Pedagogical Research
2025, 10(2), em0239
e-ISSN: 2468-4929
https://www.pedagogicalresearch.com

MODESTUM

Research Article

OPEN ACCESS

Paradigm shift in reliability estimates: Application of analysis of variance repeated measures (ANOVAM) in validation studies

Usani Joseph Ofem [1,2*], Valentine Joseph Owan [3], Cletus Ibout [3], Sylvai Victor Ovat [3]

[1] Alex Ekwueme Federal University, Ndufu-Alike, NIGERIA
[2] De Oracle Research Network, NIGERIA
[3] University of Calabar, NIGERIA
***Corresponding Author:** ofemoracle@gmail.com

Citation: Ofem, U. J., Owan, V. J., Ibout, C., & Ovat, S. V. (2025). Paradigm shift in reliability estimates: Application of analysis of variance repeated measures (ANOVAM) in validation studies. *Pedagogical Research, 10*(2), em0239. https://doi.org/10.29333/pr/16402

ARTICLE INFO

Received: 25 Jun 2024

Accepted: 10 Mar 2025

ABSTRACT

This study employed repeated measures ANOVA to assess the reliability of an instrument designed to measure utilization, awareness, and perception of AI in research among 150 undergraduate students. Validated instruments with robust psychometric properties were used for the study. Data collection occurred in three phases spaced two weeks apart, following experts recommendations for longitudinal research. Initial findings using Cronbach's alpha indicated high reliability in the first phase. However, subsequent test-retest analyses revealed decreasing reliability coefficients below acceptable thresholds for utilization, awareness, and perception constructs. Further analysis using repeated measures ANOVA showed significant differences in mean scores across the three phases, suggesting inconsistency in respondents' perceptions over time. The study underscores the dynamic nature of attitudes towards AI, necessitating careful consideration in longitudinal research designs. Methodologically, it highlights the limitations of relying solely on static reliability estimates such as Cronbach's alpha. Practically, the findings suggest the need for continuous refinement of measurement instruments to capture evolving attitudes accurately. Theoretical contributions include advancing understanding of reliability in dynamic contexts, prompting future research to explore more robust statistical methods and measurement approaches in studying attitudes towards emerging technologies.

Keywords: ANOVA repeated measures, artificial intelligence, utilization of AI, Cronbach alpha, test-retest

INTRODUCTION

Reliability stands as a fundamental pillar of empirical research, ensuring the consistency and dependability of research outcomes. It measures the degree to which an instrument produces consistent results under stable conditions. Without reliability, the validity and trustworthiness of any empirical study are jeopardized, as unreliable measurements can lead to erroneous conclusions. Trochim (2006) defines reliability as the consistency of a measurement tool or research method over time, encompassing forms like test-retest reliability, inter-rater reliability, and internal consistency. In fields such as psychology, reliability is crucial for ensuring consistent measurement of constructions. For instance, in clinical settings, reliable diagnostic tools are vital for accurately identifying mental health conditions.

However, challenges arise when measures lack consistency, potentially leading to misdiagnosis or inappropriate treatments (Koo & Li, 2016). Likewise, high inter-rater reliability is crucial in subjective studies like behavioral observations (Hallgren, 2012). While traditional measures such as Cronbach's alpha have been foundational in assessing reliability (Brown & Hudson, 2020; Smith, 2019), their applicability and limitations are increasingly scrutinized in contemporary research. For instance, Cronbach's alpha's sensitivity to item homogeneity poses challenges; if items do not measure the same construction, alpha values may be lower (Tavakol & Dennick, 2011). Moreover, the length and relevance of test items influence Cronbach's alpha, with longer tests potentially inflating reliability scores if additional items are not conceptually aligned (Cortina, 1993). Despite its utility in assessing scale homogeneity, Cronbach's alpha's sensitivity to scale length, sample size, and item variance is debated (Sijtsma, 2009).

Similarly, in test-retest reliability, the time interval between test administrations plays a crucial role. A short interval may lead respondents to recall their previous responses, artificially inflating correlations and overestimating reliability. Conversely, a long interval may allow for changes in respondents' true scores due to maturation or external factors, thereby underestimating reliability (Anastasi & Urbina, 1997). Furthermore, the stability of the construction being measured also influences test-retest reliability. Constructs that are inherently stable, such as intelligence, tend to show higher reliability estimates over time compared to constructs like mood or stress levels, which are more prone to change (Nunnally & Bernstein, 1994). Similarly, split-half reliability, another commonly used technique, has its limitations. Randomly splitting a test can yield different reliability estimates compared to systematic splitting methods like odd-even item splits. This discrepancy arises because the items in each split might not equally represent the entire content of the test, leading

to inconsistent reliability estimates (Eisinga et al., 2013). The length of the test also affects split-half reliability (Nunnally & Bernstein, 1994). Additionally, factors such as the clarity and quality of test items, respondent motivation, fatigue, and familiarity with test format can impact reliability (DeVellis, 2016; Hogan, 2007). Moreover, testing conditions, the quality of instructions given to respondents, and inconsistent administration procedures can introduce variability that is not related to the construction being measured. Standardizing administration conditions helps mitigate this source of error (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Therefore, employing robust techniques to determine reliability is essential to account for these various factors and ensure the accuracy and consistency of measurement in research studies.

