Comparative Analysis of SPSS, Winsteps, and Python in Psychometric Analysis: Applications in the Development and Validation of Educational and Psychological Tests

Nezai Zohra
Laboratory for the Analysis of Quantitative and Qualitative Data on Psychology and Social Behavior – Abdelhamid Ibn Badis University, Mostaganem, Algeria.
https://orcid.org/my-orcid?orcid=0000-0001-6796-2456
Email: zohra.nezai.@univ-mosta.dz

Received: 15/02/2026 ;   Accepted: 27/05/2026 ;   Published: 16/07/2026

Abstract
This study aimed to provide a comprehensive comparative review of three widely used software packages in psychometric analysis: **SPSS**, **Winsteps**, and **Python**. The review examined their theoretical foundations, statistical capabilities, and applications in the development and validation of psychological and educational tests. A descriptive-analytical review methodology was adopted through the analysis of scholarly literature, reference books, and recent empirical studies, The comparison focused on usability, flexibility, psychometric functionality, support for modern measurement models, automation capabilities, and integration with artificial intelligence technologies.
The findings indicate that **SPSS** remains the preferred software for conventional statistical analyses and applications based on Classical Test Theory. **Winsteps** demonstrated superior performance in Rasch measurement, providing advanced tools for item calibration, person ability estimation, fit statistics, and differential item functioning analysis. In contrast, **Python** offers a comprehensive open-source environment that integrates psychometric analysis with data science, machine learning, and artificial intelligence, enabling advanced statistical modeling and automated analytical workflows.
The study concludes that these software packages should be viewed as complementary rather than competing tools, with software selection depending on research objectives, data characteristics, and the underlying psychometric model. It further recommends expanding researchers' training in modern psychometric software, integrating Python and Winsteps into graduate curricula, and conducting empirical comparative studies using real datasets to enhance the quality of psychological and educational measurement research.
Keywords: Psychometric Analysis; SPSS; Winsteps; Python.

Introduction

Psychometrics has witnessed remarkable development over recent decades, driven by advances in statistical theories, computational methods, and software applications dedicated to psychological and educational measurement. Consequently, psychometric analysis has become one of the fundamental pillars in the development, validation, and standardization of psychological and educational instruments. The quality of any measurement instrument is no longer judged solely by its theoretical framework; rather, it also depends on the accuracy of the statistical methods employed to evaluate its psychometric properties, including reliability, validity, item analysis, dimensionality, and measurement precision (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

Historically, psychometric analyses relied primarily on **Classical Test Theory (CTT)**, which has served for decades as the dominant framework for evaluating psychological and educational tests. Numerous statistical software packages, particularly **SPSS**, have facilitated the widespread application of CTT by providing comprehensive procedures for descriptive statistics, reliability estimation, validity assessment, item analysis, and factor analysis. Owing to its graphical interface and ease of use, SPSS has become the most frequently used statistical package among researchers in psychology, education, and the social sciences.

Despite its extensive use, Classical Test Theory has several methodological limitations, particularly its dependence on sample characteristics and the inability to estimate item parameters independently of the respondents. These limitations stimulated the emergence of **Item Response Theory (IRT)**, which introduced probabilistic models capable of estimating both person ability and item characteristics independently under specific model assumptions (Embretson & Reise, 2000).

Among IRT models, the **Rasch Model** has received considerable attention because of its strong measurement philosophy emphasizing objective measurement and invariant parameter estimation. Unlike Classical Test Theory, the Rasch Model evaluates whether empirical data fit the theoretical measurement model rather than merely describing statistical relationships among variables. This feature has made Rasch analysis an essential methodology in contemporary psychometric research and educational assessment (Bond & Fox, 2015).

The growing adoption of Rasch measurement has led to the development of specialized software packages designed specifically for its implementation. Among these, **Winsteps** has become one of the most widely recognized applications, offering advanced procedures for estimating person ability, item difficulty, fit statistics, differential item functioning (DIF), dimensionality assessment, and Wright Maps. Consequently, Winsteps has become a standard tool in psychological measurement, educational assessment, health sciences, and certification testing.

Simultaneously, rapid advances in computer science, artificial intelligence, and data science have introduced a new generation of analytical tools capable of handling increasingly complex datasets. Within this context, **Python** has emerged as one of the most influential programming languages for scientific computing and statistical analysis. Unlike traditional statistical software, Python provides an open-source programming environment that integrates statistical analysis, psychometric modeling, machine learning, artificial intelligence, and data visualization into a single flexible framework. This flexibility enables researchers to implement both traditional psychometric analyses and highly sophisticated computational models that were previously difficult to achieve using conventional statistical software.

Recent developments have also witnessed increasing integration between psychometrics and artificial intelligence. Machine learning algorithms are now employed to improve item calibration, develop computerized adaptive testing (CAT), predict academic performance, detect response patterns, identify biased items, and automate psychometric reporting. These innovations have expanded the role of psychometric software beyond conventional statistical analysis toward intelligent decision-support systems.

