Environment and Social Psychology (2025) Volume 10 Issue 11 ISSN: 2424-8975 (O)
doi: 10.59429/esp.v10i11.3570 2424-7979 (P)

RESEARCH ARTICLE

Gender-based analysis of teacher empowerment scale: Examining factor structure and rasch model fit in higher education

PANELA, Teody Lester V.*

Northwest Samar State University, Philippines

*** Corresponding author:** PANELA, Teody Lester V., teodylester.panela@nwssu.edu.ph.

ABSTRACT

This study examined the measurement invariance and psychometric properties of the Teacher Empowerment Scale across gender groups in higher education. Using Rasch analysis, 86 items spanning three factors (fostering continuous improvement, teaching ownership and freedom, and work climate and conditions) were analyzed with data from 968 faculty members. Results demonstrated excellent model fit (mean infit/outfit MNSQ ≈1.00) and high reliability (α=0.90-0.93) across all factors. Differential item functioning analysis revealed minimal gender-based variations, with only 5 items in factor 1, 4 items in factor 2, and none in factor 3 showing significant differences. The scale provides fair assessment of teacher empowerment constructs for both male and female educators, supporting previous research findings. Recommendations include implementing the scale confidently while attending to items with differential functioning; refining these items to enhance gender neutrality; extending validation research to additional demographic variables; conducting longitudinal studies; and utilizing the three-factor structure for designing targeted interventions. This research addresses existing gaps regarding gender considerations in scale development, advancing equitable assessment instruments for higher education settings.

Keywords: measurement invariance; item response theory; construct validity; faculty development; organizational climate

1. Introduction

The accurate measurement of teacher empowerment across different demographic groups remains a fundamental concern for ensuring fair and effective educational assessment[1]. Although previous research has documented the significance of teacher empowerment in educational settings, questions persist regarding whether gender influences how measurement tools function. Establishing whether measurement instruments operate equivalently across gender groups is not merely a technical consideration—It determines whether comparisons between male and female teachers yield valid interpretations and whether institutional policies based on such assessments are justifiable.

Emerging evidence indicates that gender shapes how teachers experience and express empowerment in their professional contexts[2]. Research by Akpan and Ayinmoro[3] revealed that middle-aged, female, and married teachers demonstrated higher levels of empowerment compared to other demographic groups.

ARTICLE INFO
Received: 24 March 2025 | Accepted: 15 October 2025 | Available online: 20 November 2025

CITATION
PANELA TLV. Gender-based analysis of teacher empowerment scale: Examining factor structure and rasch model fit in higher education. *Environment and Social Psychology* 2025; 10(11): 3570 doi:10.59429/esp.v10i11.3570

Correspondingly, Berhanu[4] observed that female teachers with higher educational attainment exhibited greater exercise of empowerment capabilities. These patterns suggest that gender may influence not only empowerment levels but potentially how teachers interpret and respond to measurement items assessing empowerment constructs. Despite these observations, systematic examination of measurement invariance across gender groups remains conspicuously absent from the teacher empowerment literature, particularly within higher education contexts.

Teacher empowerment represents a multidimensional construct encompassing various facets of teachers' professional lives. Ahmadi and Arief[6] characterized empowered teachers as possessing the capacity to freely exercise professional judgment, effectively navigate challenges, and adapt to structural changes within educational institutions. The construct encompasses multiple dimensions including leadership and decision-making authority, opportunities for professional development, professional reputation and standing, teacher efficacy beliefs, institutional autonomy, collegiality among peers, and overall work climate [5]. This conceptual complexity underscores the importance of rigorous psychometric examination to ensure that instruments capture these dimensions accurately across different respondent groups.

Webber and Nickel[9] documented that teacher empowerment processes vary considerably based on individual characteristics, with gender potentially serving as a moderating influence. However, existing measurement research has largely overlooked this possibility. Short and Rinehart[10] developed the School Participant Empowerment Scale with six dimensions but conducted no systematic examination of whether the scale functioned equivalently for male and female teachers. Subsequent validation studies have perpetuated this limitation. Golle et al.[7] verified the validity and reliability of this instrument among teachers of gifted students but treated gender only as a descriptive variable without testing measurement invariance. Similarly, Gomes et al.[8] revalidated the scale for science and mathematics teachers, deriving six revised factors, yet failed to assess whether these factors maintained equivalent measurement properties across gender groups.

This oversight becomes particularly problematic in higher education settings, where gender dynamics intersect with unique organizational structures, professional expectations, and empowerment processes. Unlike primary and secondary education environments where some gender research exists, tertiary education presents distinctive characteristics—including greater emphasis on research productivity, different governance structures, and varied collegial relationships—that may interact with gender in ways that affect measurement (Navarro-González et al., 2024). Van Woerden et al.[11] noted that higher education institutions involve distinct forms of teamwork, collaboration, and professional autonomy that faculty members may experience differently based on gender. Yet no studies have examined whether teacher empowerment scales capture these experiences equivalently across gender lines in tertiary settings.

The critical gap in the literature centers on three interconnected issues[12]. First, despite documented gender differences in empowerment levels and processes, no studies have systematically tested whether widely used teacher empowerment scales demonstrate measurement invariance across gender groups. This absence means that researchers cannot confidently determine whether observed differences between male and female teachers reflect genuine empowerment disparities or measurement artifacts stemming from differential item functioning. Second, the specific context of higher education remains understudied regarding teacher empowerment measurement, creating uncertainty about whether instruments validated in K-12 settings function appropriately for faculty populations. Third, methodologically, most validation studies have relied on classical test theory approaches that lack the item-level sensitivity needed to detect subtle forms of measurement bias that may disadvantage one gender group. As Celik et al.[1] emphasized,

without rigorous examination of measurement equivalence across demographic groups, educational assessments risk producing misleading conclusions that undermine both research validity and policy decisions.

Addressing this gap requires methodological approaches capable of detecting item-level bias. The relationship between gender and psychometric properties of measurement instruments constitutes a critical consideration in scale development and validation. Katsikeas et al.[13] demonstrated that incorporating gender considerations during scale development yields improved measurement outcomes and enhanced construct validity. Pan et al.[14] emphasized that accounting for gender during scale development is essential for accurately measuring responses and designing gender-responsive interventions. However, these methodological insights have not been systematically applied to teacher empowerment scales in higher education contexts. The absence of such examination leaves researchers, administrators, and policymakers uncertain whether their assessments of faculty empowerment are valid or whether conclusions drawn from comparing male and female faculty are defensible.