In recent years, there has been an increasing interest in applying Analysis of Variance (ANOVA) techniques, specifically repeated measures ANOVA, to evaluate reliability in empirical research (Field, 2013). Originally designed for comparing means across different groups, ANOVA has been adapted to assess changes within subjects over multiple time points or conditions (Keppel & Wickens, 2004). Repeated measures ANOVA offers several advantages over traditional reliability measures, particularly in its capability to simultaneously account for within-subject and between-subject variability (Hays, 2018). This statistical method examines changes in a dependent variable over time or in response to various experimental conditions within the same participants (Tabachnick & Fidell, 2019). By analyzing variance across multiple measurements from the same individuals, researchers can evaluate the stability and consistency of measurements under different circumstances (Morrison, 2018). This approach yields a more nuanced understanding of how measurements behave across diverse contexts, which is crucial for establishing the reliability of instruments used in longitudinal studies or experiments involving repeated assessments.

When considering the application of ANOVA repeated measures for reliability assessment, it is essential to address the challenges and potential justifications for this approach, particularly in relation to the time interval issue inherent in test-retest reliability. Traditional test-retest reliability faces the challenge of determining an appropriate interval between test administrations, which can influence the stability of the measured construct (Anastasi & Urbina, 1997). This raises the question of whether employing ANOVA repeated measures, which involves multiple administrations over time, is justified given this challenge. Firstly, it is crucial to delineate the contexts in which ANOVA repeated measures can offer advantages. ANOVA repeated measures is a robust statistical tool that facilitates the assessment of changes in a dependent variable across multiple time points (Field, 2013). This method is particularly beneficial for examining within-subject variability and comprehending how construction evolves over time. The repeated measures design adjusts for individual differences by employing the same participants across all time points, thereby enhancing the sensitivity of the analysis (Girden, 1992).

An important advantage of ANOVA repeated measures over traditional test-retest reliability is its capability to accommodate more than two times. Traditional test-retest reliability typically involves only two administrations, which may restrict the ability to capture the stability and consistency of the construct overtime (Nunnally & Bernstein, 1994). By incorporating three or more administrations, ANOVA repeated measures can provide a more comprehensive assessment of the construct's reliability. This approach enables the identification of patterns of stability and change that might not be apparent with only two time points.

However, the issue of time intervals remains pertinent. Like test-retest reliability, careful consideration of the time intervals between administrations in ANOVA repeated measures is essential. Short intervals may lead to memory effects, where participants recall their previous responses, while extended intervals may introduce changes in the construction being measured due to maturation or external factors (Cohen, 1960). Thus, selecting appropriate time intervals is critical to ensure that observed changes reflect the true stability of the construction rather than external influences.

Despite these obstacles, employing ANOVA repeated measures can be justified and advantageous in several ways. For example, by carefully planning the study with appropriate intervals between assessments, researchers can mitigate potential biases related to memory effects and developmental changes. Additionally, the repeated measures design enable examination of both within-subject and between-subject variability, yielding a more nuanced understanding of reliability (Howell, 2010). Furthermore, ANOVA repeated measures can incorporate covariates to control for confounding variables, thereby enhancing the robustness of the analysis. For instance, researchers can include covariates such as age, gender, or previous exposure to the construction being measured to adjust for their impact on reliability estimates (Tabachnick & Fidell, 2013). This capability to include covariates represents a significant advantage over traditional test-retest reliability, which typically does not address such factors. Moreover, ANOVA repeated measures can evaluate interaction effects between time and other variables, providing insights into how various factors influence the stability of the construction over time. This depth of analysis is not achievable with traditional test-retest reliability, making ANOVA repeated measures a more adaptable and informative approach for evaluating reliability (Maxwell & Delaney, 2004). Despite this potential, many studies continue to rely on conventional methods without exploring the potential benefits of ANOVA repeated measures across diverse research contexts (Winer et al., 1991). Addressing this gap is essential for enhancing the methodological rigor of empirical research and expanding researchers' toolkit for assessing reliability. Thus, this study seeks to provide empirical evidence supporting ANOVA repeated measures as a viable alternative for assessing the stability and consistency of measurements.

LITERATURE REVIEW: OVERVIEW OF RELIABILITY MEASURES

Reliability within research methodology pertains to the consistency, stability, and trustworthiness of measurement instruments and procedures, crucial for ensuring the validity of research findings (Carmines & Zeller, 1979). Grounded in generalizability theory and measurement theory, reliability frameworks provide the basis for evaluating the consistency of measurements (Shavelson & Webb, 1991). Generalizability theory posits that observed scores encompass both true scores and measurement error, underscoring the importance of minimizing error variance to enhance reliability (Cronbach et al., 1972). Measurement theory, meanwhile, focuses on the properties and characteristics of measurement scales and methods used to assess reliability (Nunnally & Bernstein, 1994).