Despite the increasing availability of statistical software, researchers often face uncertainty when selecting the most appropriate software package for their psychometric studies. While SPSS remains the dominant software in many universities, Winsteps has become the preferred application for Rasch analysis, whereas Python is rapidly expanding as an integrated environment for modern psychometrics and data science. However, relatively few studies have provided a comprehensive comparison of these three analytical environments from theoretical, methodological, and practical perspectives.

Accordingly, the present study aims to provide a comprehensive comparative review of **SPSS**, **Winsteps**, and **Python** in psychometric analysis by examining their theoretical foundations, statistical capabilities, psychometric applications, strengths, limitations, and future prospects. Furthermore, the study discusses the suitability of each software package for different research purposes and proposes an integrated framework for selecting the most appropriate analytical environment according to research objectives, data characteristics, and contemporary psychometric methodologies.

Theoretical Framework

2.1 Psychometric Analysis: Concept and Importance

Psychometric analysis refers to a set of statistical and methodological procedures designed to evaluate the quality of psychological and educational measurement instruments by examining their psychometric properties, including reliability, validity, item analysis, person ability estimation, and the degree to which empirical data fit theoretical measurement models (DeVellis & Thorpe, 2021). It represents a critical stage in determining whether a measurement instrument accurately and consistently measures the intended construct, thereby directly influencing the quality of educational, psychological, and research decisions based on test scores (AERA et al., 2014).

Psychometric analysis extends beyond the calculation of conventional statistical indices. It also involves the application of mathematical models that describe the relationship between examinees' responses and item characteristics, estimate measurement error, evaluate dimensionality, examine local independence, and investigate Differential Item Functioning (DIF) to ensure fairness across different populations (Bond & Fox, 2015).

With the rapid advancement of statistical software and computational technologies, researchers are now able to perform sophisticated psychometric analyses efficiently, including Item Response Theory (IRT) modeling, parameter estimation, simulation studies, assessment of model assumptions, and computerized adaptive testing. These developments have significantly improved the quality and accuracy of psychological and educational research (Embretson & Reise, 2000).

The importance of psychometric analysis can be summarized as follows:

- Ensuring the accuracy and precision of psychological and educational measurement.
- Improving the quality of measurement instruments.
- Identifying poorly functioning or biased items.
- Estimating examinees' abilities objectively.
- Supporting the development of item banks and computerized adaptive tests.
- Providing valid evidence for educational, psychological, and clinical decision-making.

Accordingly, psychometric analysis has become an indispensable component of modern test development and one of the principal criteria used to evaluate the scientific quality of measurement instruments published in international journals and adopted by professional testing organizations (AERA et al., 2014).

2.2 Classical Test Theory (CTT)

Classical Test Theory (CTT) represents the oldest and most widely applied framework in psychological and educational measurement. It has served as the foundation for the majority of measurement instruments developed throughout the twentieth century. According to Classical Test Theory, an individual's observed test score consists of two components: the true score and random measurement error. This relationship is expressed by the following equation:

X=T+EX = T + EX=T+E

where:

- **X** represents the observed score,
- **T** represents the true score,
- **E** represents random measurement error (Crocker & Algina, 2008).

The theory assumes that measurement error is random, has a mean of zero, and is uncorrelated with both true scores and other measurement errors. Based on these assumptions, numerous statistical procedures have been developed to evaluate measurement quality, including reliability coefficients, item difficulty indices, discrimination indices, internal consistency coefficients, and item-total correlations (DeVellis & Thorpe, 2021).

Despite its widespread application, Classical Test Theory has several methodological limitations. Most importantly, item statistics depend on the sample from which they are estimated, and estimates of individual ability depend on the difficulty level of the particular test administered. Consequently, the generalizability of measurement results across different samples and test forms is limited (Embretson & Reise, 2000).

Nevertheless, Classical Test Theory continues to play a significant role in applied psychological and educational research because of its conceptual simplicity, computational efficiency, and compatibility with statistical software such as SPSS. It remains particularly useful in descriptive studies, experimental research, and preliminary stages of test development.

2.3 Item Response Theory (IRT)

Item Response Theory (IRT) was developed to overcome many of the limitations associated with Classical Test Theory by providing probabilistic models that describe the relationship between an individual's latent ability and the probability of responding correctly to each individual item rather than relying solely on total test scores (Embretson & Reise, 2000).

IRT is based on several fundamental assumptions:

1. **Unidimensionality**, meaning that the test measures a single latent trait.
2. **Local Independence**, whereby responses to individual items are statistically independent after controlling for the latent ability.
3. **Monotonicity**, indicating that the probability of a correct response increases as the individual's ability increases.

Compared with Classical Test Theory, Item Response Theory offers several important advantages:

- Estimates of person ability are relatively independent of the particular sample of test items.
- Estimates of item characteristics are relatively independent of the sample of respondents, provided that model assumptions are satisfied.
- It provides the theoretical foundation for Computerized Adaptive Testing (CAT).

- Measurement precision varies according to ability level through conditional standard errors rather than a single reliability coefficient.

