Journal of Entrepreneurship, Business and Economics
ISSN 2345-4695
2025, 13(1), 135–186

QUESTIONNAIRE VALIDITY AND RELIABILITY: A REVIEW WITH PRACTICAL GUIDELINES

Dalowar Hossan [1]; Bert Wolfs [1]; Noor-E-Medina Suraiya Jesmin [2]; Milos Petkovic [1]

[1] SBS Swiss Business School, Kloten-Zurich, Switzerland
[2] Department of Law and Human Right, University of Asia Pacific, Dhaka, Bangladesh
E-mail: d.hossan@research.sbs.edu, dalowarhossan.bd@gmail.com

Received April 2025; accepted June 2025

Abstract
When it comes to producing meaningful and trustworthy research, ensuring validity and reliability is absolutely essential, no matter which methodology is used. This article brings together practical advice on how to design and assess questionnaires effectively for both quantitative and qualitative studies. It highlights the value of relying on well-established scales with proven reliability, running pilot tests to perfect the instruments, and consulting subject-matter experts to make sure the questions truly reflect the concepts being measured. In quantitative research, it is especially important to check internal consistency and apply statistical techniques to test how well the scales perform, aiming for results that are both solid and broadly applicable. On the qualitative side, being transparent about the research process and staying reflective as a researcher are key to building trust and reducing bias. The article also encourages blending methods when appropriate, as this can provide richer data and help confirm findings through triangulation. Ultimately, a thoughtfully designed questionnaire, tested and validated properly, can significantly improve the strength, credibility, and overall value of research across various fields.

Research Paper

Keywords: Questionnaire Validity, Questionnaire Reliability, Quantitative Study, Qualitative Study

Reference to this paper should be made as follows: Hossan, D., Wolfs, B., Jesmin, N.M.S., & Petkovic, M. (2025). Questionnaire Validity and Reliability: A Review with Practical Guidelines. *Journal of Entrepreneurship, Business and Economics*, *13*(1), 135–186.

Introduction

Questionnaires are one of the most common tools used to collect data in both qualitative and quantitative research (Taherdoost, 2022). They offer a practical way to obtain consistent information from participants and are widely used across a variety of disciplines including medicine, natural sciences, business, and the social sciences. Despite their popularity, the usefulness and credibility of the data gathered through questionnaires hinge largely on two crucial factors: validity and reliability (Anderson et al., 2022).

Reliability refers to the consistency, stability, and repeatability of measurements over time, across different items, and among various observers (Bujang et al., 2024). It reflects how consistently an instrument produces the same results under similar conditions, reducing the impact of random errors (Tavakol & Dennick, 2011). In quantitative research, high reliability is necessary to establish the trustworthiness of results and serves as a prerequisite for validity (Adeyemi, 2024). For instance, if a scale measuring employee work engagement provides different scores for the same individual across repeated administrations without any real change in the underlying construct, the scale would be deemed unreliable. There are three types of reliability (test-retest reliability**,** internal consistency**,** inter-rater reliability**)** commonly discussed in quantitative research (Shelley et al., 2024)**.** While, in qualitative research, the notion of reliability translates to dependability (Lim, 2024). Rather than focusing on replicability in the statistical sense, dependability emphasizes the consistency of the research process. Researchers ensure dependability through an audit trail, documentation of decisions, and transparent pro-

cedural steps, allowing others to trace the path from data collection to findings. Validity refers to the degree to which an instrument accurately measures what it is intended to measure. It reflects the truthfulness, relevance, and meaningfulness of the inferences drawn from research findings (Lopez et al., 2023). In quantitative research, validity is a multidimensional concept that includes content, construct, and criterion-related validity. In qualitative research, validity is conceptualized through the lens of trustworthiness (Ahmed, 2024), encompassing several components such as, credibility, transferability, confirmability, dependability.

A reliable questionnaire yields similar results under consistent conditions and across different times, raters, or items. Without reliability, researchers cannot be certain whether observed differences in data are due to actual variation in the construct or merely measurement error (Chetwynd et al., 2022). Validity, on the other hand, concerns the accuracy and relevance of the measurement tool that examines whether the instrument truly measures the construct it purports to measure (Chetwynd et al., 2022). Importantly, a measurement tool can be reliable without being valid, but it cannot be valid unless it is reliable (Tavakol & Dennick, 2011). In quantitative research, validity and reliability are often assessed using statistical techniques such as Cronbach's alpha, factor analysis, and correlation with external benchmarks (Millar et al., 2022). In contrast, qualitative research approaches validity and reliability through a different lens known as trustworthiness (Ahmed, 2024; Golafshani, 2003).