Traditionally, Cronbach's alpha serves as a widely accepted measure of internal consistency reliability, particularly in the development and validation of scales (Cronbach, 1951). It computes the average correlation among items within a scale, indicating the degree to which items measure the same underlying construction (Smith, 2019). Higher alpha coefficients imply greater internal consistency, yet the measure is sensitive to factors such as scale length, sample size, and item homogeneity (Sijtsma, 2009). Despite its prevalence, Cronbach's alpha has faced criticism for potentially inflating reliability estimates and insufficiently addressing measurement error (Tavakol & Dennick, 2011). Cronbach's alpha is most suitable for evaluating internal consistency in scales featuring multiple items measuring a single construct (Cortina, 1993), but it may not be suitable for complex constructs or scales with heterogeneous item sets measuring different dimensions (Gliem & Gliem, 2003).

Similarly, test-retest reliability evaluates the consistency of measurements over time by administering the same test to the same group of participants on two separate occasions (Streiner & Norman, 2008). The correlation of scores between these administrations provides an indication of reliability, assuming the measured construction remains stable across tests (Field, 2013). However, test-retest reliability can be influenced by factors such as memory and practice effects, as well as changes in participants' conditions or circumstances between test sessions (Bland & Altman, 1996). While suitable for stable constructions, test-retest reliability may be less robust for variables prone to change over time (Bonett, 2002). Inter-rater reliability assesses the consistency of judgments made by different raters or observers (Hallgren, 2012). This type of reliability is commonly utilized in qualitative research, observational studies, and coding schemes where subjective assessments are involved (Koo & Li, 2016). Cohen's kappa coefficient and intraclass correlation coefficients (ICC) are frequently employed to quantify inter-rater agreement, with higher coefficients indicating greater reliability (McHugh, 2012). However, inter-rater reliability can be affected by rater biases, differences in interpretation criteria, and variations in rater expertise (Lombard et al., 2002). Establishing consensus among raters can be challenging in complex or ambiguous assessment tasks, underscoring the importance of rigorous methodological approaches (Cicchetti, 1994).

Conceptual Framework of ANOVA and Theoretical Underpinnings

ANOVA is founded on the principle of partitioning variance and employing F-statistics to evaluate differences among group means. Rooted in analysis of variance, ANOVA encompasses one-way and factorial designs tailored to diverse research inquiries and experimental setups (Keppel & Wickens, 2004). The theoretical basis of ANOVA draws from probability theory and the distribution of sample means. By comparing systematic variance (attributable to group differences) to unsystematic variance (within-group variability), ANOVA determines whether observed group differences stem from genuine effects or random sampling fluctuations (Maxwell & Delaney, 2004). This method not only offers a structured approach to hypothesis testing but also quantifies effect sizes using measures like eta-squared and partial eta-squared (Olejnik & Algina, 2003).

Repeated measures designed in ANOVA are particularly valuable, involving multiple assessments of the same subjects under varied conditions or time points. This design reduces variability between subjects, enhancing statistical power to detect smaller effects (Field, 2013). Repeated measures ANOVA extends traditional ANOVA by examining within-subject variability across levels of an independent variable, whether categorical or continuous (Greenhouse & Geisser, 1959). Widely applied in psychology, medicine, and behavioral sciences, repeated measures ANOVA is used to study changes over time, treatment effects, or interactions within individuals (Keppel & Wickens, 2004). For example, in clinical trials, it assesses treatment efficacy by comparing outcomes before and after intervention within the same patient group (Fitzmaurice et al., 2011).

ANOVA repeated measures represent an innovative approach to assessing reliability, involving multiple administrations of a measurement tool. This method aligns closely with the theoretical foundations of reliability, offering a deeper understanding of measurement consistency. Conducting three or more administrations allows researchers to capture a detailed view of reliability, revealing how consistently the tool performs across instances and conditions (Millsap & Maydeu-Olivares, 2009). This approach provides critical insights into the tool's stability over time, thereby enhancing the robustness of reliability assessments.

Empirical studies on reliability

Empirical research has extensively validated the utility of Cronbach's alpha across diverse fields. For instance, Tavakol and Dennick (2011) conducted a comprehensive analysis of Cronbach's alpha, highlighting its relevance in medical education. They emphasized that Cronbach's alpha values above 0.70 are generally acceptable, cautioning against sole reliance on this measure. Their study underscored the importance of considering the number of items and the construction's dimensionality. In a seminal study, Schmitt (1996) investigated test-retest reliability across various psychological measures over different time intervals, revealing that reliability coefficients typically decrease with longer intervals due to changes in respondents' psychological states or external influences.

Recent studies have explored advanced reliability assessment methods. Generalizability Theory (G-Theory), as studied by Brennan (2001), extends classical test theory by accounting for multiple sources of error variance. Brennan demonstrated G-Theory's applicability in educational assessments, aiding in the design of tests with optimal reliability by managing items, raters, and testing occasions. Item Response Theory (IRT), applied by Embretson and Reise (2000), models the relationship between latent traits and test item responses, offering precise reliability estimates across varying trait levels. Thissen and Wainer (2001) illustrated IRT's effectiveness in adaptive testing contexts, providing detailed insights into item characteristics and test reliability.