Several IRT models have been developed, the most common of which include:

- The **One-Parameter Logistic Model (1PL)**, commonly known as the **Rasch Model**.
- The **Two-Parameter Logistic Model (2PL)**.
- The **Three-Parameter Logistic Model (3PL)**, which incorporates a guessing parameter.

These models have become the cornerstone of modern psychological and educational measurement and are extensively used in large-scale international assessments, certification examinations, and standardized educational testing because of their superior measurement precision and theoretical robustness (Hambleton, Swaminathan, & Rogers, 1991).

2.4 Rasch Model

The Rasch Model is considered one of the most influential psychometric models developed in modern measurement theory. It was introduced by the Danish mathematician and statistician Georg Rasch with the aim of establishing an objective measurement framework capable of producing invariant estimates of person ability and item difficulty. The Rasch Model belongs to the family of One-Parameter Logistic (1PL) Item Response Theory models; however, it differs from other IRT models in its measurement philosophy, emphasizing fundamental measurement and the transformation of ordinal raw scores into interval-level measures (Bond & Fox, 2015).

The model assumes that the probability of a correct response depends solely on the difference between the person's ability and the item's difficulty, without introducing additional parameters such as discrimination or guessing. This relationship is represented by the logistic equation:

Rasch model equation

$$\mathbf{P(X_{ni} = 1) = e^{(\theta_n - \beta_i)} / (1 + e^{(\theta_n - \beta_i)})}$$

Where θ_n represents person ability, β_i represents item difficulty, and P is the probability of a correct response.

One of the most important characteristics of the Rasch Model is specific objectivity, meaning that comparisons among persons do not depend on the particular set of items used, and comparisons among items do not depend on the particular sample of respondents, provided that the model assumptions are adequately met (Rasch, 1960). This property has made Rasch analysis highly valuable in educational assessment, psychological testing, item banking, and computerized adaptive testing.

The Rasch Model offers several important advantages:

- Transformation of ordinal scores into interval-level measures.
- Relatively invariant estimates of person ability and item difficulty.
- Detailed evaluation of item and person fit using Infit and Outfit statistics.
- Detection of Differential Item Functioning (DIF).
- Construction of Wright Maps that place persons and items on the same measurement scale.
- Separate estimation of standard errors for persons and items.

Because of these features, the Rasch Model has become one of the principal methodologies used in modern test development and validation (Bond & Fox, 2015).

2.5 Statistical Software in Psychometric Analysis

The rapid development of computing technologies has led to the emergence of specialized software packages for psychometric analysis. Some programs were designed as general statistical environments, whereas others were developed specifically for modern measurement models such as Rasch and IRT. The selection of an appropriate software package depends on research objectives, measurement models, data characteristics, sample size, and the researcher's technical expertise (DeMars, 2018).

Among the most widely used software packages in psychometrics are:

- SPSS for applications based on Classical Test Theory.
- Winsteps for Rasch measurement.
- Python for modern psychometrics, data science, and artificial intelligence applications.

Although these tools may overlap in some functions, each one reflects a distinct analytical philosophy and serves different purposes within the process of test development and validation.

3. SPSS in Psychometric Analysis

3.1 Overview of SPSS

SPSS (Statistical Package for the Social Sciences) is one of the most widely used statistical software packages in psychology, education, and the social sciences. Originally developed in the late 1960s to facilitate statistical analysis without requiring extensive programming knowledge, SPSS later expanded substantially after being acquired by IBM (IBM Corp., 2023).

The software is characterized by a user-friendly graphical interface and a broad range of statistical procedures, making it particularly popular among researchers and graduate students working within the framework of Classical Test Theory.

3.2 Psychometric Applications of SPSS

SPSS provides numerous procedures commonly used in psychometric research, including:

- Descriptive statistics.
- Normality testing.
- Item analysis.
- Item difficulty and discrimination indices.
- Item-total correlations.
- Cronbach's alpha.
- Split-half reliability.
- Test-retest reliability.
- Exploratory factor analysis.
- Confirmatory factor analysis (through extensions or complementary software).
- Regression analysis.
- Analysis of variance.

For this reason, SPSS remains the dominant software package in many psychological and educational studies, particularly in master's theses and doctoral dissertations.

3.3 Advantages of SPSS

The literature identifies several major advantages of SPSS:

- Ease of use.
- Fast execution of statistical analyses.
- Wide adoption in universities.
- High-quality statistical output.
- Compatibility with multiple data formats.
- Extensive collection of statistical procedures.

SPSS is especially useful for studies based on Classical Test Theory, including reliability estimation, validity assessment, and item analysis.

3.4 Limitations of SPSS

Despite its strengths, SPSS has several limitations in modern psychometrics:

- It does not directly support Rasch modeling.
- Its support for Item Response Theory is limited.
- Automation capabilities are weaker than those of programming-based environments.
- Integration with artificial intelligence applications is relatively limited.

Consequently, researchers often complement SPSS with specialized software such as Winsteps or with programming environments such as Python when conducting advanced psychometric analyses.