As researchers across various fields place greater emphasis on methodological rigor, there is a growing need to approach questionnaire design

with equal attention to both validity and reliability (Bauer et al., 2024). This can be particularly challenging for those working with mixed-methods, where maintaining measurement quality across both qualitative and quantitative strands is essential. To address this gap, the current paper aims to offer a well-rounded discussion on these two fundamental aspects of questionnaire development. By drawing on both theory and real-world examples from established studies (such as Golafshani, 2003; Tavakol & Dennick, 2011; Choi & Pak, 2005), it brings together insights from different research traditions. The goal is to equip researchers with clear, practical guidance that can help improve the validity, reliability, and overall trustworthiness of their instruments across a wide range of academic and applied settings.

Considering the aforementioned challenges, this research seeks to discuss key issues through the following set of questions:

(1) How are validity and reliability conceptualized and applied within both quantitative and qualitative research paradigms?

(2) What challenges do researchers face when designing questionnaires, and what ethical considerations must they take into account during the process?

(3) How can a researcher ensure validity and reliability in their studies through the application of best practices (that have been recommended in this article)?

In the first part of the introduction, the researcher outlines the fundamental concepts of validity and reliability, then emphasises their importance and interrelationship. The second part discusses the application of validity

and reliability in both quantitative and qualitative research, providing practical guidelines for implementation. The third part addresses the challenges researchers encounter during questionnaire development, along with the ethical considerations they must take into account. Finally, the concluding section presents recommendations and best practices for ensuring validity and reliability in research.

Methods

The information used in this article came from the various fields of published research articles. The flowchart (figure 1) explains how these were chosen from various potential options. Figure 1 illustrates the types of publications that were considered for and rejected from further consideration at each level. All of the material was looked through using ScienceDirect/ Elsevier as the data-bases of choice. This study assesses the validity and reliability of a newly developed questionnaire using construct validity and construct reliability. Therefore, it was determined that the following search phrases would be helpful: “construct validity” and “construct reliability”.

