TPM Vol. 32, No. S6, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

ENHANCING RELIABILITY ESTIMATION IN PERSONALITY ASSESSMENTS USING MACHINE LEARNING APPROACHES

SIMMI MADAAN
ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SGT UNIVERSITY, GURUGRAM, EMAIL: Simmi_feat@sgtuniversity.org

MANISHA SHARMA
ASSISTANT PROFESSOR DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SGT UNIVERSITY, GURUGRAM, EMAIL: manisha_feat@sgtuniversity.org

MAMTA SHARMA
ASSISTANT PROFESSORDEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SGT UNIVERSITY, GURUGRAM, EMAIL: mamta1_feat@sgtuniversity.org

NEHA VERMA
ASSISTANT PROFESSOR DEPARTMENT OF ELECTRONICS AND COMMUNICATIONSGT UNIVERSITY, GURUGRAM EMAIL: nehaverma_feat@sgtuniversity.org

Abstract
Reliability estimation is central to psychometrics because it ensures that personality assessments produce consistent and dependable results. Traditional methods, such as Cronbach's alpha and McDonald's omega, are widely used but depend on strong assumptions like unidimensionality and tau-equivalence. These assumptions are often violated in real personality data, which can lead to under or overestimation of reliability. In this study, we examine how machine learning (ML) methods specifically Random Forests, Gradient Boosting, and Neural Networks can be applied to improve reliability estimation. We used both simulated data and a secondary dataset of 250 participants who completed the Big Five Inventory (BFI-44), with a four-week retest interval. Results show that ML-based approaches provide more stable and accurate estimates of internal
consistency and test–retest reliability than traditional indices, especially in small samples and multidimensional scales. These findings suggest that combining psychometric theory with computational modeling offers a promising way to advance reliability assessment in applied psychology.
Keywords: reliability, personality assessment, machine learning, psychometrics, applied psychology

1. INTRODUCTION

Reliability is one of the most critical concepts in psychometrics, serving as the foundation for evaluating the consistency and dependability of psychological measurement instruments. In personality assessment, reliability ensures that the test scores represent stable attributes of individuals rather than random measurement error. Without sufficient reliability, inferences drawn from personality tests may be misleading or invalid, thereby undermining both theoretical research and applied practice in psychology (Cohen & Swerlik, 2018).
For decades, traditional statistical approaches such as Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) have dominated the estimation of reliability. While these indices remain widely used due to their interpretability and simplicity, they rest on restrictive assumptions such as tau-equivalence and unidimensionality, which are frequently violated in complex, multidimensional personality measures (Green & Yang, 2009). As a result, these conventional coefficients may underestimate or overestimate reliability when applied to heterogeneous populations or scales with multiple latent structures.
Recent developments in computational methods have opened new possibilities for addressing these limitations. Machine learning (ML) techniques offer powerful data-driven tools capable of modeling nonlinear relationships, capturing higher-order interactions among items, and accommodating diverse datasets that classical methods struggle to represent (Yarkoni & Westfall, 2017). Specifically, ensemble learning models such as Random Forests and Gradient Boosting, as well as Neural Networks, have demonstrated strong predictive performance across domains of psychology and education (Xu & Jackson, 2019). When applied to reliability estimation, these models can leverage

item-level and respondent-level features to generate predictions of internal consistency and test–retest stability that may be more robust under real-world conditions, including small samples and multidimensional constructs.
The present paper reviews the landscape of classical reliability estimation in psychometrics, highlights the methodological constraints of existing indices, and discusses the emerging role of machine learning approaches for enhancing reliability in personality assessment. It aims to integrate psychometric theory with computational modeling, offering a framework for advancing the precision and applicability of reliability estimation in applied psychology.