In contrast, limited research has explored ANOVA repeated measures as a reliability tool. Existing studies primarily utilize this method in clinical settings to assess experimental effects. Blanchard et al. (2010), for example, employed repeated measures ANOVA to evaluate the impact of a digital literacy program on students' skills, showing significant improvements over time. Other studies (Anderson & Williams, 2017; Brown & Green, 2020; Lee & Johnson, 2019; Thompson et al., 2018) similarly found ANOVA repeated measures effective for analyzing responses longitudinally. However, ANOVA repeated measures present challenges, including the assumption of sphericity and managing missing data. Violations of sphericity assumption can lead to erroneous conclusions, mitigated by techniques like Mauchly's test and Greenhouse-Geisser correction (Girden, 1992). Addressing missing data requires advanced methods such as mixed-effects

Table 1. Psychometric properties of adopted scale of Ofem et al. (2024)

Constructs	N	α	AVE	Discriminant values
Awareness	6	0.761	0.563	0.750
Perception	3	0.779	0.870	0.933
Utilization of AI in research	6	0.760	0.606	0.7788

models or multiple imputation to ensure robust findings (Field, 2013). The scarcity of studies exploring ANOVA repeated measures for reliability assessment underscores the motivation for this study.

METHODOLOGY

The study employed a descriptive survey design, aiming to collect information about the characteristics, behaviors, attitudes, or opinions of a specific population. This design primarily focuses on describing current situations or phenomena without manipulating the study environment (Creswell, 2014). The required population size for reliability studies varies among scholars. According to Nunnally (1978), a minimum sample size of 30 participants is often recommended to estimate the reliability coefficient. However, recent recommendations suggest larger samples for more accurate estimates. Some researchers advocate for sample sizes ranging from 100 to 200 participants to ensure robust and reliable results (Kline, 2011; Tabachnick & Fidell, 2013). Therefore, this study selected 150 respondents. The criteria for selecting respondents stipulated that they must be in their final year of study, assigned a supervisor, and actively engaged in project work. This criterion ensures that participants are actively involved in rigorous research activities where AI may prove beneficial to them.

Measures

The instrument used or data collection was a scale titled ‘‘Students’ Awareness, Perceptions, and Utilisation of Artificial Intelligence in Research Scale (SAPUAR) adopted from the study of Ofem et al. (2024). The scale was validated quantitative with strong psychometric evidence. The properties of the scale are presented in **Table 1**. The instrument was adopted completely, and no item was removed, added, or adjusted for the study. this is because of the cultural or contextual setting that has no difference with the samples that were used for the study (Brislin, 1970).

Procedure for Data Collection

Phase one

During Time 1, the initial administration of the instrument assessing students’ awareness, perception, and utilization of AI for academic research involved several critical steps to ensure a robust baseline measurement. The researchers ensured the instrument was properly printed, tested for readability, and face-validated for the target population. Following the best global practices, the study protocol was submitted and approved by the Institutional Review Board of Alex Ekwueme Federal University (IRB/FUNAI/024/8321). All participants were informed about the study’s purpose, assured of data confidentiality using encryption via a Firefox Password, and required to provide their consent on administered forms. Out of the 150 selected participants, 27 declined consents, and their decision did not influence the study. Data was collected from the remaining 123 respondents, each assigned unique identifiers for confidentiality. The collected data was carefully cleaned to eliminate missing values, outliers, or inconsistencies and securely stored in a designated folder.

Phase two: Second administration of the Instrument

For studies focused on students’ awareness, perception, and utilization of AI in academic research, the ideal interval between the first and second administrations is typically 2 weeks to 1 month. This interval balances avoiding significant external influences that could alter perceptions while allowing for potential changes in responses. Research supports this interval as adequate for assessing test-retest reliability without significant memory effects (Field, 2013). Participants were reminded about the second administration, and the same paper-based questionnaire format and relaxed environment were maintained to ensure consistency in administration conditions. Researchers reiterated the study’s objectives and stressed the importance of honest and thoughtful responses. Each participant’s responses were tracked using their unique identifiers assigned during Time 1 to maintain accuracy.

Phase three: Third administration for repeated measures ANOVA

For the third administration, a similar time interval of two weeks was maintained. This consistency helps in capturing changes or stability over a reasonably short period without introducing significant external influences (Field, 2013). Here, the researcher must financially incentivise the respondents so as not to fake their responses or just commit response bias in a bid to satisfy the researchers. The researchers extensively explained to them that this administration is the last time, and the responses will be vital for the achievement of the study. They were asked to be honest and objective as before in providing the responses. To reduce the level of excitement that will affect their responses, the money was given to them before and at the end of the final submission of the questionnaire. However, the same less tensed environment was replicated as closely as possible to maintain consistency and control for external variables. The unique codes or identifiers were also used to ensure that nobody is identified and that the responses from the first to the third are not mixed. The data were collected from all the participants and securely store stored in the same excel sheet. All responses were finally obtained and not one respondent was lost in the process. The data analysis was performed using SPSS 20.0 and Jamovi.

