Robust and Intelligent Approaches to Ordinal Factor Analysis: An Empirical Comparison of Robust, Machine Learning, and Deep Learning Methods

Abuelgasim Ahmed
School of Quantitative Sciences, Universiti Utara Malaysia (UUM)

Zahayu Binti Md Yusof, PhD
School of Quantitative Sciences, Universiti Utara Malaysia (UUM)

ABSTRACT
Ordinal Likert-type indicators are ubiquitous in behavioral and social science measurement, yet applying estimators designed for continuous normal variables can bias parameters and inflate misfit under skewed category use and floor/ceiling effects. This study benchmarks traditional and robust ordinal CFA estimators (WLS, WLSMV, DWLS) using simulated datasets (n = 250, 500, 1000) and an empirical Malaysian Green Consumption dataset (N = 375). All CFA models were estimated on polychoric correlation matrices and evaluated using CFI, TLI, RMSEA, and SRMR. Estimator stability was assessed via nonparametric bootstrapping on the real dataset (B = 500), summarizing convergence rates and average 95% confidence-interval widths for standardized loadings. In addition, machine learning (RF, GBM, SVM) and deep learning (DNN, CNN, RNN) models were evaluated for outcome prediction using 5-fold cross-validation (R^2, RMSE, MAE). Results show that robust estimators consistently improve fit and stability relative to WLS in small samples (e.g., at n = 250, WLSMV achieved CFI ≈ 0.97 and RMSEA ≈ 0.05 versus WLS CFI ≈ 0.94 and RMSEA ≈ 0.08), and reduce uncertainty in loadings (mean CI width ≈ 0.14–0.15 versus 0.20 for WLS) with near-perfect bootstrap convergence. For prediction, nonlinear learners perform best, with DNN (R^2 ≈ 0.35) and GBM (R^2 ≈ 0.33) outperforming other baselines. Overall, the findings provide practical guidance for estimator choice, stability reporting, and predictive validation when analyzing ordinal data.

Keywords
Ordinal data; confirmatory factor analysis; polychoric correlation; WLSMV; DWLS; bootstrap; machine learning; deep learning; cross-validation.

1. INTRODUCTION
Ordinal Likert-type items are widely used in behavioral and social science measurement, but they encode order without equal spacing, making continuous-data assumptions problematic. When ordinal indicators are treated as continuous, Pearson correlations may attenuate associations under skewness and floor/ceiling effects, leading to biased factor loadings and distorted fit statistics in CFA [1]. A standard remedy is to treat Likert items as categorical and adopt threshold models in which each observed response arises from discretizing an underlying continuous latent response. Under this framework, polychoric correlations estimate associations between latent response propensities and typically provide more appropriate inputs for ordinal CFA than Pearson correlations [2], [3]. Empirical and simulation studies show that ignoring ordinality can distort measurement conclusions, particularly in small samples and under severe threshold imbalance [4], [5].

Robust ordinal CFA commonly relies on WLS-family estimators. Full WLS is asymptotically efficient but requires estimating a high-dimensional weight matrix that is unstable in typical sample sizes. Consequently, practical implementations often use DWLS and WLSMV, which reduce computational burden and improve convergence and recovery in ordinal settings [6], [7]. Nevertheless, these estimators can still exhibit instability and inflated interfactor relations under sparse categories or strong skewness, motivating explicit stability assessment and careful reporting [8]. Model adequacy is typically evaluated using multiple fit indices (e.g., CFI, TLI, RMSEA, SRMR) rather than a single statistic [9]. To quantify estimator robustness, nonparametric bootstrap resampling is recommended for assessing convergence and confidence-interval width of standardized loadings [10], [11].

In parallel, machine learning and deep learning methods can capture nonlinear patterns and interactions not represented by linear CFA. Tree ensembles and boosting are strong baselines for tabular prediction [12]–[14], and variational autoencoders provide flexible latent embeddings for high-dimensional data [15]. This paper provides a unified comparison of robust CFA estimators and predictive ML/DL models for ordinal data under matched validation protocols, offering practical guidance for estimator selection, stability reporting, and predictive evaluation.