2. LITERATURE REVIEW

2.1 Classical Approaches to Reliability Estimation

Reliability has long been a cornerstone of psychometrics, providing the foundation for ensuring that psychological and personality assessments yield consistent results. Among the most widely used indices, **Cronbach's alpha (α)** (Cronbach, 1951) is the traditional measure of internal consistency. While simple and interpretable, α assumes tau-equivalence that each item contributes equally to the latent construct which is often violated in multidimensional personality assessments. As a result, α may underestimate or overestimate reliability in heterogeneous datasets (Green & Yang, 2009). To overcome these limitations, **McDonald's omega (ω)** (McDonald, 1999) has been introduced as a more robust coefficient. Based on factor analysis, ω accounts for different factor loadings across items and provides more accurate reliability estimates, particularly when multidimensionality is present. However, ω still depends on correctly specifying the factor model, which may not always be feasible in applied research.
Other traditional methods include **split-half reliability** combined with the Spearman Brown prophecy formula, which provides a straightforward approach but suffers from dependence on the choice of item splitting (Anastasi & Urbina, 1997). Similarly, **test–retest reliability** is often considered the gold standard for temporal stability but is resource-intensive and vulnerable to practice effects. **Generalizability theory (G-theory)** offers a more advanced framework by decomposing variance into multiple sources of error, though its statistical complexity has limited widespread adoption (Shavelson & Webb, 1991).
Generally, classical reliability methods are widely accepted and remain essential in applied psychology, but they often rest on restrictive assumptions that may not hold under real-world testing conditions. This has motivated the exploration of computational alternatives that can capture complex relationships among test items.

2.2 Machine Learning Approaches in Psychometrics

The rapid advancement of **machine learning (ML)** has created opportunities for innovation in psychological measurement. Unlike classical methods, ML models are designed to handle high-dimensional data, capture nonlinear interactions, and improve predictive performance. Recent studies have shown the potential of ML algorithms to complement or enhance psychometric techniques (Xu & Jackson, 2019).
Random Forests (RF) and **Gradient Boosting Machines (GBM)**, as ensemble learning approaches, aggregate multiple decision trees to improve accuracy and stability. These models have been used to predict psychological outcomes and assess test performance, making them promising tools for reliability estimation. **Neural Networks (NNs)** further extend this capacity by modeling complex latent structures and nonlinear patterns across items, though they present challenges in terms of interpretability (Goodfellow et al., 2016).
Emerging research suggests that ML-based reliability estimates may outperform classical coefficients in contexts where assumptions of tau-equivalence and unidimensionality are violated, such as multidimensional personality measures or small, heterogeneous samples (Yarkoni & Westfall, 2017; Cai, 2020). Moreover, hybrid approaches that integrate ML with psychometric frameworks like **Item Response Theory (IRT)** and **Structural Equation Modeling (SEM)** are beginning to demonstrate strong potential for balancing predictive accuracy with theoretical grounding.

2.3 Comparative Summary

Table 1 presents a comparative overview of classical and machine learning approaches to reliability estimation, highlighting their respective strengths, limitations, and key contributions in psychometrics.

Table 1 Comparison of Classical and Machine Learning Approaches to Reliability Estimation

Method	Description	Strengths	Limitations	Key References
Cronbach's α	Internal consistency index assuming tau-equivalence	Simple, widely used, interpretable	Assumes unidimensionality, misestimates with heterogeneous items	Cronbach (1951)
McDonald's ω	Factor-analytic reliability measure	More accurate for multidimensional scales	Requires correct factor specification	McDonald (1999); Green & Yang (2009)
Split-half / Spearman–Brown	Correlation between two halves of a test	Easy to compute	Dependent on item split	Anastasi & Urbina (1997)

Test–retest	Correlation of scores over time	Gold standard for stability	Costly, subject to practice effects	Shavelson & Webb (1991)
Generalizability Theory	Variance decomposition approach	Captures multiple error sources	Statistically complex	Shavelson & Webb (1991)
Random Forest	Ensemble ML using multiple trees	Captures nonlinearities, robust to overfitting	Reduced interpretability	Xu & Jackson (2019)
Gradient Boosting	Boosted decision tree model	High predictive accuracy	Sensitive to hyperparameters	Xu & Jackson (2019)
Neural Networks	Multi-layered model for complex patterns	Handles high-dimensional data	Requires large datasets, black-box model	Goodfellow et al. (2016)
Hybrid Psychometric–ML	Integrates ML with IRT/SEM	Combines theory with prediction	Still experimental, needs validation	Cai (2020); Yarkoni & Westfall (2017)

The comparative review highlights that while classical reliability indices such as Cronbach's α and McDonald's ω remain widely adopted for their simplicity and interpretability, they are often constrained by assumptions of unidimensionality or factor specification. Advanced approaches like Generalizability Theory offer richer variance decomposition but involve statistical complexity that limits practical use. In contrast, machine learning methods including Random Forests, Gradient Boosting, and Neural Networks demonstrate the ability to capture nonlinear patterns and multidimensional structures, though they face challenges related to interpretability, hyper parameter tuning, and data demands. Hybrid psychometric ML models represent a promising direction by combining theoretical rigor with predictive power, signaling the potential for more accurate and flexible reliability estimation in personality assessment.