Table 2. Reliability estimates of the student's utilization, awareness, and perception of AI in research using Cronbach alpha

Measures	N	M	SD	α
UT1	100	2.58	.74	
UT2	100	2.53	.70	
UT3	100	2.58	.74	
UT4	100	2.54	.74	
UT5	100	2.55	75	
UT6	100	2.56	.75	
Utilization	100	15.43	4.36	.994
AWA1	100	3.32	.63	
AWA2	100	3.33	.63	
AWA3	100	3.32	.64	
AWA4	100	3.30	.59	
AWA5	100	3.31	.63	
AWA6	100	2.82	.83	
Awareness	100	19.37	3.67	.971
PER1	100	2.48	.64	
PER2	100	2.71	.60	
PER3	100	2.69	.61	
Perception	100	7.88	1.76	.942

M=mean, SD=Standard deviation, α=Cronbach alpha

RESULTS

The results were presented in three phases based on the data collection pattern. The initial phase aimed to assess the scale's reliability using a single administration method to determine its internal consistency. The second phase observed changes in reliability coefficients through a test-retest method with another administration. The final stage applied Analysis of Variance (ANOVA) repeated measures to assess the stability of respondents' responses and determine their consistency with the provided items.

Phase 1 Result

In the first phase, Cronbach's alpha was employed to evaluate the scale's internal consistency. This method calculates the average correlation among all possible combinations of items within a scale, offering a comprehensive measure of consistency (Tavakol & Dennick, 2011). Unlike the split-half method, which divides items into halves for correlation, Cronbach's alpha considers all potential item groupings, thereby providing a more reliable estimate across various contexts and populations (DeVellis, 2016). The reliability estimates using Cronbach's alpha for Utilization (UT), Awareness (AWA), and Perception (PER) of AI in research, as shown in **Table 2**, indicate robust internal consistency within each construct. High Cronbach's alpha values approaching 1.00 suggest strong reliability due to consistent internal alignment among scale items (Carmines & Zeller, 1979). Specifically, Utilization demonstrated a Cronbach's alpha of .994, indicating high consistency among UT1 to UT6 items measuring AI utilization. Similarly, Awareness showed a Cronbach's alpha of .971, reflecting strong internal consistency among AWA1 to AWA6 items assessing AI awareness. Perception exhibited a Cronbach's alpha of .942, signifying robust internal consistency across PER1 to PER3 items evaluating AI perception in research. These high alpha values underscore the reliability of the measurement instruments for assessing participants' utilization behaviors, awareness levels, and perceptions of AI in research settings. In summary, the study's high Cronbach's alpha values (.994 for Utilization, .971 for Awareness, and .942 for Perception) validate the internal consistency of the scales used. These findings support the reliability of the measurement instruments, thereby enhancing the validity of the study's outcomes and conclusions (Carmines & Zeller, 1979).

Phase 2 Result

The second phase of the analysis utilized the test-retest method. The findings presented in **Table 3** indicate that the mean score for Utilization across all items (UT1 to UT6) was 15.43 in the initial administration, increasing to 16.28 in the subsequent administration. The standard deviation for Utilization was 4.36 initially and 4.33 upon retest. For Awareness, the mean score across items (AWA1 to AWA6) was 19.37 initially and decreased to 17.32 upon retest, with standard deviations of 3.67 and 5.40 respectively. In Perception, the mean score across items (PER1 to PER3) was 7.88 initially and rose to 8.19 upon retest, with standard deviations of 1.76 and 2.08 respectively. The correlation coefficients (0.428 for Utilization, 0.538 for Awareness, and 0.601 for Perception) indicate the degree of stability or consistency in responses over time. Correlations above 0.50 are generally considered moderate to strong, suggesting reasonable reliability across test-retest administrations (Bland & Altman, 1996). This suggests a consistent pattern of responses or measurements over repeated administration. However, the interpretation of reliability coefficients depends on factors such as the nature of the measurement instrument, the stability of the construction being measured, and the study's objectives. Acceptable reliability levels can vary depending on the specific context and goals of the research.

Table 3. Reliability estimates of student's utilization, awareness and perception of AI in research using test-retest method

Measures	1st Mean	1st SD	2nd Mean	2nd SD	Correlation
UT1	2.58	.74	2.83	.86	
UT2	2.53	.70	2.65	.72	
UT3	2.58	.74	2.70	.75	
UT4	2.54	.74	2.71	.76	
UT5	2.55	75	2.70	.75	
UT6	2.56	.75	2.72	.77	
Utilization	15.43	4.36	16.28	4.33	.428
AWA1	3.32	.63	2.82	.87	
AWA2	3.33	.63	2.90	.92	
AWA3	3.32	.64	2.90	.91	
AWA4	3.30	.59	2.92	.91	
AWA5	3.31	.63	2.91	.92	
AWA6	2.82	.83	2.93	.93	
Awareness	19.37	3.67	17.32	5.40	.538
PER1	2.48	.64	2.65	.77	
PER2	2.71	.60	2.77	.69	
PER3	2.69	.61	2.78	.70	
Perception	7.88	1.76	8.19	2.08	.601

SD=standard deviation

Table 4. Reliability estimate using analysis of variance (ANOVA) repeated measures