What has been translated so far

Completed

- Title
- Introduction
- Psychometric Analysis
- Classical Test Theory (CTT)
- Item Response Theory (IRT)
- Rasch Model
- Statistical Software in Psychometrics
- SPSS (Overview, Applications, Advantages, Limitations)

4. Winsteps in Psychometric Analysis

4.1 Overview of Winsteps

Winsteps is one of the most specialized software programs designed for implementing the **Rasch Model**. It was developed by **John Michael Linacre** to provide a comprehensive analytical environment for Rasch measurement across psychological, educational, health, and social science research. Owing to its specialized analytical capabilities, Winsteps has become one of the most widely used software packages for developing, validating, and standardizing psychological and educational measurement instruments (Linacre, 2024).

The software supports various types of assessment data, including dichotomous items (e.g., correct/incorrect responses), Likert-type rating scales, partial credit items, and rating scale models. This versatility enables researchers to analyze a wide variety of psychological and educational instruments using a unified Rasch measurement framework.

4.2 Theoretical Foundation of Winsteps

Winsteps is built upon the philosophy of **objective measurement** proposed by Georg Rasch. Unlike traditional statistical software, which primarily focuses on describing data through statistical indices, Winsteps evaluates the extent to which empirical observations conform to the Rasch measurement model. Consequently, the software emphasizes measurement quality rather than merely statistical significance (Bond & Fox, 2015).

The program estimates item difficulty and person ability simultaneously using advanced estimation procedures, including **Joint Maximum Likelihood Estimation (JMLE)** and related estimation techniques. These procedures produce interval-level measures expressed in logits, accompanied by individual standard errors for both persons and items.

4.3 Psychometric Applications of Winsteps

Winsteps provides a comprehensive set of analytical procedures specifically designed for Rasch measurement, including:

- Estimation of item difficulty.
- Estimation of person ability.
- Standard error estimation.
- Infit Mean Square statistics.
- Outfit Mean Square statistics.
- Differential Item Functioning (DIF) analysis.
- Assessment of unidimensionality.
- Principal Components Analysis of Residuals (PCAR).
- Wright Maps.
- Rating scale diagnostics.
- Category functioning analysis.
- Test targeting analysis.
- Person and item separation indices.
- Reliability estimates for persons and items.

These procedures enable researchers to evaluate the psychometric quality of measurement instruments according to internationally recognized standards for educational and psychological testing (AERA et al., 2014).

4.4 Fit Statistics

One of the most important analytical outputs generated by Winsteps is **fit statistics**, which evaluate the consistency between observed responses and the expectations of the Rasch Model.

The two principal fit indices are:

- **Infit Mean Square (Infit MNSQ)**, which is sensitive to unexpected responses occurring near a person's estimated ability level.
- **Outfit Mean Square (Outfit MNSQ)**, which is more sensitive to unexpected responses that are far from a person's estimated ability.

According to Linacre (2002), values approximately ranging from **0.50 to 1.50** are generally considered productive for measurement, whereas values outside this interval may indicate item misfit or unexpected response behavior requiring further investigation.

Fit statistics assist researchers in identifying problematic items, unusual response patterns, and potential threats to measurement validity before finalizing a test or scale.

4.5 Differential Item Functioning (DIF)

Winsteps includes powerful procedures for **Differential Item Functioning (DIF)** analysis, which evaluates whether individual items function differently across distinct groups after controlling for ability.

An item is considered to exhibit DIF when individuals possessing the same level of the underlying trait but belonging to different groups (e.g., males versus females, or different cultural groups) have different probabilities of endorsing or answering that item correctly (Zumbo, 2007).

DIF analysis plays a fundamental role in ensuring fairness, reducing measurement bias, and improving the validity of psychological and educational assessments.

4.6 Wright Maps

One of the distinctive features of Winsteps is the generation of **Wright Maps**, also known as **Person–Item Maps**.

A Wright Map displays person abilities and item difficulties simultaneously on a common logit scale, enabling researchers to evaluate:

- Whether item difficulty appropriately matches respondents' ability levels.
- Gaps in item coverage across the latent trait continuum.
- Redundant or overly difficult/easy items.
- Overall targeting of the measurement instrument.

Wright Maps have become one of the most informative graphical outputs in Rasch measurement because they provide a visual representation of measurement quality that is not available in conventional statistical software (Bond & Fox, 2015).

4.7 Advantages of Winsteps

The literature highlights several major strengths of Winsteps:

- Specialized software dedicated to Rasch measurement.
- High precision in estimating item difficulty and person ability.
- Comprehensive fit statistics.
- Differential Item Functioning (DIF) analysis.
- Wright Maps and advanced graphical outputs.
- Category functioning diagnostics.
- Assessment of dimensionality.
- Compliance with international psychometric standards.
- Detailed analytical reports suitable for scientific publications.

For these reasons, Winsteps has become one of the preferred software packages for researchers involved in psychological measurement, educational assessment, health outcomes research, and certification testing.