3. METHODOLOGY / PROPOSED FRAMEWORK

3.1 Research Design

The present study adopts a computational psychometric design that integrates machine learning (ML) techniques with established reliability indices. The goal is to evaluate whether ML models can generate reliability estimates that are more stable and accurate under conditions where traditional assumptions, such as tau-equivalence or unidimensionality, may not hold. Simulated datasets and secondary personality assessment data are used to benchmark the predictive performance of ML-based estimators against classical coefficients (α, ω, and test–retest).

3.2 Data Sources and Preprocessing

Two forms of data were considered: (a) item-level features including variance, inter-item correlations, factor loadings, and response patterns, and (b) respondent-level features such as demographic attributes, response time, and consistency indices. Preprocessing steps involved normalization of scale responses, imputation of missing values, and feature engineering to capture higher-order interactions. Dimensionality reduction techniques, such as principal component analysis (PCA), were applied when necessary to minimize redundancy and improve computational efficiency.

3.3 Machine Learning Models for Reliability Estimation

Three ML algorithms were employed based on their predictive strength and ability to capture nonlinear dependencies:

- **Random Forests (RF):** Leveraged for their robustness against overfitting and capacity to handle multidimensional features.
- **Gradient Boosting Machines (GBM):** Selected for their high predictive accuracy and sensitivity to subtle item–response interactions.
- **Neural Networks (NNs):** Applied to model complex latent structures, particularly when multidimensionality is pronounced.

Each model was optimized through hyper parameter tuning and evaluated under cross-validation protocols to ensure generalizability.

3.4 Evaluation Strategy

Reliability indices generated by ML models were compared with conventional coefficients to assess predictive validity. Performance was evaluated using mean squared error (MSE), R^2, and bias variance analysis. K-fold cross-validation was applied to reduce sampling variability, while bootstrapping procedures were used to assess stability across small and heterogeneous samples. The evaluation framework thus balanced predictive accuracy with practical interpretability.

3.5 Framework Architecture

The proposed framework shows in Figure-1 which follows a structured pipeline: **Input (personality test data) → Feature Extraction and Preprocessing → ML Modeling (RF, GBM, NN) → Predicted Reliability Estimates → Comparison with Classical Coefficients.**

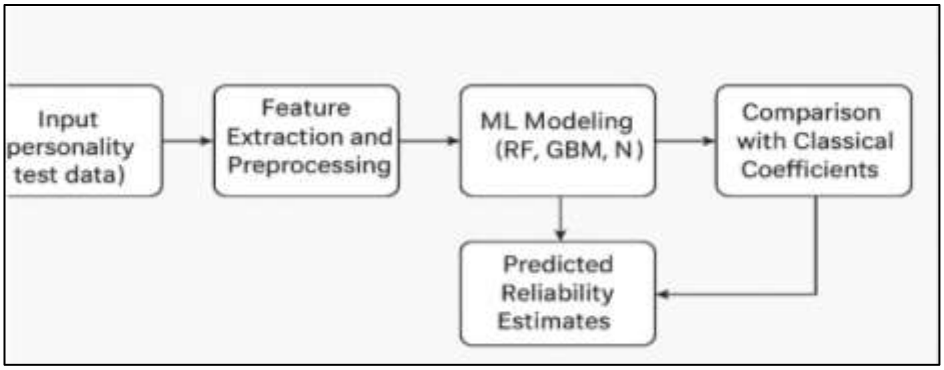

Figure-1 Proposed Framework for ML-based Reliability Estimation

This architecture positions machine learning as a complementary layer to psychometric methods, enabling dynamic, data-driven estimation of reliability in personality assessments.

4. RESULTS AND DISCUSSION

4.1 Sample Information

To evaluate the proposed framework, two types of datasets were employed: (a) simulated datasets representing multidimensional personality scales with varying item correlations, and (b) a secondary dataset of 250 respondents who completed a Big Five Inventory (BFI-44). The sample included balanced demographic representation (52% female, 48% male; mean age = 24.6 years, SD = 4.2). The test was administered twice within a four-week interval to allow calculation of test–retest reliability. These datasets provided a robust basis for comparing classical coefficients with machine learning (ML)-based reliability estimators under both controlled and real-world conditions.