2. RELATED WORK
2.1 Ordinal Measurement and Polychoric-Based CFA
Ordinal Likert-type variables encode rank order but not equal distances between categories. Treating ordinal responses as continuous can attenuate associations under skewness or floor/ceiling effects, leading to biased loadings and misleading fit behavior in CFA [4], [5]. A common solution is the threshold (latent response) view of ordinal measurement, where observed categories are discretized realizations of an underlying continuous response. Under this framework, polychoric correlations estimate associations between the latent responses and generally provide a more appropriate correlation structure for ordinal factor models than Pearson correlations [2], [3], particularly when category distributions are imbalanced [5]. Nevertheless, polychoric estimation can become unstable when categories are sparse, which may lead to non–positive-definite correlation matrices and downstream estimation problems [3].

2.2 Robust Ordinal CFA Estimators: WLS, DWLS, and WLSMV
After estimating a polychoric correlation matrix, ordinal CFA is commonly estimated using weighted least squares (WLS) criteria. Full WLS is asymptotically efficient, but it requires estimation of a high-dimensional weight matrix that is difficult to stabilize in typical sample sizes. This has motivated DWLS

and WLSMV as practical alternatives for ordinal CFA [6], [7]. Simulation evidence shows that DWLS and WLSMV often improve parameter recovery and fit compared with continuous-data estimators, especially when indicators are ordinal and distributions are skewed [4], [6], [7]. However, performance depends on sample size and threshold imbalance: under small N or extreme skew, robust ordinal estimators may still exhibit instability or inflated associations and therefore benefit from additional validation [7], [8].

2.3 Stability Assessment and Reporting Practices

Because ordinal CFA depends on estimated thresholds and polychoric correlations, uncertainty can propagate into loadings, fit indices, and convergence behavior. Nonparametric bootstrapping provides an empirical method for assessing stability by repeatedly resampling cases, refitting the model, and summarizing convergence rates and confidence intervals for key parameters [10]. Comparative work emphasizes that bootstrap confidence intervals for standardized loadings can reveal instability not apparent from point estimates alone and should be reported, particularly in small-to-moderate samples [11]. For model fit reporting best practice is to report multiple indices (e.g., CFI/TLI with RMSEA/SRMR) rather than relying on a single statistic, using established guidelines for interpretation [9].

2.4 Machine Learning and Deep Learning for Ordinal Prediction and Representation

Machine learning methods provide flexible modeling of nonlinear patterns and interactions that linear CFA does not represent. Tree ensembles such as Random Forests and Gradient Boosting Machines are strong baselines for structured/tabular prediction tasks [12], [13]. XGBoost is a widely used scalable boosting implementation that often performs strongly on high-dimensional predictor sets and provides importance measures that can support exploratory assessment of item relevance [14]. Deep learning offers representation learning approaches that compress observed indicators into low-dimensional embeddings; variational autoencoders (VAE) learn latent representations using probabilistic regularization and can capture complex dependence structures [15]. While these methods can improve predictive performance, they do not natively produce confirmatory measurement outputs (e.g., loadings, thresholds, and fit indices), which limits their direct use for theory-driven construct validation.

2.5 Summary and Gap Addressed by This Study

Prior work establishes that robust ordinal CFA estimators (DWLS/WLSMV) are preferable to continuous-data CFA for Likert indicators, but still face instability under small samples and threshold imbalance [6]–[8]. ML/DL approaches can capture nonlinear structure useful for prediction, yet lack confirmatory measurement semantics [12]–[15]. Therefore, applied researchers often need both: (i) robust measurement validation and (ii) predictive modeling under matched validation protocols. This study addresses that gap by providing a unified benchmark of robust ordinal CFA estimators and ML/DL predictive models, using consistent validation procedures and transparent reporting, implemented using standard SEM tooling (e.g., lavaan) [16].

3. METHODOLOGY

3.1 Data Sources

Simulated data: We generated 5-point ordinal indicators from a one-factor latent response model with $p = 6$ indicators. For each sample size $n \in \{250, 500, 1000\}$, we generated an underlying factor $F \sim N(0,1)$ and continuous responses $y_{ij} = \lambda_j F_i + \varepsilon_{ij}$, with $\varepsilon_{ij} \sim N(0, \theta_j)$ and target standardized loadings around 0.7. Observed ordinal responses were obtained by thresholding y into five categories using (i) approximately even thresholds (yielding near-uniform category use) and (ii) uneven thresholds (producing floor/ceiling effects). Each condition was replicated 500 times to summarize fit, recovery, and stability trends.