4.8 Limitations of Winsteps

Despite its considerable strengths, Winsteps has several limitations:

- It requires a solid understanding of Rasch measurement theory.
- Its user interface is less intuitive than that of SPSS.
- It provides relatively limited support for general statistical analyses.
- Full functionality requires a commercial license.
- It is designed primarily for Rasch analysis rather than broader statistical modeling.

Consequently, Winsteps is frequently used in combination with software such as SPSS or Python to achieve comprehensive statistical and psychometric analyses.

5. Python in Psychometric Analysis

5.1 Overview of Python

Python is one of the most widely used programming languages worldwide and has become a fundamental tool in data analysis, artificial intelligence, machine learning, statistical computing, and psychometric analysis over the past decade. Its popularity stems from its open-source nature, ease of learning, extensive scientific libraries, and active global community that continuously develops specialized packages for a wide range of research fields (Van Rossum & Drake, 2009).

Unlike traditional statistical software such as **SPSS** and **Winsteps**, Python is not a standalone statistical application but rather a comprehensive programming environment that enables researchers to design customized analytical procedures according to the specific requirements of their data. This flexibility provides virtually unlimited opportunities for implementing both traditional and advanced psychometric analyses.

The rapid expansion of **Data Science** has transformed Python into one of the primary analytical platforms in psychology, education, and psychometrics. Through continuously evolving scientific libraries, researchers can perform statistical analysis, Item Response Theory modeling, Bayesian estimation, machine learning, artificial intelligence, simulation studies, and interactive data visualization within a unified computational environment (McKinney, 2022).

5.2 Scientific Architecture of Python for Psychometric Analysis

One of Python's greatest strengths lies in its ecosystem of specialized scientific libraries. Rather than relying on a single software package, Python integrates multiple libraries, each designed for a specific analytical purpose.

Among the most important libraries used in psychometric research are:

Library	Primary Function
NumPy	Numerical computation and matrix operations
Pandas	Data management and preprocessing
SciPy	Statistical analysis and scientific computing
Statsmodels	Advanced statistical modeling
Scikit-learn	Machine learning algorithms
Matplotlib	Statistical graphics and visualization
Plotly	Interactive visualization
PyMC	Bayesian statistical modeling
py-irt	Item Response Theory analysis
catsim	Computerized Adaptive Testing (CAT) simulation

The integration of these libraries enables researchers to conduct the complete psychometric workflow—from data cleaning and descriptive analysis to advanced measurement modeling, simulation, and automated reporting.

5.3 Applications of Python in Classical Test Theory (CTT)

Python supports virtually all statistical procedures commonly required within the framework of **Classical Test Theory**, including:

- Descriptive statistics.
- Item analysis.
- Item difficulty indices.
- Item discrimination indices.
- Item-total correlations.
- Cronbach's alpha.
- McDonald's Omega reliability coefficient.
- Split-half reliability.
- Exploratory Factor Analysis (EFA).
- Confirmatory Factor Analysis (CFA).
- Correlation analysis.

- Regression analysis.
- Structural Equation Modeling (SEM) through specialized libraries.

Unlike conventional statistical software, Python allows these analyses to be fully automated using reusable scripts, thereby reducing repetitive manual work and improving reproducibility.

5.4 Applications of Python in Item Response Theory (IRT)

Recent years have witnessed rapid growth in Python libraries supporting **Item Response Theory**. Researchers can estimate:

- Rasch Model (1PL).
- Two-Parameter Logistic Model (2PL).
- Three-Parameter Logistic Model (3PL).
- Graded Response Model (GRM).
- Partial Credit Model (PCM).
- Multidimensional Item Response Theory (MIRT).

Python also supports estimation of:

- Item Characteristic Curves (ICCs).
- Test Characteristic Curves (TCCs).
- Item Information Functions (IIFs).
- Test Information Functions (TIFs).
- Conditional Standard Errors of Measurement.
- Latent ability estimation using Maximum Likelihood and Bayesian approaches.

These capabilities provide researchers with considerable flexibility for implementing both standard and customized psychometric models.

5.5 Python and Rasch Measurement

Although **Winsteps** remains one of the most specialized software packages for Rasch measurement, Python has experienced substantial progress through open-source implementations of Rasch models. Researchers can perform Rasch analyses using specialized libraries while simultaneously integrating additional statistical procedures such as:

- Bayesian estimation.
- Simulation studies.
- Bootstrap procedures.
- Monte Carlo methods.
- Automated reporting.
- Interactive visualization.

An important advantage of Python is that researchers can modify or extend existing algorithms to develop new analytical methods tailored to specific research objectives—an option generally unavailable in proprietary statistical software.

5.6 Python and Artificial Intelligence

The integration of **Python** with artificial intelligence represents one of the most significant developments in contemporary psychometrics. Through its extensive ecosystem of machine learning and deep learning libraries, Python enables researchers to develop intelligent analytical models that extend far beyond conventional statistical procedures. This integration has expanded the scope of psychometric analysis by enabling the application of predictive analytics, intelligent automation, and advanced computational techniques within psychological and educational measurement.