The present study addresses these interconnected gaps by conducting a comprehensive analysis of the Teacher Empowerment Scale's psychometric properties across gender groups within higher education contexts. Employing Rasch analysis provides sophisticated methods for detecting differential item functioning—Instances where items operate differently for equally empowered male and female faculty members—While simultaneously evaluating overall measurement quality. This analytical approach offers item-level diagnostic information unavailable through traditional validation methods, enabling identification of specific items that may introduce bias. The findings will advance both theoretical understanding of how gender intersects with empowerment measurement and practical application by informing development of more equitable assessment tools. Ultimately, establishing measurement invariance across gender groups will enable more valid comparisons, support evidence-based institutional policies, and facilitate targeted interventions that genuinely address empowerment disparities rather than measurement artifacts in tertiary educational settings.

2. Purpose of the present study

This study aimed to examine the measurement invariance and psychometric properties of the Teacher Empowerment Scale across gender groups using Rasch analysis. The research addressed the following objectives:

1. To evaluate the overall Rasch fit statistics and reliability coefficients of the scale's three factors
2. To examine differential item functioning across gender groups
3. To assess the scale's measurement invariance and construct validity.

3. Research design

The study employed Rasch analysis to examine the psychometric properties and measurement invariance of the Teacher Empowerment Scale [16]. This analytical approach was selected to address well-documented limitations of Classical Test Theory (CTT), which include sample-dependent calibrations, problematic assumptions about equal intervals in Likert-type response formats, and inability to evaluate whether response categories function optimally across different respondent groups. The Rasch model proves particularly suitable for this investigation because it enables sophisticated item-level examination of how measurement instruments function across demographic groups, offering advantages over traditional approaches that assume equal item discrimination parameters[15]. Unlike CTT approaches that produce

ordinal-level estimates, Rasch measurement generates interval-level measures with the property of specific objectivity, meaning that item difficulties can be estimated independently of the particular sample tested and person abilities can be estimated independently of the particular items administered.

The analysis focused specifically on testing measurement invariance across teachers' gender. Measurement invariance addresses whether a construct maintains equivalent meaning and measurement properties across different groups, which constitutes a prerequisite for valid between-group comparisons [19]. The analytical sequence began with establishing a configural model in which all dimensions were freely estimated across gender groups. Adequate fit of a configural model provides initial evidence that the same latent factors define the construct for both male and female teachers, establishing the baseline for subsequent invariance testing. Subsequently, a metric invariance model was tested by constraining factor loadings to equality across gender groups, examining whether items contribute equivalently to the underlying construct for both groups[17]. When configural and metric invariance models demonstrated acceptable fit, additional constraints were imposed to test scalar invariance, which assesses whether item intercepts differ systematically between groups. Failure to establish scalar invariance would indicate that observed mean differences between male and female respondents reflect measurement bias rather than true group differences in the underlying empowerment construct[18].

However, several methodological limitations warrant acknowledgment regarding the invariance testing approach employed. First, the sequential constraint-based approach assumes correct specification of the baseline configural model; if this initial model exhibits misspecification, subsequent invariance tests may yield misleading conclusions regardless of whether constraints are supported statistically. Second, the study's reliance on traditional chi-square difference testing for evaluating invariance remains sensitive to sample size, with large samples potentially flagging trivial non-invariance as statistically significant despite negligible practical implications for measurement comparability. Third, the combination of Rasch analysis and multi-group confirmatory factor analysis represents somewhat redundant methodological approaches to invariance assessment, as both methods address similar research questions through different psychometric frameworks. The study does not articulate how conflicting results between these two approaches would be reconciled should they yield discrepant conclusions about measurement equivalence.

The analysis concentrated on three established factors of the Teacher Empowerment Scale: fostering continuous improvement (comprising 51 items), teaching ownership and freedom (comprising 32 items), and work climate and conditions (comprising 3 items). This factor structure emerged from prior exploratory and confirmatory factor analyses, providing an empirically-supported foundation for the current gender-based invariance testing. However, the substantial imbalance in items across factors—with the first two factors containing substantially more items than the third—presents potential limitations for model estimation and interpretation. Factors with fewer indicators may demonstrate less stable parameter estimates and reduced reliability, particularly when subjected to equality constraints across groups[20]. The three-item work climate factor may prove especially problematic for invariance testing, as the minimum of three indicators per factor represents the lower bound for factor identification and leaves no degrees of freedom for evaluating factor-specific misspecification. Furthermore, the large number of items in the first two factors (51 and 32 respectively) raises questions about whether the instrument may benefit from item reduction to enhance efficiency without sacrificing measurement precision, particularly given concerns about respondent burden in survey research.

4. Participants and sampling

Data were collected from 968 higher education faculty members from state universities and colleges in Region VIII of the Philippines. The sample size substantially exceeds contemporary recommendations for factor analytic procedures and measurement invariance testing. While classical guidelines suggested 300-450 participants for acceptable pattern comparability in factor analysis, more recent simulation studies indicate that adequate sample sizes depend on multiple factors including communalities, number of indicators per factor, and model complexity[20]. Contemporary research demonstrates that sample size requirements for confirmatory factor analysis vary considerably based on model characteristics, with simple models requiring as few as 200 participants while more complex models may require 500 or more for stable parameter estimation [22]. For Rasch analysis and multi-group invariance testing specifically, current methodological literature recommends minimum sample sizes of 250-500 per group for stable parameter estimation and adequate statistical power to detect meaningful differences in item functioning[21].

Nevertheless, several sampling limitations warrant explicit acknowledgment. First, the study does not report sample sizes disaggregated by gender, representing a critical omission that prevents evaluation of whether adequate statistical power existed for detecting meaningful differences in item functioning between male and female faculty. If the gender distribution proved highly unbalanced—for instance, with substantially fewer male than female participants or vice versa—the smaller group may have insufficient sample size for reliable parameter estimation regardless of the total sample size. Second, the stratification variables employed in the proportionate stratified random sampling procedure were not explicitly specified in the methodology, leaving unclear whether stratification accounted for potentially important characteristics such as institutional size, disciplinary composition, faculty rank distribution, or years of teaching experience that might influence empowerment experiences differently for male and female faculty members.

The sampling procedure employed proportionate stratified random sampling to ensure adequate representation across institutions. This approach was selected because it enhances population coverage by providing researchers greater control over subgroup representation, reducing sampling error compared to simple random sampling while maintaining probability-based selection[23]. Stratified sampling proves particularly valuable when population subgroups differ substantially on the characteristic being measured, as it ensures that all important subgroups are adequately represented in the final sample[24]. Computer-generated selection maintained randomness within stratified groupings, thereby minimizing potential selection bias that could arise from systematic selection procedures. However, the stratification approach introduces additional complexity to data analysis, as stratified samples require weighted analyses or multilevel modeling approaches to appropriately account for the sampling design, yet the study does not indicate whether such design-appropriate analytical adjustments were implemented.