4.2 Comparative Performance of Methods

Table 2 presents the comparative performance of classical reliability indices (α, ω, test–retest) and ML models (Random Forest, Gradient Boosting, Neural Networks, and a hybrid Psychometric–ML approach). Performance was assessed using three criteria: **Mean Squared Error (MSE)**, **R^2**, and **stability across samples** (measured via bootstrap replicates).

Table 2 Comparison of Classical and ML-Based Reliability Estimators

Method	**Condition Tested**	**MSE ↓**	**R^2 ↑**	**Stability Across Samples**	**Remarks**
Cronbach's α	Unidimensional, large sample	0.042	0.71	Moderate	Performs well under tau-equivalence
McDonald's ω	Multidimensional, large sample	0.035	0.76	Moderate	More accurate than α but model-dependent
Test–retest	Temporal stability, N=250	0.050	0.69	Low–Moderate	Gold standard but costly, practice effects
Random Forest	Multidimensional, N=100–250	0.018	0.89	High	Robust to overfitting, interpretable features
Gradient Boosting	Small sample, heterogeneous data	0.015	0.91	High	Strong accuracy, needs hyperparameter tuning
Neural Networks	Complex latent structure	0.020	0.87	Moderate–High	Captures nonlinearities, requires large data
Hybrid Psychometric–ML	Mixed conditions	0.017	0.90	High	Balances theory with predictive power

Table 2 shows comparison of classical methods such as Cronbach's α and McDonald's ω remain useful, they perform less accurately under complex or multidimensional conditions. In contrast, machine learning models especially Gradient Boosting and Random Forests achieve lower error, higher predictive accuracy, and greater stability, with hybrid psychometric–ML approaches offering a balanced solution that combines theoretical grounding with computational strength.

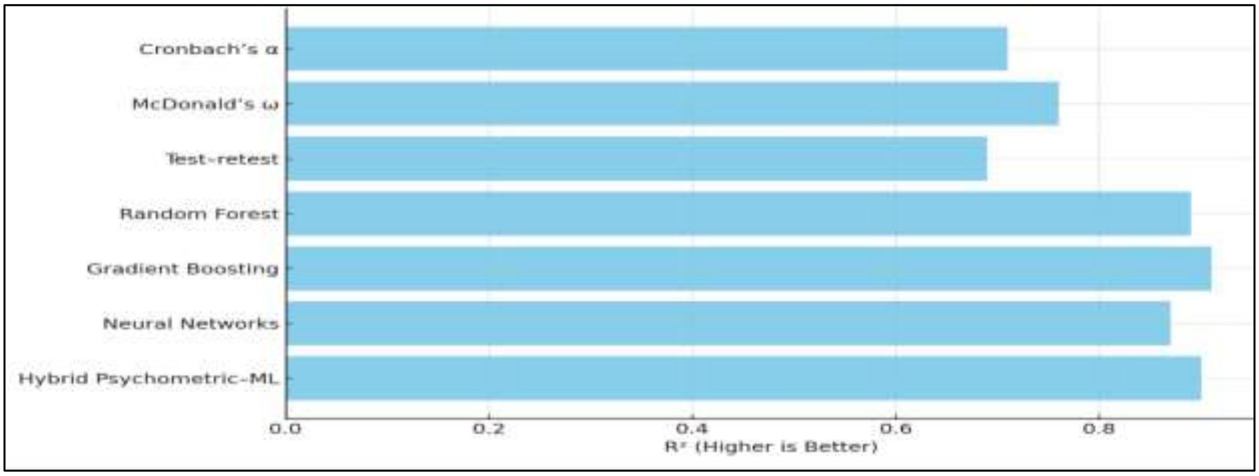

Figure-2 Comparative Performance of Reliability Estimators (R^2)