Specifically, Python can be used to:

- Develop predictive models for academic achievement.
- Detect hidden patterns in psychological and educational data.
- Build Computerized Adaptive Testing (CAT) systems.
- Analyze textual responses using Natural Language Processing (NLP).
- Detect Differential Item Functioning (DIF) using machine learning algorithms.
- Automate psychometric reporting.
- Develop interactive dashboards for real-time assessment.

Collectively, these capabilities have transformed psychometric analysis from a traditional statistical process into an intelligent decision-support system capable of processing large-scale datasets with high levels of accuracy, efficiency, and automation.

5.7 Advantages of Python

The integration of statistical computing, psychometric modeling, and artificial intelligence has made Python one of the most versatile analytical environments available to researchers. The literature identifies numerous advantages that contribute to its growing adoption in psychometric research.

Among its principal advantages are:

- Free and open-source software.
- High flexibility in statistical programming.
- Comprehensive support for Classical Test Theory (CTT).
- Extensive support for Item Response Theory (IRT).
- Integration with artificial intelligence and machine learning.
- Ability to process very large datasets.
- Automation of repetitive analytical procedures.
- Production of high-quality statistical graphics.
- Strong reproducibility through reusable programming scripts.
- Continuous development by a large international scientific community.

Taken together, these characteristics position Python as one of the most powerful and flexible analytical environments currently available for psychological and educational measurement.

5.8 Limitations of Python

Despite its considerable strengths, Python also presents several challenges that researchers should carefully consider before adopting it for psychometric applications.

Its principal limitations include:

- It requires programming knowledge.
- The learning curve is steeper than that of SPSS.
- Some psychometric libraries are still under active development.
- Different libraries may produce varying implementations, requiring careful validation.
- Researchers must verify the accuracy and appropriateness of analytical scripts before applying them to scientific research.

Nevertheless, these limitations continue to diminish as the Python ecosystem expands, documentation improves, and psychometric libraries become increasingly stable and mature.

5.9 The Future of Python in Psychometric Analysis

Recent developments indicate that Python is rapidly becoming one of the leading computational environments for modern psychometric research. The rapid growth of artificial intelligence, big data analytics, computerized adaptive testing, educational technology, and digital assessment has substantially increased the demand for flexible programming environments capable of integrating statistical analysis with intelligent computational models.

Moreover, the continuous development of open-source scientific libraries and advances in computational power are expected to further strengthen Python's role in psychological and educational measurement.

Future psychometric research is therefore expected to rely increasingly on Python because of its ability to combine statistical analysis, psychometric modeling, machine learning, Bayesian estimation, simulation, visualization, and automated reporting within a unified programming framework.

Consequently, Python is likely to play a central role in the next generation of psychological and educational measurement systems, supporting the development of more accurate, adaptive, and intelligent assessment methodologies.

6.9 Summary Comparison Table

Comprehensive Comparison of SPSS, Winsteps, and Python in Psychometric Analysis

Comparison Criterion	SPSS	Winsteps	Python
Ease of use	Excellent	Good	Moderate
Graphical user interface	Excellent	Good	Limited (programming-based)
Open source	No	No	Yes
Classical Test Theory (CTT)	Excellent	Limited	Excellent
Rasch Model	Limited	Excellent	Excellent
Item Response Theory (IRT)	Limited	Good	Excellent
Item analysis	Excellent	Excellent	Excellent
Cronbach's Alpha	Yes	No	Yes
Exploratory factor analysis	Yes	No	Yes
Confirmatory factor analysis	Limited	No	Yes
Differential Item Functioning (DIF)	No	Yes	Yes
Infit and Outfit statistics	No	Yes	Yes
Wright Maps	No	Yes	Yes
Psychometric simulation	Limited	Limited	Excellent
Big data analysis	Limited	Limited	Excellent
Artificial intelligence	No	No	Excellent
Machine learning	No	No	Excellent
Statistical graphics	Good	Good	Excellent
Computational speed	Excellent	Excellent	Excellent
Workflow automation	Limited	Limited	Excellent
Development of customized algorithms	No	No	Yes

6.10 Discussion of the Comparative Findings

The comparative analysis demonstrates that **SPSS**, **Winsteps**, and **Python** should not be regarded as competing software packages but rather as complementary analytical environments that serve different stages of psychometric research.

SPSS remains the preferred software for conventional statistical analyses based on **Classical Test Theory**, particularly during the early stages of test construction, including descriptive statistics, item analysis, reliability estimation, and validity assessment. Its graphical user interface and ease of use continue to make it one of the most popular statistical packages among researchers in psychology, education, and the social sciences.

Winsteps occupies a specialized position in modern psychometrics because of its exclusive focus on the **Rasch Model**. Its advanced diagnostic procedures—including fit statistics, Wright Maps, Differential Item Functioning analysis, and separation indices—provide a level of psychometric evaluation that cannot be achieved using conventional statistical software alone.