Real data: The empirical dataset was the Malaysian Green consumption behaviour dataset (n=375) reported by Mohd Ghani et al. The analysis excluded demographic variables to concentrate on ordinal measurement and prediction.

3.2 CFA Estimators and Model Fit

CFA models were estimated using WLS, WLSMV, and DWLS on polychoric correlation matrices [4],[5]. Fit was evaluated using CFI, TLI, RMSEA, and SRMR, following common cutoffs [15]. Estimation was implemented in R using lavaan [16].

3.3 Predictive Models (ML/DL)

Machine learning models included Random Forest (RF), Support Vector Machine (SVM with RBF kernel), and Gradient Boosting Machine (GBM). RF models used ntree = 500 with mtry tuned by grid search; SVM tuned C and γ on log-scaled grids; GBM tuned learning rate (η), maximum depth, and number of boosting rounds with early stopping. Models were evaluated using 5-fold cross-validation with R^2, RMSE, and MAE metrics [10],[11],[12]. Deep learning baselines included a Dense Neural Network (DNN), a CNN, and an RNN. The DNN used two hidden layers with dropout (0.2–0.5) and early stopping on validation loss; CNN/RNN used minimal architectures as baselines. All neural models were tuned via constrained random search with early stopping to reduce overfitting [13].

3.4 Validation Procedures

Estimator stability for CFA was assessed with a nonparametric bootstrap (B=500 resamples) on the real dataset to compute convergence rates and average 95% CI widths for standardized loadings [8],[9]. Predictive models were assessed with 5-fold cross-validation, and all models were evaluated on consistent splits for fair comparison.

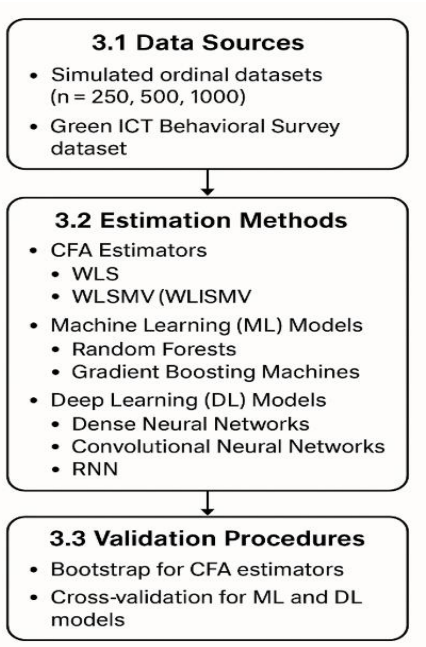

Fig. 1. Overview of methodology

4. RESULTS

4.1 CFA Model Fit Across Estimators

Table 1 reports fit indices across estimators and sample sizes. At n=250 (simulated), WLS yielded poorer fit (CFI≈0.94; RMSEA≈0.08) relative to WLSMV (CFI≈0.97; RMSEA≈0.05) and DWLS (CFI≈0.96; RMSEA≈0.06). By n=1000, performance differences diminished. In the real dataset (n=375), RMSEA increased across estimators, with WLSMV and DWLS remaining superior to WLS.

Parameter recovery in simulation corroborated the fit results. At n = 250, WLS showed moderate bias (e.g., underestimating a low-loading item, $\hat{\lambda} \approx 0.40$ vs. true $\lambda = 0.50$) and factor correlations deviated from generating values by ≈ 0.10. Robust estimators reduced these errors (mean absolute loading error ≈ 0.03). The correlation between true and estimated factor scores improved from r ≈ 0.92 (WLS) to r ≈ 0.96 (WLSMV) and r ≈ 0.95 (DWLS). By n = 500, factor-score correlations exceeded r > 0.98 for all estimators.

Table 1. CFA model fit indices by estimator and sample size (simulated n=250, 500, 1000; real n=375).

Sample	Estimator	CFI	TLI	RMSEA	SRMR
250 (sim)	WLS	0.94	0.92	0.08	0.07
250 (sim)	WLSMV	0.97	0.96	0.05	0.04
250 (sim)	DWLS	0.96	0.95	0.06	0.05
500 (sim)	WLS	0.96	0.95	0.05	0.05
500 (sim)	WLSMV	0.99	0.99	0.03	0.03
500 (sim)	DWLS	0.98	0.98	0.04	0.03
1000 (sim)	WLS	0.99	0.99	0.02	0.02
1000 (sim)	WLSMV	1.00	1.00	0.01	0.01
1000 (sim)	DWLS	0.99	0.99	0.01	0.01
375 (real)	WLS	0.92	0.90	0.08	0.07
375 (real)	WLSMV	0.95	0.93	0.06	0.05
375 (real)	DWLS	0.93	0.91	0.07	0.06

Table note: Fit indices computed using polychoric correlations. Cutoff guidelines follow [15].