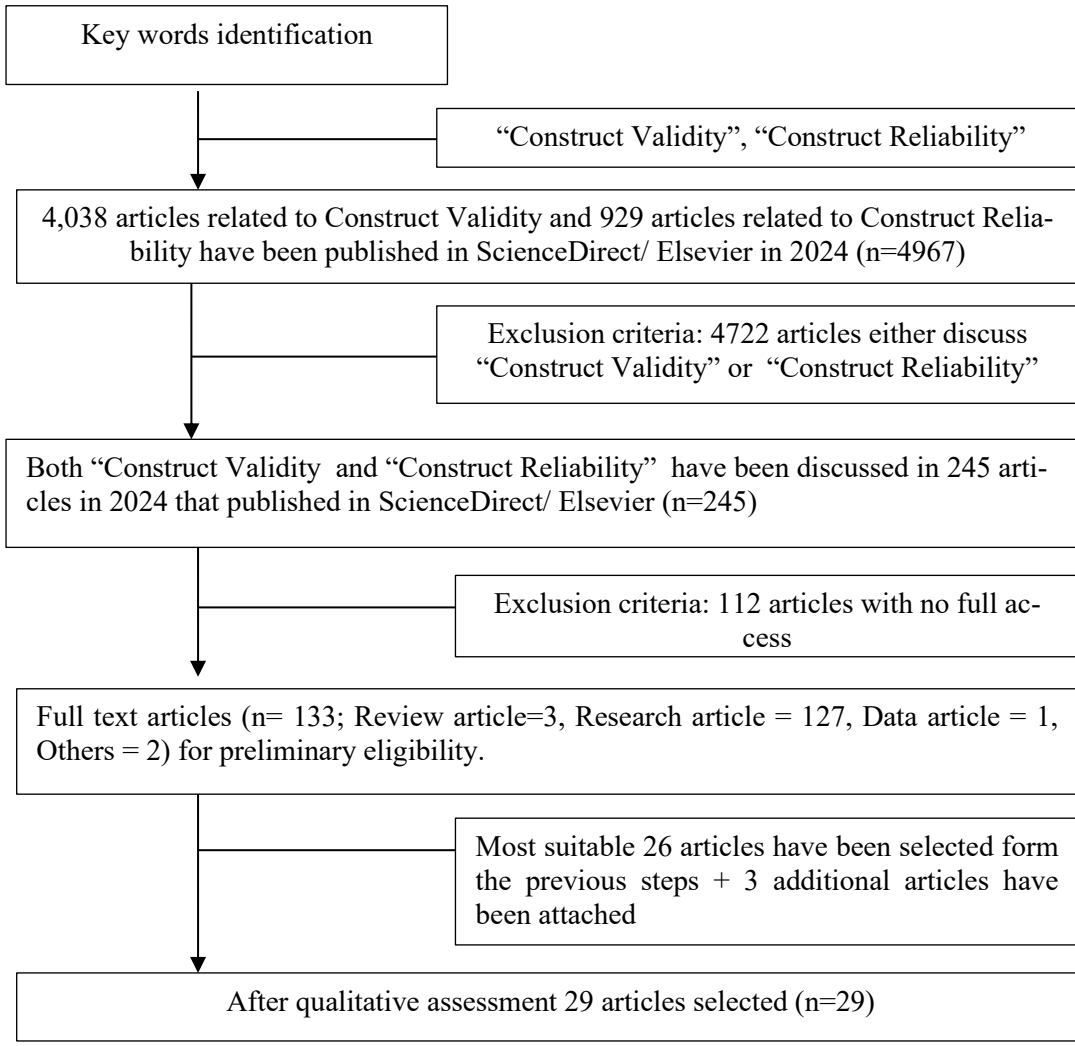


Figure 1. Sampling Process

All duplicate records and irrelevant documents were meticulously screened to ensure the reliability and rigor of this systematic review. The initial search identified a total of 4,038 articles related to construct validity and 929 articles related to construct reliability, resulting in a total of 4,967 articles

published in 2024 on ScienceDirect/Elsevier. After applying exclusion criteria to remove 4,722 articles that discussed only one of the two constructs, a total of 245 articles that addressed both "construct validity" and "construct reliability" were retained for further consideration. Among these, 112 articles were excluded due to lack of full-text access, leaving 133 eligible articles (comprising 3 review articles, 127 research articles, 1 data article, and 2 others). To assure the quality and relevance of the literature, article abstracts were rigorously evaluated, followed by a full-text analysis to determine eligibility. Ultimately, based on qualitative assessment and purposive sampling aligned with the study's objectives, 26 articles with additional 3 articles, n=29 were selected for inclusion (list in appendix). The selection process ensured that only the most relevant and illustrative studies were incorporated, enhancing the robustness of the analysis. This methodological rigor ensures that the conclusions drawn are grounded in high-quality and comprehensive academic evidence.

Discussion

Validity and Reliability in Quantitative Research

Types of Reliability in Quantitative Research

In quantitative research, reliability ensures that the measurement tool consistently produces the same results when used under similar conditions. It is a key element for establishing the credibility of quantitative findings. Different types of reliability play distinct roles, depending on the type of instrument and the context in which data is collected. This section explores three

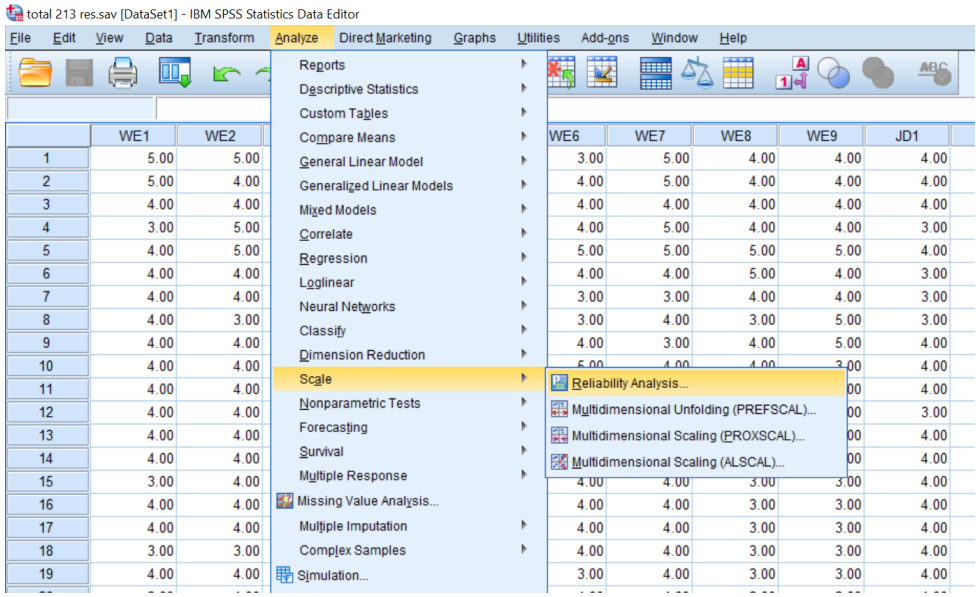
main types of reliability that are particularly relevant to questionnaire-based research.