Participants represented diverse academic disciplines, ranks, and years of teaching experience, ostensibly enhancing generalizability of findings within the regional higher education context. However, critical limitations constrain the generalizability and interpretability of study findings. First, restriction to Region VIII substantially limits generalizability to faculty in other Philippine regions where institutional structures, resource availability, cultural norms regarding gender roles, and professional empowerment dynamics may differ markedly. Generalization to international higher education contexts proves even more problematic, as faculty governance structures, promotion systems, gender equity policies, and organizational cultures vary dramatically across national educational systems. Second, the study provides no descriptive statistics for demographic variables disaggregated by gender, precluding assessment of whether male and female subsamples were comparable on characteristics potentially relevant to empowerment including

academic rank, years of experience, disciplinary affiliation, employment status (full-time versus part-time), and institutional type. Such comparability checking constitutes standard practice in invariance research, as pre-existing group differences on relevant covariates may confound interpretation of measurement non-invariance [25].

Third, the sampling frame—state universities and colleges—excludes private institutions, which may operate under different governance structures and resource constraints that influence faculty empowerment differently. Fourth, no information was provided regarding response rates overall or by gender, precluding assessment of potential nonresponse bias. If response rates differed systematically between male and female faculty, the participating sample may not represent the target population adequately for either gender group, threatening both internal and external validity of invariance testing conclusions. Nonresponse bias represents a pervasive concern in survey research, as nonrespondents often differ systematically from respondents on key variables of interest[26]. Finally, the temporal scope of data collection during pandemic-related restrictions introduces additional limitations regarding the representativeness and generalizability of findings, as faculty empowerment experiences during crisis periods may differ substantially from typical operational contexts.

Due to pandemic-related restrictions during data collection, all survey administration occurred online through secured digital platforms. While this approach ensured participant safety and data integrity during exceptional circumstances, exclusive reliance on online administration introduced several methodological limitations. First, online survey administration creates potential coverage bias by systematically excluding faculty with limited internet access, inadequate technological infrastructure, or insufficient digital literacy—factors that may correlate with both gender and empowerment levels in resource-constrained Philippine higher education settings. Research on digital divides in developing countries demonstrates that internet access and technological proficiency often vary by gender, socioeconomic status, and geographic location, potentially introducing systematic bias into samples recruited exclusively through online channels[27].

Second, online survey administration can produce different response patterns compared to paper-based or interview-based approaches, particularly for lengthy instruments requiring sustained attention. Research indicates that online respondents may exhibit greater satisficing behavior (providing minimally acceptable responses rather than optimal responses), higher rates of item nonresponse, and different response distributions compared to respondents completing identical instruments via alternative administration modes [28]. These mode effects become especially pronounced for lengthy instruments like the 86-item Teacher Empowerment Scale employed in this study, raising questions about measurement equivalence across administration modes that remain unexamined. Third, the lack of interviewer presence in self-administered online surveys eliminates opportunities for clarifying ambiguous items, encouraging complete responses, and maintaining respondent engagement throughout lengthy questionnaires—factors that may affect data quality differentially for male and female respondents if systematic gender differences exist in help-seeking behavior or question interpretation strategies.

5. Data collection

Data collection utilized the 86-item Teacher Empowerment Scale, developed through qualitative exploration, expert validation, and psychometric testing. The scale employs a 5-point Likert response format ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Prior to main data collection, the instrument underwent pretesting with 325 college teachers from state universities and colleges in Region VIII, yielding high internal consistency reliability (Cronbach's alpha = .947). While this coefficient indicates strong inter-item correlation, excessively high alpha values (particularly those exceeding .90) may paradoxically signal

measurement problems rather than psychometric excellence. Specifically, very high alpha values often indicate item redundancy, suggesting that multiple items essentially ask the same question using slightly different wording, thereby adding respondent burden without contributing unique measurement information [29]. Contemporary psychometric research emphasizes that alpha values substantially exceeding .90 frequently reflect unnecessarily long scales that could be shortened without sacrificing reliability or construct coverage, particularly when coefficient omega or other model-based reliability estimates suggest that alpha overestimates true score reliability due to violations of tau-equivalence assumptions[30].

Furthermore, Cronbach's alpha represents only a lower-bound estimate of reliability under restrictive assumptions of essential tau-equivalence and uncorrelated measurement errors—assumptions that are rarely verified in practice and frequently violated in applied research [31]. The coefficient actually estimates the proportion of variance in observed scores attributable to systematic variance (including both true score variance and any systematic method variance), rather than true score variance exclusively. Consequently, elevated alpha values may reflect systematic response biases such as acquiescence bias, halo effects, or common method variance rather than genuine reliability. The study would have benefited from reporting complementary reliability estimates such as omega total (ωt) or omega hierarchical (ωh) that relax tau-equivalence assumptions and provide more accurate reliability estimation under realistic measurement conditions[32]. Additionally, the 86-item length raises serious concerns about respondent burden and potential fatigue effects that may compromise response quality, particularly given that the instrument was administered entirely online without interviewer support to maintain engagement.

The pretesting phase helped ensure items were contextually appropriate and meaningful for the target population before full-scale administration. However, several data collection limitations merit careful consideration. First, the 86-item instrument represents substantial respondent burden that may have induced fatigue effects, potentially compromising response quality especially for items appearing later in the survey. Survey methodology research consistently demonstrates that respondent fatigue increases with questionnaire length, manifesting as decreased response variability, increased item nonresponse, greater satisficing behavior (selecting responses with minimal cognitive effort), and diminished attention to item content [33]. These fatigue effects become particularly salient when lengthy instruments are administered online without interviewer presence to maintain engagement and motivation. Critically, if male and female respondents exhibited different patterns of survey fatigue or adopted different satisficing strategies when cognitively fatigued, measurement non-invariance could reflect differential fatigue effects rather than genuine gender differences in how empowerment items function psychometrically.

Second, the three-month data collection period, while accommodating faculty schedules and ensuring adequate institutional representation, introduced temporal variation that could confound results if institutional events, policy changes, or external circumstances occurring during this window affected male and female faculty differentially. For instance, if budget announcements, leadership transitions, promotion decisions, or workload changes occurred during data collection and impacted male and female faculty differently, observed differences in item responses might reflect these contextual factors rather than stable gender-based measurement properties. Third, the study does not report whether data collection occurred during academic term or between semesters, a potentially important consideration as faculty stress levels, workload pressures, and empowerment perceptions may vary systematically across the academic calendar.

Survey distribution occurred through official institutional channels following formal permissions from university administrators. Participants received detailed information about the study's purpose and were assured of confidentiality and anonymity. Digital consent forms were embedded within the survey platform,

requiring acknowledgment before accessing questionnaire items—a procedure consistent with ethical standards for online [29]. To maximize response rates and minimize missing data, automated reminders were sent to non-respondents, and partial responses were preserved to allow completion across multiple sessions. While these procedures align with contemporary best practices for online survey administration, the study did not report response rates overall or disaggregated by gender, precluding assessment of potential nonresponse bias that represents a pervasive threat to validity in survey research.