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	η^2	ObservedPower[a]
Utilization	Sphericity Assumed	77.207	2	38.603	3.065	.049	.030	.587
	Greenhouse-Geisser	77.207	1.820	42.410	3.065	.054	.030	.559
	Huynh-Feldt	77.207	1.853	41.672	3.065	.053	.030	.564
	Lower-bound	77.207	1.000	77.207	3.065	.083	.030	.411
Error	Sphericity Assumed	2494.127	198	12.597				
	Greenhouse-Geisser	2494.127	180.227	13.839				
	Huynh-Feldt	2494.127	183.421	13.598				
	Lower-bound	2494.127	99.000	25.193				
Awareness	Sphericity Assumed	1242.407	2	621.203	49.671	.000	.334	.873
	Greenhouse-Geisser	1242.407	1.957	634.953	49.671	.000	.334	.889
	Huynh-Feldt	1242.407	1.996	622.542	49.671	.000	.334	1.00
	Lower-bound	1242.407	1.000	1242.407	49.671	.000	.334	.822
Error	Sphericity Assumed	2476.260	198	12.506				
	Greenhouse-Geisser	2476.260	193.712	12.783				
	Huynh-Feldt	2476.260	197.574	12.533				
	Lower-bound	2476.260	99.000	25.013				
Perception	Sphericity Assumed	125.407	2	62.703	18.737	.000	.159	.904
	Greenhouse-Geisser	125.407	1.870	67.060	18.737	.000	.159	.874
	Huynh-Feldt	125.407	1.905	65.839	18.737	.000	.159	876
	Lower-bound	125.407	1.000	125.407	18.737	.000	.159	.990
Error	Sphericity Assumed	662.593	198	3.346				
	Greenhouse-Geisser	662.593	185.137	3.579				
	Huynh-Feldt	662.593	188.569	3.514				
	Lower-bound	662.593	99.000	6.693				

η^2=Eta Size, df=degree of freedom

Phase 3 Result

The final phase of the analysis applied repeated measures Analysis of Variance (ANOVA) to examine the stability of responses across three administrations of the instrument. This approach was chosen based on theoretical insights emphasizing the importance of consistent measurement over time to ensure reliability and accuracy (DeVellis, 2016).

Descriptive statistics from **Table 4** showed varying mean scores and standard deviations for the utilization of AI across three conditions: Utilization1 (M = 15.43, SD = 4.35), Utilization2 (M = 16.28, SD = 4.33), and Utilization3 (M = 15.07, SD = 4.78). The assumption of sphericity was violated according to Mauchly's test, χ^2 (2) = 10.174, p = .006, thus Greenhouse-Geisser correction was applied. Within-subjects' analysis revealed a significant main effect, F (1.820, 180.227) = 3.065, p = .031, η^2 = .030. Post-hoc tests indicated a significant quadratic contrast, F (1, 99) = 4.804, p = .031, η^2 = .046, suggesting a non-linear trend in utilization over time. Between-subjects' effects also showed a significant main effect of utilization, F (1, 99) = 2066.239, p < .001, η^2 = .954.

For awareness, mean scores were 19.37 (SD = 3.68) for Awareness1, 17.32 (SD = 5.40) for Awareness2, and 14.41 (SD = 5.83) for Awareness3. A multivariate analysis of variance (MANOVA) indicated a significant effect of awareness conditions, Wilks' Lambda Λ = .524, F (2, 98) = 44.487, p < .001, η^2 = .476. Specifically, a significant linear contrast was found, F (1, 99) = 89.878, p < .001, η^2 = .476, indicating a linear trend in awareness across conditions. Mauchly's test showed no violation of sphericity for awareness conditions, χ^2 (2) = 2.193, p = .334, confirming valid assumptions for repeated measures ANOVA.

Participants' perceptions of AI varied across conditions with mean scores of 7.88 (SD = 1.77) for Perception1, 8.19 (SD = 2.08) for Perception2, and 6.69 (SD = 2.43) for Perception3. A MANOVA revealed an overall significant effect of perception conditions, Wilks' Lambda $\Lambda = .759$, $F(2, 98) = 15.597$, $p < .001$, $\eta^2 = .241$. Mauchly's test indicated no violation of sphericity for perception conditions, $\chi^2(2) = 7.057$, $p = .029$, $\varepsilon = .935$, meeting assumptions for repeated measures ANOVA. Within-subjects' effects confirmed a significant main effect of perception, $F(1.870, 185.137) = 18.737$, $p < .001$, $\eta^2 = .159$ (Greenhouse-Geisser corrected), with a significant linear contrast, $F(1, 99) = 37.475$, $p < .001$, $\eta^2 = .159$. These results indicate that participants' utilization, awareness, and perception of AI showed varying patterns across the different conditions tested, underscoring the importance of repeated measures ANOVA in understanding the stability of responses over time.

DISCUSSION OF FINDINGS

In the realm of research assessing attitudes and perceptions towards Artificial Intelligence (AI), ensuring the reliability of measurement instruments is crucial for drawing valid conclusions and making meaningful interpretations. This discussion explores the findings of a study that employed Cronbach's alpha, test-retest reliability, and repeated measures ANOVA to assess the reliability of measures related to Utilization, Awareness, and Perception of AI across multiple administrations.