Table 1. shows At n=250, robust estimators achieve acceptable-to-good fit (e.g., WLSMV CFI≈0.97; RMSEA≈0.05) relative to WLS (CFI≈0.94; RMSEA≈0.08). By n=1000, differences shrink, indicating large samples mitigate weighting instability and threshold noise.

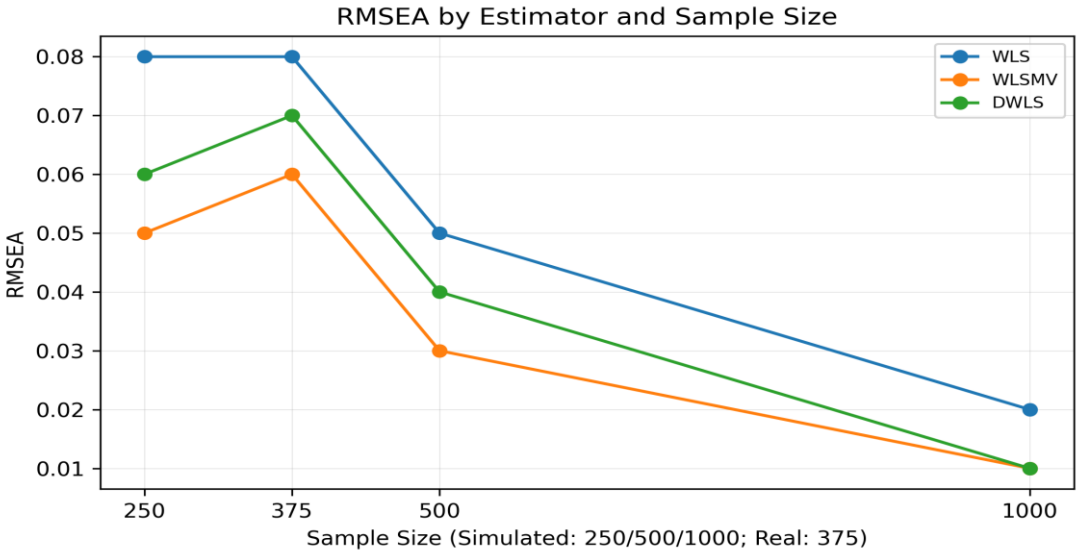

Fig. 2. RMSEA across estimators and sample sizes (simulated & real).

Fig. 2 shows that RMSEA decreases as sample size increases, and WLSMV consistently yields the lowest RMSEA, reflecting superior absolute fit in small-to-moderate samples.

4.2 Bootstrap Stability of Loadings (Real Data)

Bootstrap resampling (500 resamples) quantified estimator stability. WLS converged in 94% of resamples, whereas WLSMV and DWLS converged in virtually all. Mean 95% CI widths for standardized loadings were ~0.20 (WLS), ~0.14 (WLSMV), and ~0.15 (DWLS), indicating improved precision for robust estimators.

Table 2. Bootstrap convergence and CI width for loadings (real data; 500 resamples).

Estimator	Bootstrap Convergence (%)	Mean CI Width (Loading)
WLS	94%	0.20
WLSMV	100%	0.14
DWLS	100%	0.15

Table note: CI width summarizes average 95% bootstrap intervals for standardized loadings [8],[9].