Nonresponse bias occurs when individuals who choose not to participate differ systematically from those who do participate on key variables of interest, potentially producing samples unrepresentative of the target population [35]. If response rates differed substantially between male and female faculty, the participating samples for each gender group may not adequately represent their respective populations, threatening both the internal validity of invariance testing and external validity of generalization. For example, if female faculty responded at higher rates than male faculty (or vice versa), and if response propensity correlated with empowerment levels, observed gender differences in item functioning could reflect selection artifacts rather than genuine measurement properties. Contemporary survey methodology emphasizes that response rate information constitutes essential evidence for evaluating potential bias and should be routinely reported to enable readers to judge the credibility of study conclusions[34].

Additionally, the study does not describe procedures for handling missing data, despite the fact that even with automated reminders and the ability to complete surveys across multiple sessions, item-level nonresponse invariably occurs in lengthy questionnaires. The treatment of missing data can substantially impact parameter estimates, standard errors, and model fit statistics in both Rasch analysis and confirmatory factor analysis frameworks[36]. Modern missing data theory distinguishes between data missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), with different analytical approaches appropriate for different missing data mechanisms. If missing data patterns differed between male and female respondents—for instance, if certain empowerment items produced higher nonresponse rates for one gender—and if these patterns were not appropriately addressed through missing data techniques such as full information maximum likelihood or multiple imputation, parameter estimates and invariance tests could be biased. The study's failure to report missing data rates, patterns, or handling procedures represents a notable methodological limitation that hinders evaluation of potential bias in study findings.

Finally, the exclusive reliance on self-report measures introduces potential common method variance that may inflate relationships among study variables and contribute to elevated internal consistency estimates. While self-report remains appropriate and often necessary for measuring psychological constructs like empowerment, triangulation with alternative data sources such as supervisor ratings, behavioral indicators, or archival records of faculty participation in governance activities could strengthen construct validity and reduce common method bias. The potential for social desirability bias also warrants consideration, particularly if male and female faculty differ in their tendencies to present themselves favorably when reporting empowerment-related attitudes and behaviors.

6. Data analysis

The inclusion of the Rasch Model addresses well-documented psychometric limitations of Classical Test Theory, which include sample-dependent and item-dependent calibrations, problematic assumptions about equal intervals for ordinal Likert-scale data, and the assumption that chosen response categories function appropriately for all respondents without empirical verification[37]. The Rasch model, a foundational approach within item response theory, aims to describe the probabilistic relationship between person ability levels and item difficulty parameters, offering several advantages for psychometric analysis including

sample-independent parameter estimation, interval-level measurement properties, and explicit mechanisms for evaluating whether data meet measurement model requirements[38]. The Rasch analysis comprised multiple diagnostic procedures: item polarity assessment, item fit evaluation, item characteristic curve examination, differential item functioning detection, response category diagnostics, and person-item map construction. Item polarity was evaluated using point-measure correlation coefficients (PTMEA CORR), with acceptable values ranging from 0.30 to 0.80, indicating that items measure a unidimensional construct consistently. For item fit assessment, mean-square (MNSQ) values between 0.50 and 2.0 were considered acceptable, with values below 0.50 suggesting item overfit or redundancy and values exceeding 2.0 indicating substantial misfit that degrades measurement quality[15]. Item characteristic curves graphically depict the relationship between respondent ability and response probability, with item difficulty reflected in horizontal positioning along the ability continuum[39]. However, fit statistics exhibit sensitivity to sample size, with large samples potentially flagging trivial misfit as statistically significant despite negligible practical impact on measurement.

Differential item functioning (DIF) analysis evaluated whether items exhibit bias by comparing response patterns between male and female respondents matched on overall ability levels. The fundamental assumption tested was that response probability depends solely on empowerment level, not on gender after controlling for ability [42]. DIF detection employed three complementary statistics: the Mantel-Haenszel chi-square statistic, Standardized Liu-Agresti Cumulative Common Log-Odds Ratio (LOR Z), and Liu-Agresti Cumulative Common Log-Odds Ratio (L-A LOR). Items with Mantel statistics exceeding 3.84 ($p \leq .05$) were flagged as potentially exhibiting DIF, LOR Z values outside ±1.96 provided additional DIF evidence, and the L-A LOR classified DIF magnitude with values below 0.53 indicating negligible DIF (Class A), values between 0.53 and 0.74 representing moderate DIF (Class B), and values exceeding 0.74 signifying substantial DIF (Class C) warranting item exclusion. Critical limitations characterize this DIF detection approach. Statistical DIF detection does not automatically imply problematic bias requiring item elimination, as items may function differently across groups for substantive reasons related to genuine construct differences rather than measurement artifacts[43]. The study employs multiple DIF detection methods but does not articulate how discrepancies among these methods would be resolved, as different DIF detection methods can yield conflicting conclusions about which items exhibit DIF and the magnitude of such DIF[40]. The DIF analysis framework assumes that the matching variable (total test score) accurately represents the construct of interest; if the total score exhibits substantial measurement error, the DIF analysis may fail to detect genuine item bias or may falsely flag unbiased items[41].

For each subscale, response category statistics were examined by aggregating items into dimensional groupings. Optimal category functioning requires monotonic increases in average measures across the 5-point scale from category 1 (not empowered) to category 5 (empowered) without threshold disordering—a condition where intermediate categories become less probable than adjacent categories at any ability level, indicating respondents cannot reliably discriminate between adjacent response options[39]. Threshold disordering suggests that the chosen number or labeling of response categories exceeds respondents' discrimination capacity, indicating that category reduction or relabeling may improve measurement precision. Person-item maps provided visual representations of respondent ability distributions relative to item difficulty distributions, with ideal targeting occurring when item difficulties span the full range of respondent abilities [44]. Substantial mismatches indicate targeting problems that reduce measurement efficiency, leaving portions of the ability continuum poorly measured.

Testing measurement invariance across genders began with constructing a baseline structural model representing teacher empowerment dimensions. Multi-group confirmatory factor analysis (MG-CFA)

examined invariance through a hierarchical sequence: configural invariance (establishing equivalent factor structure), metric invariance (constraining factor loadings equal), scalar invariance (constraining item intercepts equal), and strict invariance (constraining residual variances equal)[45]. Invariance evaluation relied on hierarchical comparisons of model-data fit indices and chi-square difference tests between successive models. Following contemporary guidelines, changes in comparative fit index (ΔCFI) exceeding .010, changes in root mean square error of approximation (ΔRMSEA) exceeding .015, or statistically significant chi-square differences ($p < .05$) indicated invariance violations [46]. When invariance held, measurement equivalence across gender groups was supported, permitting valid between-group comparisons. In cases of partial invariance violations, sources and implications of non-invariance were to be explored through examination of parameter differences between gender groups, with model comparisons using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to identify optimal constrained models[47].