Internal Consistency

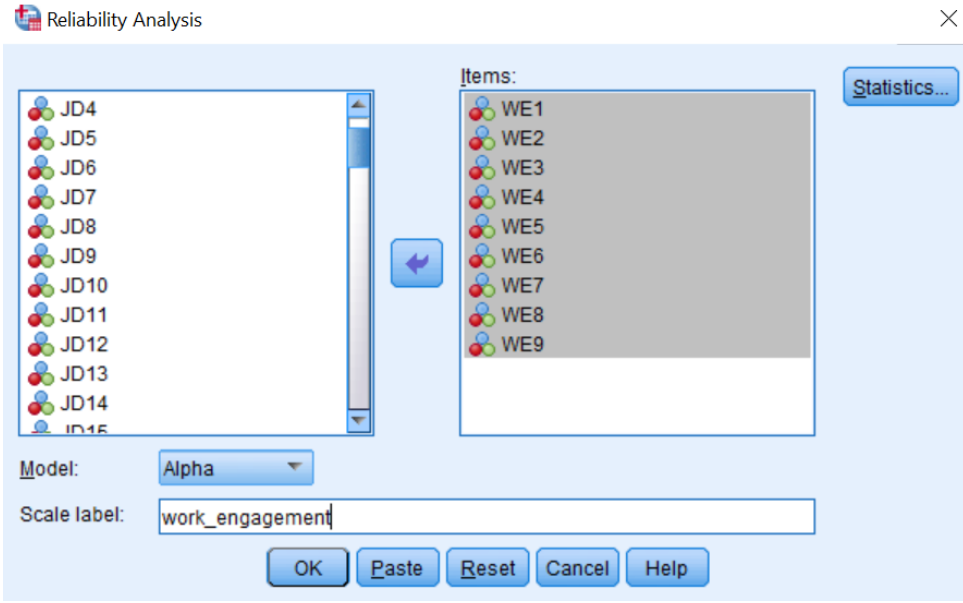
Internal consistency refers to how closely the items within a questionnaire relate to one another and measure the same underlying concept (Kang et al., 2024). It assesses whether the items work together to evaluate the same idea or construct. Cronbach's alpha is the most commonly used statistical method to determine internal consistency (Nurhidayati & Sukri, 2025). A value of 0.70 or higher is typically considered acceptable, suggesting that the instrument has adequate internal reliability (Adeniran, 2025).

For instance, a questionnaire designed to assess employee work engagement might include items related to vigour, dedication, and absorption. Internal consistency ensures that these items are strongly correlated and collectively represent the broader concept of employee work engagement. High internal consistency enhances the precision and clarity of the measurement, making the results easier to interpret and trust. However, it is important to note that an excessively high Cronbach's alpha (e.g., above 0.90) could indicate that some items are too similar, leading to redundancy and diminishing the unique value of each item (Adeniran, 2025). Researchers need to find a balance, ensuring that the items are connected but not repetitive (Tristan-Lopez, 2025).