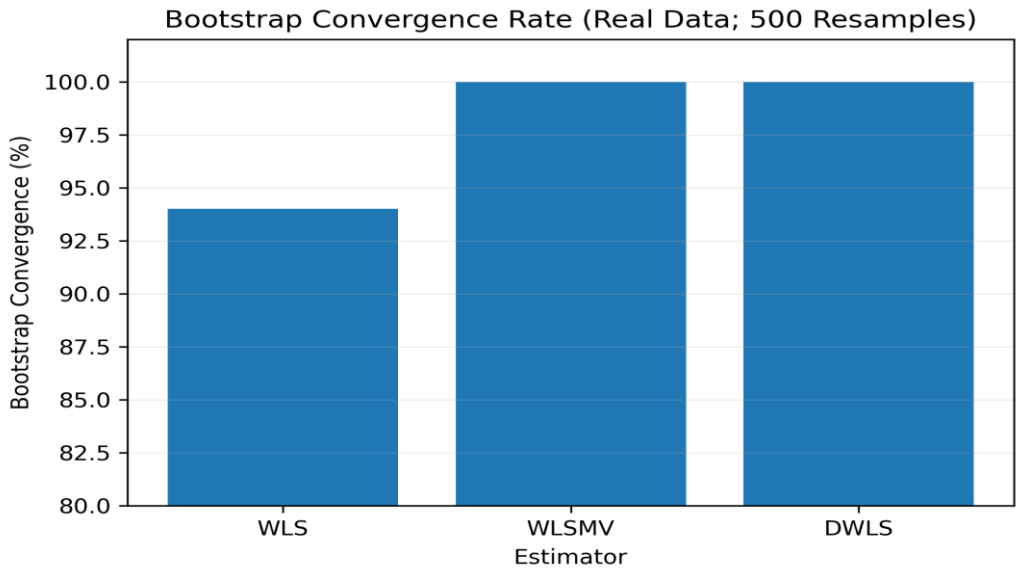

Fig. 3. Bootstrap convergence rate by estimator (real data; 500 resamples).

Fig.3 confirms that WLSMV yields the tightest loading intervals, while WLS shows the widest intervals, reflecting higher sampling variability.

4.3 Predictive Modeling Results (Real Data)

Table 3 reports 5-fold cross-validated performance for ML and DL models predicting Green ICT behavior. DNN achieved the

highest accuracy ($R^2 \approx 0.35$), followed by GBM ($R^2 \approx 0.33$). CNN and RNN baselines did not improve prediction for tabular data.

Table 3. Cross-validated prediction performance (n=375; 5-fold CV).

Model	R^2	RMSE	MAE
Random Forest (RF)	0.30	6.3	4.8
Support Vector Machine (SVM)	0.28	6.5	5.0
Gradient Boosting Machine (GBM)	0.33	6.1	4.5
Deep Neural Network (DNN)	0.35	5.9	4.4
Convolutional Neural Network (CNN)	0.26	6.7	5.1
Recurrent Neural Network (RNN)	0.27	6.6	5.0

Table note: Metrics averaged across 5 folds; tuning and validation follow [10],[11],[12],[13].

Table 3 reports 5-fold cross-validated performance for ML and DL models predicting Green ICT behavior. DNN achieved the highest accuracy ($R^2 \approx 0.35$), followed by GBM ($R^2 \approx 0.33$). CNN and RNN baselines did not improve prediction for tabular data.

5. DISCUSSION

Results show that robust ordinal estimators (WLSMV, DWLS) provide better fit and greater stability than WLS in small-to-moderate samples. Bootstrap resampling confirms improved convergence and tighter loading intervals for WLSMV/DWLS. In prediction, nonlinear learners (DNN, GBM) achieve the strongest performance under cross-validation, suggesting that nonlinear structure can be exploited for behavioral outcomes.

Taken together, the results confirm two practical conclusions. First, for measurement modeling with ordinal indicators, WLSMV (and DWLS) offer superior stability and fit relative to WLS, especially in small-to-moderate samples. Second, for prediction tasks, boosting and neural networks deliver strong performance in cross-validation, suggesting that real datasets often contain nonlinear structure beyond the linear CFA approximation. Applied researchers should therefore pair robust ordinal CFA with resampling-based stability checks and, where prediction is central, consider complementary ML/DL models evaluated with cross-validation.

Practically, the findings suggest a workflow in which (i) robust ordinal CFA (WLSMV or DWLS) is used for measurement validation and reporting, supported by bootstrap stability checks, and (ii) ML/DL models are used when prediction is central, evaluated under matched cross-validation. This paired strategy enables researchers to preserve interpretability while exploiting nonlinear predictive structure when needed.

For predictive tasks, the cross-validated results (Fig. 3 and Table 3) show that nonlinear learners—especially GBM/XGBoost and DNN—achieve higher R^2 than linear baselines, suggesting that real datasets contain interaction and nonlinear structure beyond the linear CFA approximation [10],[12]. However, these models do not yield confirmatory measurement parameters such as loadings, thresholds, and fit indices. Therefore, ML/DL should be viewed as complementary tools for prediction, while robust ordinal CFA remains the appropriate framework for theory-driven measurement validation.