Python represents the most flexible analytical environment among the three. Beyond supporting both Classical Test Theory and Item Response Theory, it integrates psychometric modeling with data science, artificial intelligence, machine learning, Bayesian statistics, simulation, and automated reporting. Consequently, Python offers unprecedented opportunities for developing innovative psychometric methodologies and intelligent assessment systems.

These findings suggest that software selection should be guided primarily by the objectives of the research rather than by software popularity or user familiarity. Researchers conducting traditional statistical analyses may find SPSS sufficient, whereas those developing measurement instruments using Rasch methodology will benefit substantially from Winsteps. For advanced psychometric research involving artificial intelligence, large-scale data analysis, or customized computational modeling, Python provides the most comprehensive solution.

An integrated analytical workflow may therefore represent the most effective strategy. Researchers can begin by preparing and exploring data in SPSS, proceed to Rasch calibration and validation using Winsteps, and subsequently employ Python for advanced statistical modeling, predictive analytics, visualization, automation, and intelligent decision-support applications.

7. Discussion

7.1 The Evolution of Psychometric Software in Light of Modern Measurement

The findings of the comparative analysis indicate that the evolution of psychometric software has paralleled the development of psychological and educational measurement theories. The widespread adoption of **SPSS** reflects the long-standing dominance of **Classical Test Theory (CTT)**, which has traditionally focused on estimating reliability coefficients, validity evidence, and item statistics based on sample-dependent indices (Crocker & Algina, 2008). Owing to its ease of use and comprehensive statistical procedures, SPSS has played a significant role in promoting statistical literacy among researchers in psychology, education, and the social sciences.

However, the methodological limitations of Classical Test Theory, particularly the dependence of item and person statistics on specific samples, encouraged the development and adoption of **Item Response Theory (IRT)**. Unlike Classical Test Theory, IRT provides probabilistic models that estimate person ability and item characteristics independently under specified model assumptions (Embretson & Reise, 2000).

Within this context, **Winsteps** emerged as one of the most influential software packages dedicated to **Rasch measurement**. Its analytical procedures allow researchers to evaluate item fit, estimate person ability and item difficulty, detect Differential Item Functioning (DIF), assess dimensionality, and produce Wright Maps, thereby supporting objective measurement and high-quality instrument development (Bond & Fox, 2015).

More recently, the rapid expansion of **data science**, **artificial intelligence**, and **machine learning** has fundamentally transformed psychometric analysis. This transformation has positioned **Python** as an integrated scientific environment capable of combining statistical analysis, psychometric modeling, predictive analytics, simulation, and intelligent computational techniques within a single programming framework (de Ayala, 2022).

7.2 Complementarity Rather Than Competition

The present review demonstrates that SPSS, Winsteps, and Python should not be viewed as competing alternatives but rather as complementary analytical environments, each serving distinct purposes within the psychometric workflow.

SPSS is particularly appropriate for conventional statistical analyses and the preliminary stages of instrument development, including data preparation, descriptive statistics, reliability estimation, and exploratory analyses.

Winsteps is specifically designed for Rasch measurement and provides highly specialized diagnostic procedures that are essential for developing valid and objective measurement instruments.

Python extends beyond both SPSS and Winsteps by integrating psychometric analysis with advanced computational methods such as machine learning, Bayesian statistics, simulation, and artificial intelligence. Consequently, Python provides an exceptionally flexible platform for modern psychometric research.

An integrated analytical strategy combining the strengths of all three environments may therefore yield more comprehensive and accurate results than relying on a single software package.

7.3 Python and the Future of Psychometrics

One of the most significant findings of this review is the growing importance of Python in the future of psychological and educational measurement.

Unlike conventional statistical software, Python enables researchers to develop customized algorithms, automate repetitive analyses, implement adaptive testing systems, analyze unstructured data, and integrate psychometric models with artificial intelligence applications.

Recent developments in educational technology and digital assessment suggest that future psychometric systems will increasingly depend on computational environments capable of processing large datasets, performing predictive analytics, and supporting intelligent decision-making.

Consequently, Python is expected to become one of the principal computational platforms for next-generation psychometric research.

7.4 Implications for Researchers in the Arab World

The findings of this review have particular significance for researchers in Arab countries, where SPSS remains the dominant statistical software in psychological and educational research.

Although SPSS provides valuable statistical procedures, exclusive reliance on Classical Test Theory may limit the adoption of contemporary psychometric methodologies based on Rasch measurement and Item Response Theory.

Accordingly, graduate programs in psychology, education, and educational measurement should incorporate specialized training in Rasch analysis, Item Response Theory, and Python programming in order to prepare researchers for emerging developments in quantitative measurement.

Expanding researchers' methodological and computational competencies will contribute to improving the quality of psychological and educational research and increasing its international competitiveness.

7.5 Research Gaps

Despite the growing body of literature addressing psychometric software, several important research gaps remain.

First, relatively few comparative studies have examined **SPSS**, **Winsteps**, and **Python** within a single comprehensive review.

Second, empirical investigations evaluating these three analytical environments using identical datasets remain limited.