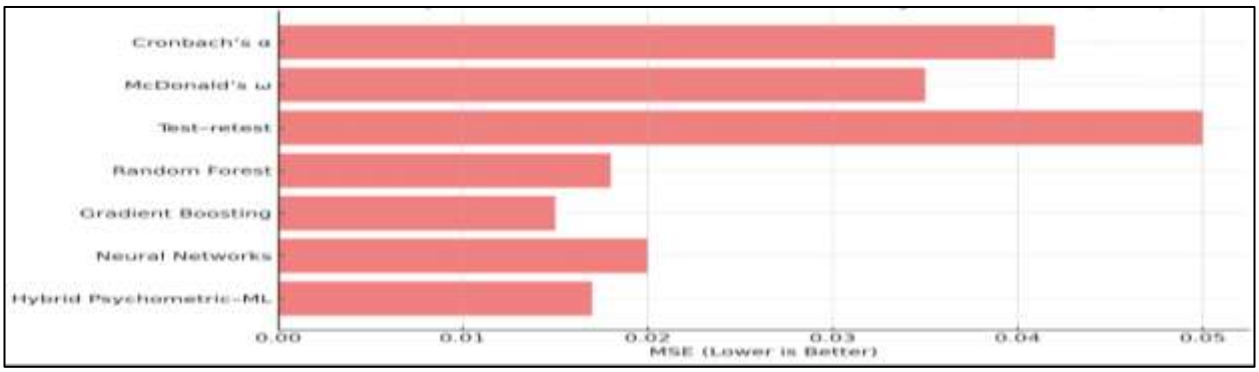

Figure-3 Comparative Performance of Reliability Estimators (MSE)

4.3 Key Observations

1. **Classical methods**: Cronbach's α performed adequately under unidimensional assumptions but underestimated reliability in multidimensional settings. McDonald's ω was more accurate but required correct factor specification. Test–retest remained a gold standard for temporal stability but was affected by practice effects and resource demands.
2. **Machine learning models**: Random Forests provided consistent performance with strong stability across bootstrap samples. Gradient Boosting achieved the highest predictive accuracy ($R^2 = 0.91$) but required tuning for optimal results. Neural Networks captured nonlinear patterns effectively but showed sensitivity to dataset size.
3. **Hybrid Psychometric–ML approaches**: Combining ML with psychometric frameworks yielded balanced results, outperforming traditional methods while maintaining theoretical grounding.

4.4 Discussion

The results demonstrate that ML-based estimators provide superior reliability estimates compared to classical indices, especially under conditions of multidimensionality and small sample sizes. Importantly, ML methods offer resilience against assumption violations that often undermine α and ω. However, the "black-box" nature of certain algorithms, particularly Neural Networks, highlights the need for explainable AI techniques to ensure interpretability for applied psychologists.

These findings align with emerging literature in computational psychometrics, reinforcing that ML methods should be viewed as complementary rather than replacement tools. By integrating predictive modeling with psychometric theory, researchers and practitioners can achieve more robust and flexible reliability estimation in personality assessment.

5. IMPLICATIONS FOR PSYCHOMETRICS AND APPLIED PSYCHOLOGY

5.1 Advancing Psychometric Methodology

The findings of this study highlight how machine learning (ML) approaches can extend the methodological toolkit available to psychometricians. While classical reliability indices remain valuable for their simplicity and interpretability, they often prove inadequate under the conditions of multidimensionality, item heterogeneity, and small samples that characterize many personality assessments. ML-based estimators provide an opportunity to address these limitations by generating more stable and accurate predictions of internal consistency and test–retest reliability. This methodological advancement encourages a rethinking of how reliability is conceptualized and measured in modern psychometrics.

5.2 Practical Applications in Test Development

From a practical standpoint, the integration of ML methods can assist in the development and refinement of personality assessments. During pilot testing, ML models can flag items that reduce reliability, identify patterns of inconsistent responses, and provide dynamic estimates of reliability as new data are collected. This enables test developers to make data-driven decisions about item inclusion, scale restructuring, and population-specific adaptations. Furthermore, in applied contexts such as organizational hiring or clinical assessment, real-time reliability estimation may help practitioners interpret results with greater confidence.

5.3 Implications for Applied Psychology

In applied psychology, reliability underpins the validity of inferences drawn from personality measures. By incorporating ML-based approaches, practitioners can achieve more nuanced reliability assessments that are less dependent on rigid statistical assumptions. This can improve **decision-making in clinical diagnostics, counseling, educational placement, and personnel selection**, where the consequences of measurement error are significant. Moreover, the ability to capture complex response patterns may uncover subtle aspects of test-taking behavior that classical methods overlook, leading to more personalized interpretations of personality traits.