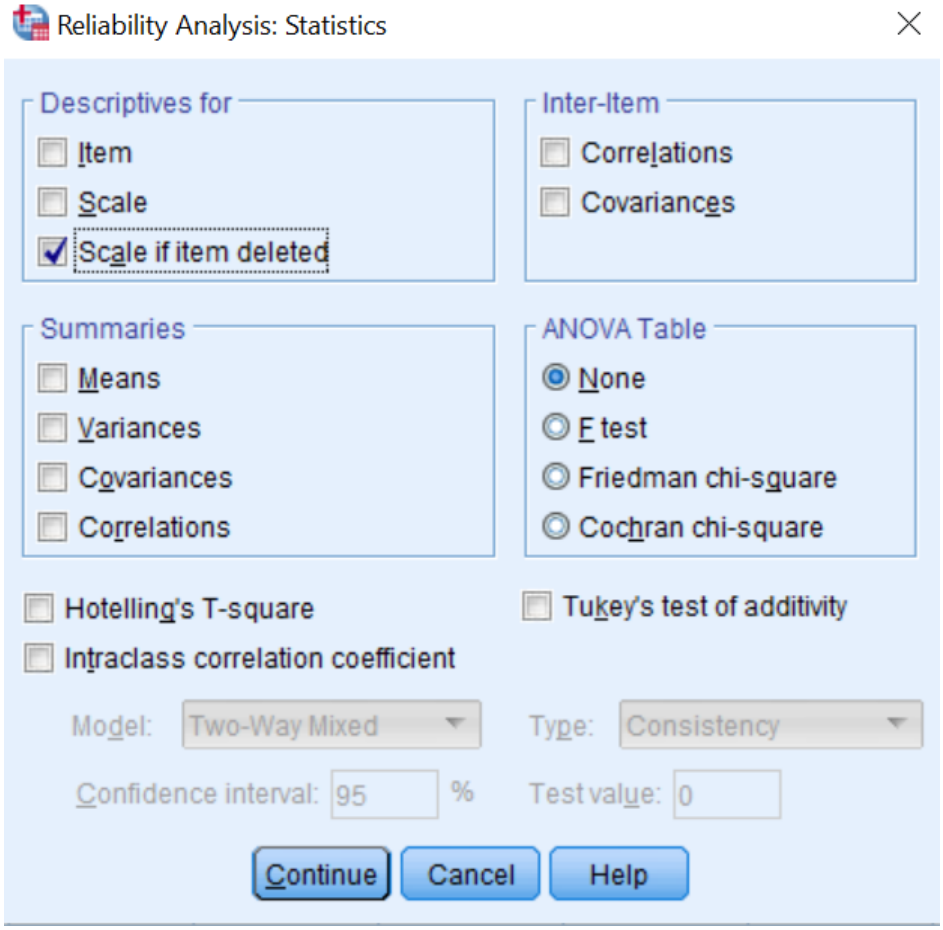
The steps for determining internal consistency are given below:



Step 1. Internal Consistency (Analyze > Scale> Reliability analysis)



Step 2. Internal Consistency (Transfer the Items of Work Engagement from Left to Right)



Step 3. Internal Consistency (Statistics > Descriptive for Scale if Item Deleted)

The internal consistency of the work engagement scale, as assessed by Cronbach's Alpha, is 0.761, indicating acceptable reliability for research purposes. The scale comprises 9 items, and most items show moderate to strong corrected item-total correlations, suggesting that they are appropriately aligned with the overall construct. Notably, item WE3 has the highest item-total correlation (0.692), indicating strong consistency with the total scale,

while item WE6 shows a very low correlation (0.040), suggesting it may not fit well with the other items.

Reliability

Scale: work_engagement

Case Processing Summary

		N	%
Cases	Valid	213	100.0
	Excluded[a]	0	.0
	Total	213	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

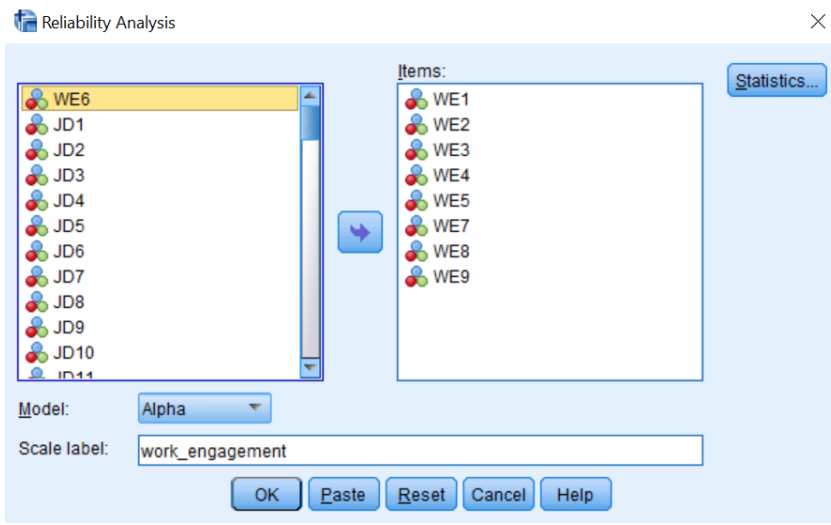
Cronbach's Alpha	N of Items
.761	9

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
WE1	31.7793	9.012	.380	.748
WE2	31.7324	8.876	.528	.729
WE3	31.6854	7.773	.692	.696
WE4	31.8028	8.546	.560	.722
WE5	31.8826	9.444	.230	.771
WE6	31.8169	10.160	.040	.798
WE7	31.6995	8.843	.368	.751
WE8	31.9765	7.617	.725	.689
WE9	31.9718	8.244	.535	.723

SPSS Output Viewer Window 1: Internal Consistency

If WE6 were removed, Cronbach's Alpha would increase to 0.798, indicating improved internal consistency. This suggests that removing WE6 could enhance the scale's reliability.