Initially, the study utilized Cronbach's alpha to evaluate the internal consistency of measures administered to participants. Cronbach's alpha is a widely used statistic that assesses how closely related a set of items are as a group, providing a measure of reliability by indicating the extent to which items in a scale are homogenous in measuring the same construct (Cronbach, 1951). During the first administration of the measures, Cronbach's alpha coefficients were found to be close to 1.00 for all three constructs: Utilization, Awareness, and Perception of AI. This initial result suggested high internal consistency among the items within each scale. For instance, the Cronbach's alpha coefficient of .94 for Utilization indicated that the items measuring how participants utilized AI in research were highly correlated and measured the construct reliably at that point in time (Field, 2013).

Following the assessment with Cronbach's alpha, the study proceeded to evaluate test-retest reliability to assess the stability of participants' responses over time. Test-retest reliability measures the consistency of scores across two or more administrations of the same test to the same group of participants under the same conditions (Streiner, 2003). During the second administration of the measures, the findings revealed lower test-retest reliability coefficients: .428 for Utilization, .538 for Awareness, and .601 for Perception of AI. These coefficients fell below the generally accepted threshold for reliability, indicating that participants' responses did not demonstrate stable patterns over time. Factors such as changes in participants' knowledge or experiences related to AI, as well as contextual influences, could have contributed to the variability observed in their responses (DeVellis, 2016). The discrepancy between high Cronbach's alpha coefficients initially and low test-retest reliability coefficients raises important considerations about the temporal stability of the measures. While Cronbach's alpha assesses internal consistency at a single point in time, test-retest reliability evaluates the consistency of responses over time, providing insights into the reliability of measures across different temporal contexts (Tabachnick & Fidell, 2019).

To further explore the consistency of participants' responses across different administrations, the study employed repeated measures ANOVA. Repeated measures ANOVA are used to analyze differences in means of a dependent variable across two or more time points or conditions within the same subjects (Field, 2013). The findings revealed significant differences in mean scores for all three constructs across the administrations, suggesting a lack of reliability in participants' responses over time. Several factors may contribute to the observed lack of consistency in responses as indicated by the repeated measures ANOVA results. Firstly, the dynamic nature of attitudes and perceptions towards AI could influence participants' responses across different time points. Research indicates that societal attitudes towards technology, including AI, can fluctuate due to media portrayal, technological advancements, and public discourse (Bostrom & Yudkowsky, 2014).

Moreover, individual experiences and exposure to AI technologies may vary over time, influencing how participants perceive and utilize AI in research contexts. Studies have shown that personal experiences with technology can significantly shape individuals' attitudes and behaviors towards their use (Venkatesh et al., 2012). Additionally, methodological considerations such as measurement error, participant fatigue, or changes in survey conditions across administrations could also contribute to the lack of response consistency. Variations in survey administration protocols or differences in contextual factors between sessions may introduce systematic biases or measurement artifacts (Field, 2013).

Research on the stability of attitudes and perceptions over time provides further insights into the challenges observed in achieving response consistency in this study. For instance, studies examining the longitudinal stability of attitudes towards emerging technologies have found that opinions can change as individuals gain new information or experiences (Kaiser et al., 2019). Furthermore, meta-analytic reviews of test-retest reliability across various psychological constructs suggest that temporal changes in attitudes and behaviors are common, underscoring the need for robust measurement strategies that account for such variability (Bland & Altman, 1996). On the other hand, some studies suggest that with careful design and control, it is possible to achieve reliable measures of perceptions and attitudes over time. Longitudinal studies using sophisticated statistical techniques, such as latent growth modelling or structural equation modelling, have demonstrated that stable constructions can be reliably measured despite potential fluctuations in responses (Little et al., 2007).

CONCLUSION/IMPLICATIONS OF THE STUDY

The study initially demonstrated strong reliability in measuring utilization, awareness, and perception of AI using Cronbach's alpha during the first administration, indicating internal consistency and reliability at that time. However, subsequent administrations showed

decreased reliability with coefficients falling below acceptable thresholds for test-retest reliability. The final analysis using repeated measures ANOVA highlighted significant differences in mean scores across administrations for utilization, awareness, and perception, suggesting instability and variability in the constructs measured. Thus, indicating insufficient evidence for reliability.

Studying carries both practical and theoretical implications. Firstly, it underscores the need for robust measurement techniques beyond Cronbach's alpha, especially for constructs prone to change over time, such as attitudes towards emerging technologies like AI. Future research should consider dynamic measures or longitudinal approaches to accurately capture shifts in perceptions. In practical applications, such as longitudinal studies or program evaluations involving AI attitudes, researchers should be cautious about assuming response stability over time based solely on initial reliability assessments. Continuous monitoring and adjustment of measurement strategies may be crucial to account for temporal fluctuations.

Theoretically, the study contributes to understanding psychological measurement reliability by demonstrating the limitations of static reliability estimates in capturing temporal dynamics. It advocates for an integrated approach that considers both reliability and validity across multiple time points for accurate data interpretation. Future studies could explore alternative statistical methods or measurement models to better accommodate variability in AI perceptions. This might involve advanced psychometric techniques or qualitative methods to complement quantitative findings and offer deeper insights into the evolution of perceptions.

Author contributions: UJO: conceptualisation, methodology writing of the original draft, validation, formal analysis; **VJO:** conceptualization, methodology, review and editing, supervision, formal analysis; **CI:** supervision, data curation, formal analysis; **SVO:** conceptualization, data curation, supervision. All authors have agreed to publish the findings of the study.