Several analytical limitations warrant acknowledgment. The hierarchical invariance testing approach assumes correct specification of the baseline configural model; if this initial model exhibits misspecification, subsequent invariance tests may yield misleading conclusions [48]. The study's reliance on chi-square difference testing for invariance evaluation remains highly sensitive to sample size, with large samples potentially flagging trivial non-invariance as statistically significant despite negligible practical impact [50]. The combination of Rasch analysis and multi-group confirmatory factor analysis represents somewhat redundant approaches to invariance testing, raising questions about how conflicting results between approaches would be reconciled. The analysis focused exclusively on gender as a grouping variable, precluding examination of potential invariance violations across other demographic characteristics such as academic rank or discipline that may interact with gender in complex ways. The study does not address statistical power for detecting non-invariance, despite the fact that invariance tests exhibit varying power depending on sample size and effect size[49]. Finally, the study does not discuss how identified cases of non-invariance would be addressed—Whether through partial invariance modeling, item elimination, or substantive interpretation—leaving unclear how measurement non-invariance would impact final study conclusions.

7. Ethical considerations

This research adhered to rigorous ethical standards throughout its implementation. Prior to data collection, the study protocol received approval from the university's Institutional Research Ethics Committee. This approval ensured that the research methodology, data collection procedures, and data management plans met established ethical guidelines for research involving human participants. Informed consent was obtained from all participants before they completed the survey. The consent form detailed the study's purpose, procedures, potential risks and benefits, confidentiality measures, and the voluntary nature of participation. Participants were informed of their right to withdraw from the study at any point without penalty or consequence to their professional standing.

To protect participant confidentiality, all personally identifiable information was removed during data processing. Data were stored on password-protected servers with encrypted access limited to the research team. The reporting of results maintained anonymity by presenting only aggregated findings without identifying specific institutions or individuals. Special consideration was given to the power dynamics inherent in educational institutions. The research team ensured that institutional administrators had no access to individual responses, protecting participants from potential professional repercussions. Additionally, the online administration method allowed participants to complete the survey in private settings, minimizing

potential influence from colleagues or supervisors. The study design also considered the principle of beneficence by ensuring that findings would contribute meaningful insights to the improvement of teacher empowerment in higher education, potentially benefiting participants and the broader educational community through enhanced understanding of gender-related factors in professional empowerment.

8. Results

The findings of this study must be interpreted within the broader context of contemporary research on measurement invariance and gender equity in educational assessment. The overall good model fit observed across all three factors (with mean MNSQ values approximating 1.00 and item reliabilities ranging from 0.90 to 0.93) suggests that the Teacher Empowerment Scale demonstrates robust psychometric properties consistent with recent validation studies employing Rasch methodology in educational contexts [38]. These results align with emerging evidence that well-constructed empowerment scales can maintain measurement stability across demographic groups when items are carefully developed and empirically validated [19]. The minimal differential item functioning detected across factors—With only 5 items in Factor 1, 4 items in Factor 2, and no statistically significant DIF in Factor 3—Indicates that the scale largely achieves measurement equivalence across gender groups, a critical prerequisite for making valid comparisons between male and female faculty members' empowerment experiences [15]. This pattern of results suggests that the construct of teacher empowerment, as operationalized in this instrument, transcends gender-specific interpretations for the majority of items, supporting the theoretical proposition that empowerment represents a universal psychological construct applicable across diverse populations.

However, the presence of significant DIF in specific items warrants careful consideration of how gender may influence interpretation of particular empowerment-related experiences rather than the underlying construct itself. Contemporary DIF research emphasizes that statistically significant differential functioning does not automatically indicate problematic bias; rather, it may reflect genuine differences in how male and female educators experience specific organizational practices or institutional structures[51]. For instance, items exhibiting DIF in Factors 1 and 2 may capture aspects of continuous professional development and teaching autonomy that intersect with gendered institutional norms, disciplinary cultures, or career trajectory patterns documented in recent higher education literature[17]. The absence of significant DIF in Factor 3 (work climate and conditions) proves particularly noteworthy, as it suggests that perceptions of organizational support, resource availability, and workplace unity operate similarly for male and female faculty—a finding that contrasts with some earlier research reporting gender disparities in workplace climate perceptions but aligns with recent evidence of converging workplace experiences in contemporary higher education settings[18]. These nuanced patterns of invariance and non-invariance underscore the importance of examining measurement equivalence at the item level rather than assuming uniform functioning across entire instruments, as recommended by current best practices in cross-group validation research[21].

The study had 968 teachers from State Universities and Colleges and utilized online data collection since it was the start of the COVID-19 Pandemic and health restrictions were intense at that time. The participants who were part of the baseline test, the in-depth interview, pilot testing, and the actual survey in other phases were excluded in the actual administration of the instrument. After the conduct of the administration of the scale, checking for item homogeneity was done. In this study, Rasch analysis was used in verifying that items reflect homogeneity in terms of gender, a trade-off between respondent's perceived factors affecting teacher empowerment. For factor 1 (fostering continuous development), a total of 51 items were checked for fist statistics. The overall fit analysis is presented in **Table 1**.

Table 1. Overall Rasch fit statistics and reliability coefficient for factor 1.

	Min	Max	Mean	SD
Infit MNSQ	0.87	1.31	1.00	0.09
Outfit MNSQ	0.85	1.22	0.99	0.09
Item Reliability			0.90	

The table presents the overall Rasch fit statistics and reliability coefficient of a set of items. The mean values of the Infit MNSQ and Outfit MNSQ are both close to 1.00, which indicates a good fit of the items to the Rasch model. The standard deviations of both Infit MNSQ and Outfit MNSQ are also relatively small, which further indicates that the fit is consistent across items[52]. The minimum and maximum values of the Infit MNSQ and Outfit MNSQ indicate that there are some items with slightly poorer fit than others, but overall, the fit is good. The item reliability coefficient is 0.90, which is considered to be high. This indicates that the set of items is reliable in measuring the construct of interest[53]. Overall, the table suggests that the set of items has a good fit to the Rasch model and is reliable in measuring the construct of interest.

According to Goretzko et al.[52], an item should have infit and outfit mean squares of 1.0 to have a perfect fit, or between 0.5-1.5 to be productive for measurement. All items have infit and outfit mean squares inside the productive for measurement range. Further analysis highlights the differential item functioning (DIF) analysis for factor 1 according to gender. The analysis measures whether there are any differences in how males and females respond to the test items in Factor 1, which could indicate bias or unfairness in the test. Each row corresponds to a different test item, and the columns show various measures related to DIF. Overall, it appears that several items show significant DIF for gender, including item number 2, 10, 77, 79, and 86. The magnitude and direction of the DIF varies across items, with some items being associated more strongly with females and others with males.