Bootstrap results strengthen these conclusions by quantifying sampling uncertainty rather than relying on single-sample point estimates. The higher bootstrap convergence rates and narrower loading confidence intervals for WLSMV/DWLS indicate better numerical robustness and more precise measurement under sampling variability [8],[9]. In applied ordinal CFA, this implies that reporting bootstrap intervals is not optional but rather an important diagnostic of estimator reliability, especially when category imbalance is present.

The fit trends in Fig. 2 (RMSEA) highlight an important practical point: differences among estimators are largest in small samples and diminish as sample size increases. At n=250, robust estimators provide an acceptable-to-good fit (CFI/TLI closer to 0.95 and RMSEA closer to 0.05), whereas WLS can be borderline under typical cutoffs [15]. By n=1000, all estimators converge toward an excellent fit, indicating that large samples partially compensate for weighting instability and threshold noise.

From a measurement perspective, the superiority of WLSMV and DWLS over WLS is expected because robust ordinal estimators rely on polychoric associations and reduced/adjusted weighting schemes that are less sensitive to small-sample instability in the full weight matrix [6],[7],[14] [16]. In contrast, full WLS requires reliable estimation of a high-dimensional weight matrix and may become numerically fragile when categories are sparse or thresholds are highly imbalanced, leading to poorer fit and reduced convergence in resampling scenarios.

6. CONCLUSION

This paper provides a matched comparison of robust ordinal CFA estimators and predictive ML/DL models for ordinal survey data. Robust estimators (WLSMV/DWLS) are recommended for ordinal CFA, especially in small-to-moderate samples, and resampling-based validation should be reported to quantify stability. For predictive tasks, DNN and GBM offer strong performance under cross-validation. Future work should investigate integrated hybrid workflows that combine measurement and prediction within a single framework.

Overall, the findings support two practical takeaways: (i) for theory-driven measurement with Likert indicators, WLSMV/DWLS yield more reliable fit and more stable loading estimates than WLS, particularly under threshold imbalance; and (ii) for outcome prediction, nonlinear learners can exploit structure not captured by linear CFA, but they should complement—not replace—confirmatory measurement. We recommend reporting multiple fit indices alongside bootstrap confidence intervals for key loadings, and using cross-validation for predictive models to ensure fair and reproducible comparisons.

7. REFERENCES

[1] A. Agresti, Analysis of Ordinal Categorical Data, 2nd ed. Wiley, 2010.

[2] D. B. Flora and P. J. Curran, "An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data," *Psychological Methods*, vol. 9, no. 4, pp. 466–491, 2004.

[3] F. P. Holgado-Tello, S. Chacón-Moscoso, I. Barbero-García, and E. Vila-Abad, "Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables," *Quality & Quantity*, vol. 44, no. 1, pp. 153–166, 2010.

[4] U. Olsson, "Maximum likelihood estimation of the polychoric correlation coefficient," *Psychometrika*, vol. 44, no. 4, pp. 443–460, 1979.

[5] K. G. Jöreskog, "On the estimation of polychoric correlations and their asymptotic covariance matrix," *Psychometrika*, vol. 59, no. 3, pp. 381–389, 1994.

[6] C. G. Forero, A. Maydeu-Olivares, and D. Gallardo-Pujol, "Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation," *Structural Equation Modeling*, vol. 16, no. 4, pp. 625–641, 2009.

[7] C.-H. Li, "Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares," *Behavior Research Methods*, vol. 48, no. 3, pp. 936–949, 2016.

[8] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. Chapman & Hall/CRC, 1993.

[9] C. DiStefano and G. B. Morgan, "A comparison of diagonally weighted least squares robust estimation techniques for ordinal data," *Structural Equation Modeling*, vol. 21, no. 3, pp. 425–438, 2014.

[10] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2009.

[11] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. KDD, pp. 785–794, 2016.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. ICLR, 2014.

[14] D. L. Bandalos, "Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation," *Structural Equation Modeling*, vol. 21, no. 1, pp. 102–116, 2014.

[15] L.-T. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Structural Equation Modeling*, vol. 6, no. 1, pp. 1–55, 1999.

[16] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012.