Acknowledgments: All parties that contributed to this study are largely acknowledged.

Funding: Authors declared that there is no funding source reported for this study.

Ethical statement: The authors stated that the study was conducted in accordance with ethical research principles. The authors further stated that the research involved a survey that did not require formal ethical approval by the institution, as it posed minimal risk to participants and did not involve sensitive or personally identifiable information. Participation was entirely voluntary, with respondents providing informed consent before completing the survey. All data was collected anonymously and used solely for academic purposes. Confidentiality and privacy were strictly maintained throughout the research processed.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data will be made available at reasonable request.

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.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.

Anderson, T., & Williams, K. (2017). The effects of high-intensity interval training on endurance performance in athletes. *Journal of Sports Science, 35*(5), 467-478. https://doi.org/10.1080/02640414.2016.1172729

Blanchard, M. R., Harris, R. J., & Jones, M. K. (2010). Impact of a digital literacy program on students' technological skills. *Computers & Education, 55*(3), 888-896. https://doi.org/10.1016/j.compedu.2010.03.004

Bland, J. M., & Altman, D. G. (1996). Statistics notes: Measurement error and correlation coefficients. *BMJ, 313*(7048), 41-42. https://doi.org/10.1136/bmj.313.7048.41

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine, 21*(9), 1331-1335. https://doi.org/10.1002/sim.1108

Bostrom, N., & Yudkowsky, E. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brown, A. D., & Green, P. C. (2020). Evaluating the effectiveness of leadership training on managerial skills. *Journal of Organizational Behavior, 41*(2), 159-175. https://doi.org/10.1002/job.2430

Brown, J. D., & Hudson, T. (2020). Reliability. In J. D. Brown, & T. Hudson (Eds.), *The encyclopedia of language and linguistics* (2nd ed., Vol. 10, pp. 106-109). Elsevier. https://doi.org/10.1016/B978-0-08-102297-1.10527-8

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185-216. https://doi.org/10.1177/135910457000100301

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage Publications. https://doi.org/10.4135/9781412985642

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approach* (4th ed.). SAGE Publications

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed.). Sage Publications.

Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health, 58*(4), 637-642. https://doi.org/10.1007/s00038-012-0416-3

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates. https://doi.org/10.1037/10519-153

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage Publications.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). Wiley. https://doi.org/10.1002/9781119513469

Girden, E. R. (1992). *ANOVA: Repeated measures*. Sage Publications.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for likert-type scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education* (pp. 82-88). Ohio State University.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95-112. https://doi.org/10.1007/BF02289823

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hays, W. L. (2018). *Statistics*. Cengage Learning.

Hogan, T. P. (2007). *Psychological testing: A practical introduction* (2nd ed.). John Wiley & Sons.

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Wadsworth.

Kaiser, F. G., Hübner, G., & Bogner, F. X. (2019). Contrasting the theory of planned behavior with the value-belief-norm model in explaining conservation behavior. *Journal of Applied Social Psychology, 39*(7), 1550-1570.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Pearson.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Lee, J., & Johnson, K. (2019). Impact of advertising campaigns on brand perception: A repeated measures study. *Journal of Consumer Research, 46*(3), 621-635. https://doi.org/10.1093/jcr/ucz029

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development, 31*(4), 357-365. https://doi.org/10.1177/0165025407077757

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587-604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410609243

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

Millsap, R. E., & Maydeu-Olivares, A. (Eds.). (2009). *The SAGE handbook of quantitative methods in psychology*. SAGE Publications. https://doi.org/10.4135/9780857020994

Morrison, D. F. (2018). *Multivariate statistical methods*. John Wiley & Sons.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Ofem, U. J., Iyam, M. A., Ovat, S. V., Nworgwugwu, E. C., Anake, P. M., Udeh, M. R., & Otu, B.D. (2024) Artificial intelligence (AI) in academic research. A multi-group analysis of students' awareness and perceptions using gender and programme type. *Journal of Applied Learning & Teaching, 7*(1), 76-92. https://doi.org/10.37074/jalt.2024.7.1.9

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. P*sychological Methods, 8*(4), 434-447. https://doi.org/10.1037/1082-989X.8.4.434

Schmitt, N. (1996). Uses and abuses of coefficient alpha. Psychological Assessment, 8(4), 350-353. https://doi.org/10.1037/1040-3590.8.4.350

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications. https://doi.org/10.1037/10109-051

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120. https://doi.org/10.1007/s11336-008-9101-0

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102-111. https://doi.org/10.1037/1040-3590.12.1.102

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103. https://doi.org/10.1207/S15327752JPA8001_18

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford University Press.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thissen, D., & Wainer, H. (2001). *Test scoring*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410604729

Thompson, L. A., Wilson, T. K., & Brown, C. E. (2018). Tracking cognitive decline in older adults: A longitudinal neuropsychological study. *Neuropsychology, 32*(3), 345-356. https://doi.org/10.1037/neu0000441

Trochim, W. M. (2006). *Research methods knowledge base*. Atomic Dog Publishing.

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly, 36*(1), 157-178. https://doi.org/10.2307/41410412

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). McGraw-Hill.