Step 4. Internal Consistency (Transfer WE6 from Right to Left)

If WE5 were removed, Cronbach's Alpha would increase to 0.811, indicating improved internal consistency.

Reliability Statistics

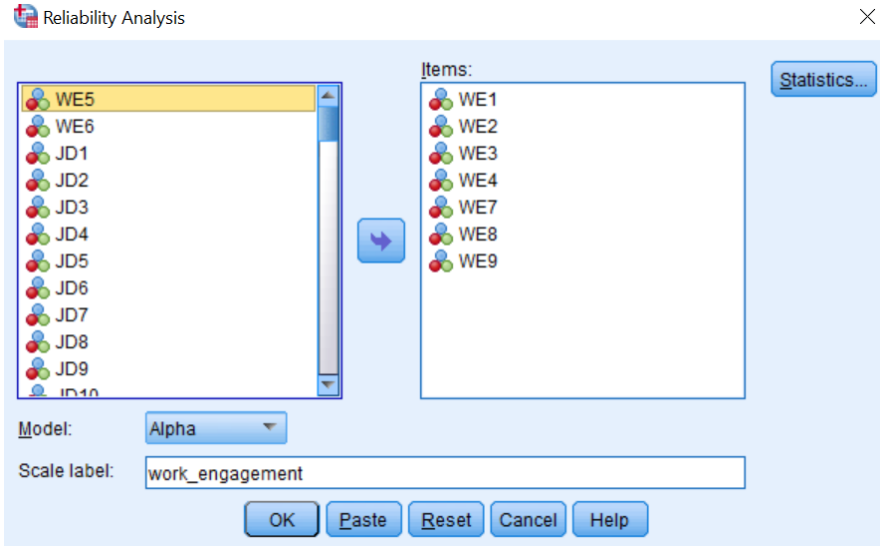
Cronbach's Alpha	N of Items
.798	8

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
WE1	27.8028	8.367	.418	.788
WE2	27.7559	8.327	.540	.773
WE3	27.7089	7.217	.717	.740
WE4	27.8263	8.125	.531	.772
WE5	27.9061	8.765	.270	.811
WE7	27.7230	8.296	.376	.796
WE8	28.0000	7.198	.705	.741
WE9	27.9953	7.731	.539	.770

SPSS Output Viewer Window 2: Internal Consistency

Item WE5 shows a very low correlation (0.270), suggesting it may not fit well with the other items.



Step 5. Internal Consistency (Transfer WE5 from Right to Left)

This suggests that removing WE6 and WE5 could enhance the scale's reliability, and their inclusion should be reconsidered in future refinements of the instrument.

Reliability Statistics

Cronbach's Alpha	N of Items
.811	7

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
WE1	23.8920	7.087	.418	.807
WE2	23.8451	7.169	.495	.795
WE3	23.7981	6.011	.726	.751
WE4	23.9155	6.785	.562	.784
WE7	23.8122	6.889	.416	.810
WE8	24.0892	6.006	.710	.754
WE9	24.0845	6.549	.523	.791

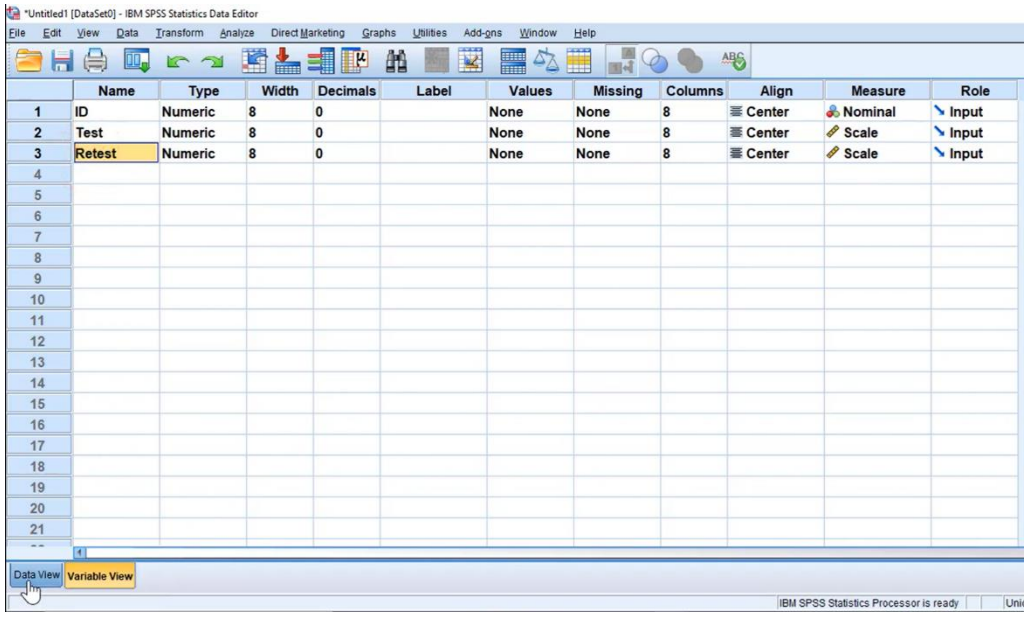
SPSS Output Viewer Window 3: Internal Consistency