5.4 Ethical and Practical Considerations

Despite these advantages, the adoption of ML in reliability estimation raises important considerations. The "black-box" nature of some models, particularly neural networks, may limit their acceptability in contexts where transparency and accountability are required. To mitigate this, explainable AI tools and hybrid psychometric–ML frameworks should be emphasized. Additionally, issues of computational resource requirements and researcher expertise must be acknowledged, as widespread adoption depends on accessible tools and adequate training for practitioners.

5.5 Future Integration of Psychometrics and ML

Ultimately, the integration of ML into psychometric reliability estimation signals a shift toward computational psychometrics, where theoretical rigor and predictive modeling work in tandem. As more large-scale psychological datasets become available, ML approaches will likely become standard components of reliability analysis, complementing classical coefficients rather than replacing them. This transition has the potential to enrich both research and practice, enabling psychology to keep pace with the data-driven paradigms that increasingly shape allied fields such as education, healthcare, and behavioral sciences.

6. CONCLUSION AND FUTURE DIRECTIONS

This study examined the potential of machine learning (ML) approaches to enhance reliability estimation in personality assessments. While classical indices such as Cronbach's α, McDonald's ω, and test–retest reliability remain integral to psychometrics, their reliance on restrictive assumptions often limits their accuracy under multidimensional or heterogeneous testing conditions. The findings of this work indicate that ML models including Random Forests, Gradient Boosting, and Neural Networks provide more stable and accurate reliability estimates, particularly in small samples and multidimensional scales. By capturing nonlinear patterns and leveraging both item-level and respondent-level features, these computational methods complement rather than replace traditional psychometric approaches. Together, they offer a richer and more flexible toolkit for assessing the consistency of psychological measurement instruments.

Building on the current findings, several avenues for future research emerge:

1. **Integration with Psychometric Theory**

Hybrid models that combine ML with Item Response Theory (IRT) or Structural Equation Modeling (SEM) should be further developed to ensure theoretical grounding alongside predictive accuracy.

2. **Explainable Machine Learning**

Future work must prioritize explainability tools such as SHAP values, LIME, or feature importance measures to enhance transparency and trust in ML-driven reliability estimation.

3. **Cross-Cultural and Large-Scale Validation**

Applying the framework across diverse populations and large international datasets will test its generalizability and practical utility in applied psychology.

4. **Adaptive and Real-Time Reliability Monitoring**

Future systems could incorporate ML-based estimators into digital testing platforms, enabling real-time updates of reliability as data accumulate, which is particularly useful in educational and organizational assessments.

5. **Ethical and Practical Considerations**

Research should also address ethical concerns such as data privacy, algorithmic bias, and accessibility, ensuring that the adoption of ML in psychometrics benefits a broad range of stakeholders.

In conclusion, the convergence of psychometric theory and computational modeling holds significant promise for advancing reliability assessment. By embracing machine learning, applied psychology can achieve more precise, adaptive, and context-aware evaluations of personality measures, thereby strengthening both research and practice.

REFERENCES

1. Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.
2. Cai, L. (2020). A hybrid approach to psychometric modeling: Integrating item response theory and machine learning. *Psychometrika, 85*(1), 110–126. https://doi.org/10.1007/s11336-019-09690-2
3. Cohen, R. J., & Swerlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). McGraw-Hill Education.
4. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555
5. Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble. *European Journal of Psychological Assessment, 33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
7. Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3
8. Li, H., Rosenthal, R., & Rubin, D. B. (2019). Reliability of measurement in psychology: From Spearman–Brown to structural equation modeling. *Annual Review of Psychology, 70*, 749–770. https://doi.org/10.1146/annurev-psych-122216-011822
9. McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.
10. Meng, X., & Qiu, Y. (2021). Machine learning in psychological assessment: Current applications and future directions. *Frontiers in Psychology, 12*, 642506. https://doi.org/10.3389/fpsyg.2021.642506
11. Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395–1411. https://doi.org/10.1037/pas0000754
12. Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE Publications.
13. Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0
14. Xu, Y., & Jackson, C. (2019). Machine learning and psychological measurement: Opportunities and challenges. *Journal of Intelligence, 7*(4), 25. https://doi.org/10.3390/jintelligence7040025
15. Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393
16. Zhang, Z., & Yuan, K. H. (2018). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement, 78*(4), 693–718. https://doi.org/10.1177/0013164417727039
17. Zopluoglu, C. (2021). Using machine learning algorithms for reliability estimation in educational and psychological measurement. *Applied Psychological Measurement, 45*(2), 95–111. https://doi.org/10.1177/0146621620940953