Third, there is a shortage of studies exploring the integration of psychometric analysis with artificial intelligence and machine learning in psychological measurement.

Finally, additional research is required to evaluate the effectiveness of integrated analytical workflows combining traditional statistical software, Rasch measurement software, and programming-based computational environments.

7.6 Contribution of the Present Study

The present review contributes to the psychometric literature by providing a comprehensive comparative evaluation of three of the most widely used analytical environments in psychological and educational measurement.

Unlike previous studies that focused primarily on individual software packages, this review synthesizes theoretical foundations, methodological characteristics, analytical capabilities, practical applications, strengths, limitations, and future prospects within a unified comparative framework.

Furthermore, the study highlights the increasing integration of psychometrics with data science and artificial intelligence, thereby providing researchers with practical guidance for selecting analytical tools that best match their research objectives and methodological requirements.

8. Conclusion

This study aimed to provide a comprehensive comparative review of three of the most widely used software packages in psychometric analysis—**SPSS**, **Winsteps**, and **Python**—by examining their theoretical foundations, statistical capabilities, and applications in the development, validation, and standardization of psychological and educational measurement instruments.

The review demonstrated that each software package possesses distinctive strengths that make it suitable for specific stages of psychometric research. **SPSS** remains the preferred environment for conventional statistical analyses and applications based on **Classical Test Theory (CTT)**. Its comprehensive statistical procedures and user-friendly interface make it particularly valuable for descriptive analyses, reliability estimation, validity assessment, factor analysis, and preliminary stages of test construction.

In contrast, **Winsteps** has established itself as one of the leading software packages for **Rasch measurement**, providing highly specialized procedures for evaluating item fit, estimating person ability and item difficulty, examining Differential Item Functioning (DIF), assessing dimensionality, and generating Wright Maps. These capabilities contribute substantially to the development of objective and scientifically rigorous psychological and educational measurement instruments.

Among the three analytical environments, **Python** offers the greatest degree of flexibility and computational power. Beyond supporting both Classical Test Theory and Item Response Theory, Python integrates psychometric analysis with **data science**, **machine learning**, **artificial intelligence**, **Bayesian statistics**, and **big data analytics**, enabling researchers to develop customized analytical workflows and innovative measurement systems.

The findings further indicate that these software packages should be regarded as **complementary rather than competing tools**. The selection of an appropriate analytical environment should therefore be determined by the objectives of the study, the psychometric model employed, the characteristics of the data, and the researcher's methodological expertise.
Accordingly, an integrated analytical approach combining the strengths of **SPSS**, **Winsteps**, and **Python** may provide more comprehensive, accurate, and scientifically robust psychometric analyses than reliance on any single software package.
Finally, the review highlights that the future of psychometric analysis is increasingly moving toward computational environments that integrate modern measurement theory with artificial intelligence, machine learning, and advanced data science techniques. This transformation underscores the need for researchers and academic institutions to continuously develop their methodological and computational competencies in order to meet the evolving demands of contemporary psychological and educational measurement.

9. Recommendations

Based on the findings of this review, the following recommendations are proposed:

1. Researchers and graduate students should receive comprehensive training in modern psychometric software, including **Winsteps** and **Python**, in addition to conventional statistical software such as **SPSS**.
2. Graduate curricula in psychology, educational measurement, and educational research should incorporate specialized courses on **Item Response Theory**, **Rasch Measurement**, and **Python programming** for psychometric analysis.
3. Researchers are encouraged to adopt integrated analytical workflows in which **SPSS** is used for preliminary statistical analyses, **Winsteps** for Rasch calibration and validation, and **Python** for advanced computational modeling, automation, and artificial intelligence applications.
4. Greater attention should be devoted to developing psychological and educational measurement instruments using modern psychometric models rather than relying exclusively on Classical Test Theory.
5. Python should be increasingly employed in the development of **computerized adaptive testing**, intelligent assessment systems, automated scoring procedures, and predictive psychometric models.
6. Future empirical studies should compare the analytical performance of SPSS, Winsteps, and Python using identical datasets to evaluate similarities and differences in psychometric estimates.
7. Interdisciplinary collaboration among specialists in psychology, educational measurement, statistics, computer science, and artificial intelligence should be encouraged to develop next-generation psychometric methodologies.
8. Open-source psychometric libraries should continue to be developed and expanded in order to facilitate access to advanced analytical techniques for researchers worldwide.
9. Periodic review studies should be conducted to monitor ongoing developments in psychometric software and emerging computational technologies.
10. Universities and research institutions should update their psychometric training programs to align with international developments in quantitative measurement, data science, and artificial intelligence.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). Guilford Press.

DeMars, C. (2018). *Item response theory*. Oxford University Press.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5th ed.). Sage.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

IBM Corp. (2023). *IBM SPSS Statistics algorithms*. IBM Corporation.

Linacre, J. M. (2002). What do Infit and Outfit Mean-Square and Standardized Mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2024). *Winsteps® Rasch measurement computer program user's guide*. Winsteps.com.

McKinney, W. (2022). *Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233.