For factor 2 (teaching ownership and freedom), a total of 32 items were checked for fist statistics. The overall fit analysis is presented in **Table 2** while the item measures. The Infit MNSQ ranges from 0.60 to 1.18, with a mean of 1.00 and standard deviation of 0.10. The Outfit MNSQ ranges from 0.61 to 1.20, with a mean of .96 and standard deviation of 0.11. These values suggest that overall, the 32 items included in Factor 2 fit the Rasch model reasonably well. The mean values of the Infit MNSQ and Outfit MNSQ are both close to 1.00, which indicates a good fit of the items to the Rasch model[54]. The standard deviations of both Infit MNSQ and Outfit MNSQ are also relatively small, which further indicates that the fit is consistent across items [52]. The minimum and maximum values of the Infit MNSQ and Outfit MNSQ indicate that there are some items with slightly poorer fit than others, but overall, the fit is good. The item reliability coefficient is 0.90, which is considered to be high. This suggests that the items are measuring a common underlying construct [53]. Overall, based on the Rasch analysis results provided, Factor 2 appears to be a reliable and valid measure of the construct being assessed by the 32 items included in this factor.

Table 2. Overall Rasch fit statistics and reliability coefficient for factor 2.

	Min	Max	Mean	SD
Infit MNSQ	0.60	1.18	1.00	0.10
Outfit MNSQ	0.61	1.20	0.96	0.11
Item Reliability			0.90	

According to Goretzko et al.[52], an item should have infit and outfit mean squares of 1.0 to have a perfect fit, or between 0.5-1.5 to be productive for measurement. All items have infit and outfit mean squares inside the productive for measurement range. Further analysis highlights the differential item functioning

(DIF) analysis for factor 2 according to gender. The analysis measures whether there are any differences in how males and females respond to the test items in Factor 2, which could indicate bias or unfairness in the test.

DIF occurs when the probability of responding to an item correctly differs between groups, in this case, males and females, even if they have the same underlying ability or construct being measured. Overall, it appears that several items show significant DIF for gender, including item number 62, 64, 71, and 72. The magnitude and direction of the DIF varies across items, with some items being associated more strongly with females and others with males. Another potential explanation for DIF is that the wording or content of the item is interpreted differently by males and females, leading to different probabilities of responding correctly.

For factor 3 (work climate and conditions), a total of 3 items were checked for fist statistics. The overall fit analysis is presented in **Table 3** while the item measures and item fit statistics is shown in **Table 4**.

Table 3. Overall Rasch fit statistics and reliability coefficient for factor 3.

	Min	**Max**	**Mean**	**SD**
Infit MNSQ	0.94	0.99	0.97	0.02
Outfit MNSQ	0.75	0.81	0.79	0.03
Item Reliability		0.93		

The Infit MNSQ ranges from 0.94 to 0.99, with a mean of 0.97 and standard deviation of 0.02. The Outfit MNSQ ranges from 0.75 to 0.81, with a mean of 0.79 and standard deviation of 0.03. These values suggest that overall, the 3 items included in Factor 3 fit the Rasch model reasonably well. The mean values of the Infit MNSQ and Outfit MNSQ are both close to 1.00, which indicates a good fit of the items to the Rasch model [54]. The standard deviations of both Infit MNSQ and Outfit MNSQ are also relatively small, which further indicates that the fit is consistent across items. The minimum and maximum values of the Infit MNSQ and Outfit MNSQ indicate that there are some items with slightly poorer fit than others, but overall, the fit is good. The item reliability coefficient is 0.93, which is considered to be high. This suggests that the items are measuring a common underlying construct[53].

Overall, based on the Rasch analysis results provided, Factor 3 appears to be a reliable and valid measure of the construct being assessed by the 3 items included in this factor. According to Goretzko et al.[52], an item should have infit and outfit mean squares of 1.0 to have a perfect fit, or between 0.5-1.5 to be productive for measurement. As seen on Table 27, all items have infit and outfit mean squares inside the productive for measurement range.

Table 4. Item measures and item fit statistics for factor 3.

Items	**Item Measure**	**Standard Error**	**Infit MNSQ**	**Outfit MNSQ**
66. No one supports my decision related to school obligation.	-0.13	0.08	0.98	0.75
67. The school has limited teaching resources.	0.40	0.07	0.99	0.81
68. There is no unity at work.	-0.27	0.08	0.94	0.80

Table 5 highlights the differential item functioning (DIF) analysis for factor 3 according to gender. The analysis measures whether there are any differences in how males and females respond to the test items in Factor 3, which could indicate bias or unfairness in the test. This table shows the results of a DIF analysis for Factor 3 by gender. Moreover, it appears that item number 66 has very little DIF, while item number 67 has moderate DIF in favor of females, and item number 68 has moderate DIF in favor of males. Overall, the

results suggest that while there is some DIF for Factor 3 by Gender for these three items, the differences are generally not statistically significant since all of the p-values were greater than the conventional threshold set at 0.05.

Table 5. Differential item functioning for factor 3 by gender.

Items	Female DIF Measure	Male DIF Measure	DIF Contrast	Joint SE	Welch T-value	p-value
66. No one supports my decision related to school obligation.	-0.07	-0.17	0.10	0.15	0.65	0.51
67. The school has limited teaching resources.	0.50	0.33	0.17	0.15	1.15	0.25
68. There is no unity at work.	-0.43	-0.15	-0.28	-0.15	1.81	0.07

This study provides extensive evidence for the measurement invariance and psychometric quality of the Teacher Empowerment Scale across gender groups in higher education settings. The Rasch analysis results demonstrate that the scale functions effectively regardless of gender, while also identifying specific areas where gender-based considerations may be relevant. The minimal differential item functioning across factors suggests that the instrument provides fair assessment of teacher empowerment constructs for both male and female educators, supporting its use in diverse educational contexts.

The findings align with previous research by Giguère et al. (2022) regarding gender differences in teacher empowerment, while demonstrating that these differences do not significantly impact the scale's measurement properties. Similarly, the absence of significant DIF in the work climate and conditions factor supports observations by Doganaksoy et al.[55] that organizational climate factors may be experienced more uniformly across demographic groups. The strong item reliability coefficients across all factors (ranging from 0.90 to 0.93) indicate that the scale provides consistent measurement regardless of gender, meeting the standards recommended by Tesio et al.[53] for high-quality psychometric instruments.

The study contributes significantly to addressing the gap identified by Gomes et al.[8] and Pan et al.[14] regarding the need for gender considerations in scale development and validation. By establishing the gender neutrality of a comprehensive teacher empowerment measurement tool while acknowledging specific areas where gender-specific patterns emerge, this research advances the development of equitable assessment instruments for higher education settings. Future research should build upon these findings by exploring additional demographic variables influencing measurement invariance, conducting longitudinal studies of gender-based patterns, and extending validation across diverse cultural contexts.

