women's Empowerment Metric for National Statistical Systems (WEMNS): Development and psychometric assessment of a face-to-face survey module. *PLOS ONE*, 21(5), e0345742. <https://doi.org/10.1371/journal.pone.0345742>