9. Conclusion and recommendation

This study provides comprehensive evidence for the measurement invariance and psychometric quality of the Teacher Empowerment Scale across gender groups in higher education settings. Through rigorous Rasch analysis of 86 items across three factors, we found consistently good model fit and high reliability coefficients (ranging from 0.90 to 0.93). The analysis revealed minimal differential item functioning across gender groups, with only a few items showing significant gender-based variations in factors 1 and 2, and none in factor 3. These findings demonstrate that the scale functions effectively regardless of gender, while also identifying specific areas where gender-based considerations may be relevant.

The key takeaway is that the Teacher Empowerment Scale provides fair and accurate assessment of teacher empowerment constructs for both male and female educators in higher education. The strong item reliability coefficients and good fit statistics across all factors confirm that the instrument meets the standards

recommended by Tesio et al.[53] and Goretzko et al.[52] for high-quality psychometric instruments. Additionally, the findings align with previous research by Giguère et al. (2022) and Doganaksoy et al.[55] regarding gender differences in teacher empowerment, while demonstrating that these differences do not significantly impact the scale's overall measurement properties.

Recommendations for researchers and educational practitioners include:

1. Implement the Teacher Empowerment Scale with confidence across gender groups in higher education settings, while remaining attentive to the specific items identified with differential functioning.
2. Consider refinement of the few items showing significant DIF to enhance gender neutrality in future versions of the scale.
3. Extend validation research to examine measurement invariance across additional demographic variables such as age, years of experience, and educational attainment.
4. Conduct longitudinal studies to examine the stability of the scale's psychometric properties and gender invariance over time.
5. Utilize the three-factor structure (fostering continuous improvement, teaching ownership and freedom, and work climate and conditions) as a framework for designing targeted teacher empowerment interventions in higher education.
6. Develop comparative studies examining empowerment patterns across different types of higher education institutions using this validated instrument.
7. Explore the relationship between teacher empowerment scores and educational outcomes to further establish the practical utility of the scale.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Celik, O. T., Sari, T., & Karagozoglu, A. A. (2024). A Systematic Literature review of research on teacher Empowerment. Urban Education, 60(10), 2728–2763. https://doi.org/10.1177/00420859241301073
2. Arta, N. G. Y. (2024). Asesmen dalam Pendidikan: Konsep, Pendekatan, Prinsip, Jenis, dan Fungsi. Jurnal Pendidikan Bahasa Dan Budaya, 3(3), 170–190. https://doi.org/10.55606/jpbb.v3i3.3925
3. Akpan, W. M., & Ayinmoro, A. D. (2024). Age Difference between Spouses and Women Empowerment in Nigeria. International Journal of Research and Innovation in Social Science, VIII(II), 996–1011. https://doi.org/10.47772/ijriss.2024.802070
4. Berhanu, K. Z. (2023). Development and validation of teachers' psychological empowerment scale in Ethiopian context. Management in Education. https://doi.org/10.1177/08920206231215264
5. Kabat, M. (2024). Empowerment jako usprawnienie aktywności nauczyciela. Kultura I Edukacja, 2024(2 (144)), 131–150. https://doi.org/10.15804/kie.2024.02.07
6. Ahmadi, R., & Arief, N. F. (2022). Teacher empowerment to improve the quality of education and school progress. EDUTEC Journal of Education and Technology, 6(2), 431–439. https://doi.org/10.29062/edu.v6i2.498
7. Golle, J., Schils, T., Borghans, L., & Rose, N. (2022). Who is considered gifted from a teacher's perspective? A representative Large-Scale study. Gifted Child Quarterly, 67(1), 64–79. https://doi.org/10.1177/00169862221104026
8. Gomes, C. M. A., Farias, H. B., & Jelihovschi, E. G. (2024). Invariance across sex, school, and educational level to Learning Approaches Scale (EABAP). Psico-USF, 29. https://doi.org/10.1590/1413-827120242901e262990
9. Webber, C., & Nickel, J. (2022). A vibrant and empowering context for teacher leaders. International Journal for Leadership in Learning, 22(1), 1–27. https://doi.org/10.29173/ijll2

10. Short, P. M., & Rinehart, J. S. (1992). School Participant Empowerment Scale: Assessment of Level of Empowerment within the School Environment. Educational and Psychological Measurement, 52(4), 951–960. https://doi.org/10.1177/0013164492052004018
11. Van Woerden, R., Van Goch, M. M., Schruijer, S. G. L., & Van Der Tuin, I. (2025). Students' teamwork behaviour in multidisciplinary student teams: an ethnographic case study. Higher Education Research & Development, 44(5), 1259–1274. https://doi.org/10.1080/07294360.2025.2468394
12. Hibdon, J., Schafer, J., & Kyle, M. (2023). Introduction to the special issue on measurement and methodology: addressing challenges and exploring opportunities. Journal of Crime and Justice, 47(1), 1–3. https://doi.org/10.1080/0735648x.2023.2211561
13. Katsikeas, C. S., Madan, S., Brendl, C. M., Calder, B. J., Lehmann, D. R., Baumgartner, H., Weijters, B., Wang, M., Huang, C., & Huber, J. (2022). Commentaries on "Scale use and abuse: Toward best practices in the deployment of scales." Journal of Consumer Psychology, 33(1), 244–258. https://doi.org/10.1002/jcpy.1319
14. Pan, L., Lu, L., & Zhang, T. (2020). Destination gender: Scale development and cross-cultural validation. Tourism Management, 83, 104225. https://doi.org/10.1016/j.tourman.2020.104225
15. Svetina, D., Rutkowski, L., & Rutkowski, D. (2019). Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/semTools Packages. Structural Equation Modeling a Multidisciplinary Journal, 27(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776
16. Panela, T. L. V. (2023). Development and validation of teacher empowerment scale: Examining factor structure and Rasch model fit in higher education [PhD Dissertation]. Leyte Normal University.
17. Piltz, L. M., Carpendale, E. J., & Laurens, K. R. (2023). Measurement invariance across age, gender, ethnicity, and psychopathology of the Psychotic-Like Experiences Questionnaire for Children in a community sample. International Journal of Methods in Psychiatric Research, 32(4), e1962. https://doi.org/10.1002/mpr.1962
18. Barbosa-Leiker, C., Burduli, E., Arias-Losado, R., Muller, C., Noonan, C., Suchy-Dicey, A., Nelson, L., Verney, S. P., Montine, T. J., & Buchwald, D. (2022). Testing gender and longitudinal measurement invariance of the SF-36 in American Indian older adults: The strong heart study. Psychological Assessment, 34(9), 870–879. https://doi.org/10.1037/pas0001153
19. Schlechter, P., & Neufeld, S. a. S. (2024). Longitudinal and gender measurement invariance of the General Health Questionnaire-12 (GHQ-12) from adolescence to emerging adulthood. Assessment, 31(8), 1687–1701. https://doi.org/10.1177/10731911241229573
20. Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013a). Sample size requirements for structural equation models. Educational and Psychological Measurement, 73(6), 913–934. https://doi.org/10.1177/0013164413495237
21. Liu, X. (2023). Detecting differential item functioning with multiple causes: A comparison of three methods. International Journal of Testing, 24(1), 53–79. https://doi.org/10.1080/15305058.2023.2286381
22. Kyriazos, T. A. (2018). Applied Psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. Psychology, 09(08), 2207–2230. https://doi.org/10.4236/psych.2018.98126
23. Reddy, K. G., & Khan, M. (2023). Constructing efficient strata boundaries in stratified sampling using survey cost. Heliyon, 9(11), e21407. https://doi.org/10.1016/j.heliyon.2023.e21407
24. Lohr, S. L. (2021). Sampling: Design and Analysis. Chapman and Hall/CRC. https://doi.org/10.1201/9780429298899
25. Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. Journal of Psychoeducational Assessment, 29(4), 347–363. https://doi.org/10.1177/0734282911406661
26. Andrade, C. (2020). The limitations of online surveys. Indian Journal of Psychological Medicine, 42(6), 575–576. https://doi.org/10.1177/0253717620957496
27. Daikeler, J., Bošnjak, M., & Manfreda, K. L. (2019). Web versus Other survey Modes: An updated and extended Meta-Analysis comparing response rates. Journal of Survey Statistics and Methodology, 8(3), 513–539. https://doi.org/10.1093/jssam/smz008
28. Ball, H. L. (2019). Conducting online surveys. Journal of Human Lactation, 35(3), 413–417. https://doi.org/10.1177/0890334419848734
29. Zakariya, Y. F. (2022). Cronbach's alpha in mathematics education research: Its appropriateness, overuse, and alternatives in estimating scale reliability. Frontiers in Psychology, 13, 1074430. https://doi.org/10.3389/fpsyg.2022.1074430
30. Taber, K. S. (2017). The use of Cronbach's Alpha when developing and reporting research instruments in science education. Research in Science Education, 48(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2
31. Sijtsma, K. (2008). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. Psychometrika, 74(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0
32. McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412–433. https://doi.org/10.1037/met0000144

33. Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opinion Quarterly, 73(2), 349–360. https://doi.org/10.1093/poq/nfp031
34. Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method, 4th Edition. Wiley. https://eric.ed.gov/?id=ED565653
35. Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. Public Opinion Quarterly, 70(5), 646–675. https://doi.org/10.1093/poq/nfl033
36. Enders, C. K. (2010). Applied missing data analysis. http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=020418619&sequence=000002&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA
37. Bedgood, D. R., Guisard, Y., Howitt, J., Prenzler, P., Barril, C., & Ryan, D. (2016). Rasch analysis of exams: A way to document graduate outcomes to employers? Proceedings of the Australian Conference on Science and Mathematics Education (Formerly UniServe Science Conference), 24. https://openjournals.library.sydney.edu.au/index.php/IISME/article/download/10800/11342
38. Hagquist, C., Bruce, M., & Gustavsson, J. P. (2008). Using the Rasch model in nursing research: An introduction and illustrative example. International Journal of Nursing Studies, 46(3), 380–393. https://doi.org/10.1016/j.ijnurstu.2008.10.007
39. Andrich, D. (2012). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "Threshold disorder controversy." Educational and Psychological Measurement, 73(1), 78–124. https://doi.org/10.1177/0013164412450877
40. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: AnRPackage for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. Journal of Statistical Software, 39(8), 1–30. https://doi.org/10.18637/jss.v039.i08
41. Chalmers, R. P., Counsell, A., & Flora, D. B. (2015). It might not make a big DIF. Educational and Psychological Measurement, 76(1), 114–140. https://doi.org/10.1177/0013164415584576
42. Penfield, R. D., & Camilli, G. (2006). 5 Differential item functioning and item bias. In Handbook of statistics (pp. 125–167). https://doi.org/10.1016/s0169-7161(06)26005-x
43. Salfran, D., & Spiess, M. (2018). Generalized Additive Model multiple imputation by chained equations with package ImputeRobust. The R Journal, 10(1), 61. https://doi.org/10.32614/rj-2018-014
44. Wright, B. D., & Stone, M. H. (1979). Best Test Design. Rasch Measurement. https://eric.ed.gov/?id=ED436552
45. 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
46. 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
47. Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference. Sociological Methods & Research, 33(2), 261–304. https://doi.org/10.1177/0049124104268644
48. Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2019). Confirmatory Factor Analysis (CFA), Exploratory Structural Equation Modeling (ESEM), and SET-ESEM: optimal balance between goodness of fit and parsimony. Multivariate Behavioral Research, 55(1), 102–119. https://doi.org/10.1080/00273171.2019.1602503
49. Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory Factor Analysis: Some consequences for model fit and standardized parameter estimates. Multivariate Behavioral Research, 49(6), 518–543. https://doi.org/10.1080/00273171.2014.933762
50. 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
51. Li, Z., Shin, J., Kuang, H., & Huggins-Manley, A. C. (2024). Exploring the evidence to interpret differential item functioning via response process data. Educational and Psychological Measurement, 85(4), 783–813. https://doi.org/10.1177/00131644241298975
52. Goretzko, D., Siemund, K., & Sterner, P. (2023). Evaluating model fit of measurement models in confirmatory factor analysis. Educational and Psychological Measurement, 84(1), 123–144. https://doi.org/10.1177/00131644231163813
53. Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2023). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. Disability and Rehabilitation, 46(3), 604–617. https://doi.org/10.1080/09638288.2023.2169772
54. Lane, K. L., Oakes, W. P., Buckman, M. M., Lane, N. A., Lane, K. S., Fleming, K., Romine, R. E. S., Sherod, R. L., Chang, C., Jones, J., Cantwell, E. D., & Crittenden, M. (2023). Examination of the factor structure and measurement invariance of the SRSS-IE. Remedial and Special Education, 45(3), 152–172. https://doi.org/10.1177/07419325231193147

55. Doganaksoy, N., Meeker, W. Q., & Hahn, G. J. (2023). Product reliability: How statistics fits in. Significance, 20(2), 28–32. https://doi.org/10.1093/jrssig/qmad029
56. Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013b). Sample size requirements for structural equation models. Educational and Psychological Measurement, 73(6), 913–934. https://doi.org/10.1177/0013164413495237
57. Wright, B. D., & Stone, M. H. (1979). Best Test Design. Rasch Measurement. https://eric.ed.gov/?id=ED436552
58. Zakariya, Y. F. (2022). Cronbach's alpha in mathematics education research: Its appropriateness, overuse, and alternatives in estimating scale reliability. Frontiers in Psychology, 13, 1074430. https://doi.org/10.3389/fpsyg.2022.1074430