Test-Retest Reliability

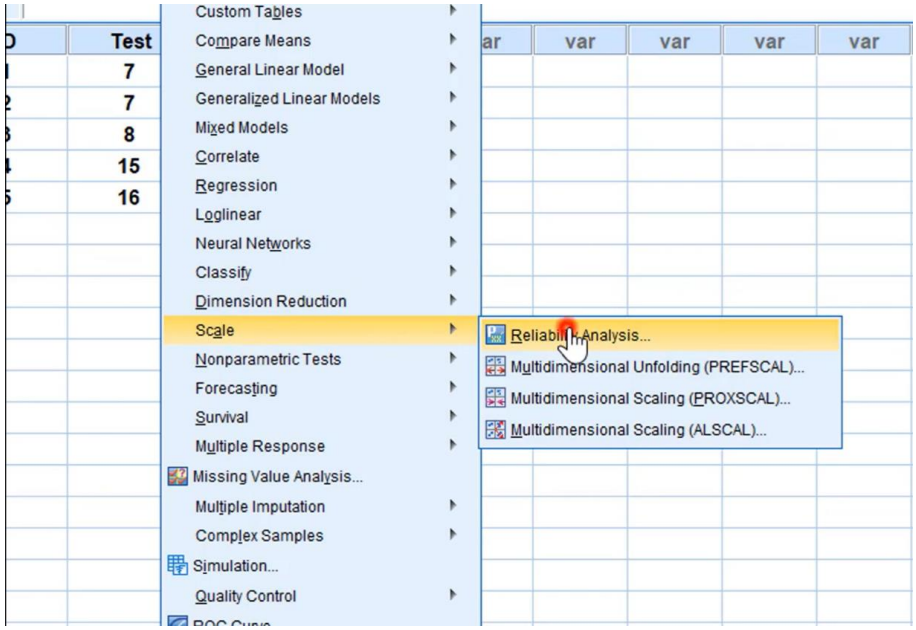
Test-retest reliability measures the stability and consistency of a questionnaire over time. This involves administering the same instrument to the same group of respondents on two different occasions and then comparing the results (Vilagut, 2024). A high test-retest reliability indicates that the instrument consistently measures a construct that remains stable over time (Fu et al., 2024; Jung et al., 2024). This type of reliability is especially important in longitudinal studies or when monitoring changes in stable constructs, such as personality traits or attitudes (Gelino et al., 2024). The time gap between the two test administrations must be just right, not so short that memory effects influence the results, and not so long that actual changes in the construct

are likely. A strong correlation (e.g., Pearson's r > 0.70) between the initial and follow-up scores suggests good temporal stability.

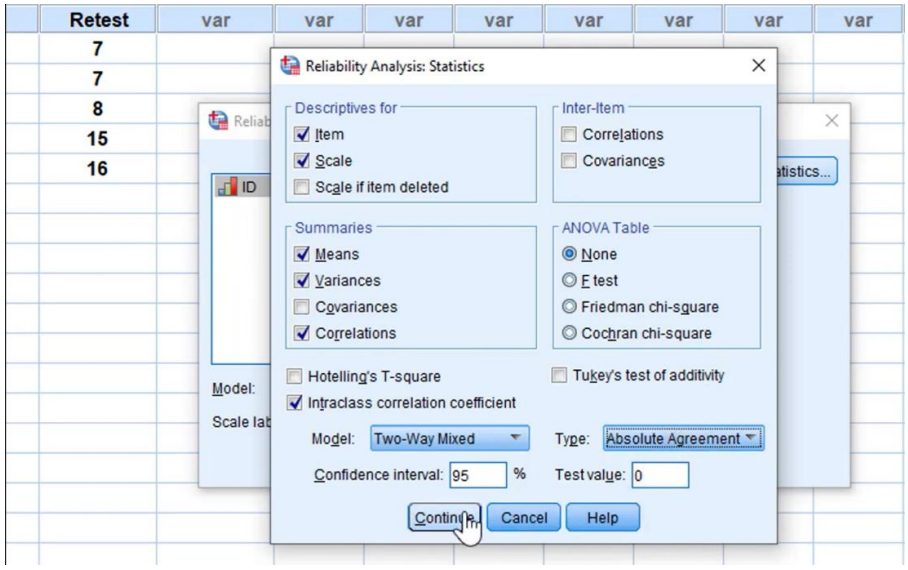
The following steps outline the process for assessing test-retest reliability:



Step 1. Test-Retest Reliability (Variable View)

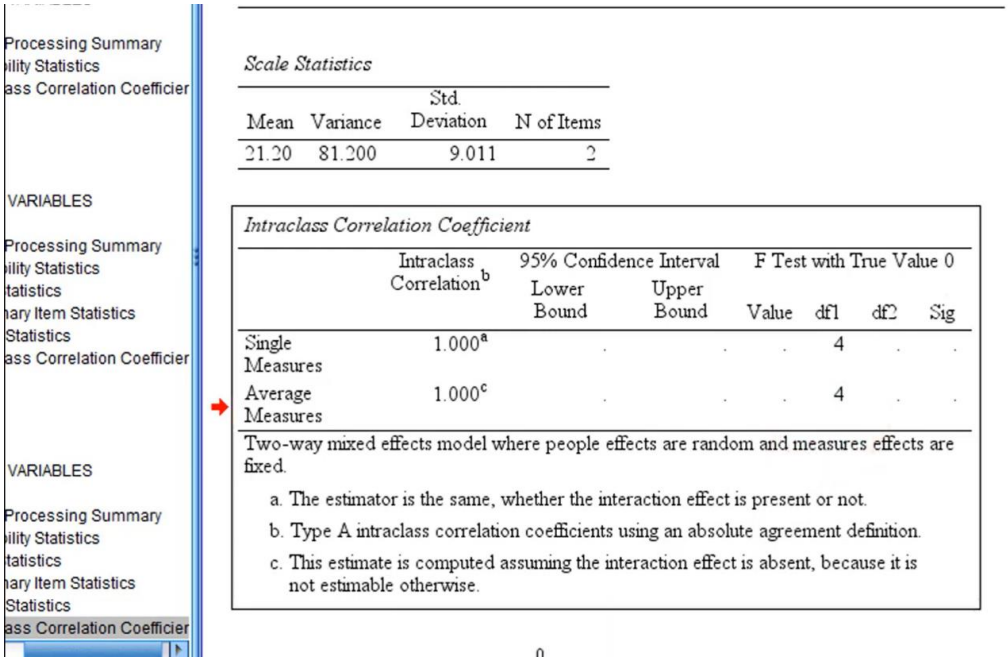


Step 2. Test-Retest Reliability (Analyse > Scale> Reliability analysis)



Step 3. Test-Retest Reliability (Descriptive for Items and Scale> Summaries: Means, Variances, Correlations> Intraclass Correction coefficient> Model: Two-Way Mixed> Type: Absolute Agreement)

The test-retest reliability of the instrument was assessed using the Intraclass Correlation Coefficient (ICC), which evaluates the degree of consistency or reproducibility of measurements over time. The ICC results show a perfect reliability score of 1.000 for both single and average measures, indicating excellent test-retest reliability. This result implies that the responses were perfectly stable across the two measurement points. The analysis used a two-way mixed effects model, assuming that people effects are random and measurement effects are fixed. Such a high ICC suggests that the instrument yields highly consistent results over time, demonstrating strong temporal stability. However, perfect reliability is unusual and may warrant further investigation to ensure no methodological issues, such as identical repeated inputs, influenced the result.



Processing Summary
ility Statistics
ass Correlation Coefficier

VARIABLES

Processing Summary
ility Statistics
tatistics
ary Item Statistics
Statistics
ass Correlation Coefficier

VARIABLES

Processing Summary
ility Statistics
tatistics
ary Item Statistics
Statistics
ass Correlation Coefficier

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
21.20	81.200	9.011	2

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	1.000[a]	.	.	.	4	.	.
Average Measures	1.000[c]	.	.	.	4	.	.

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

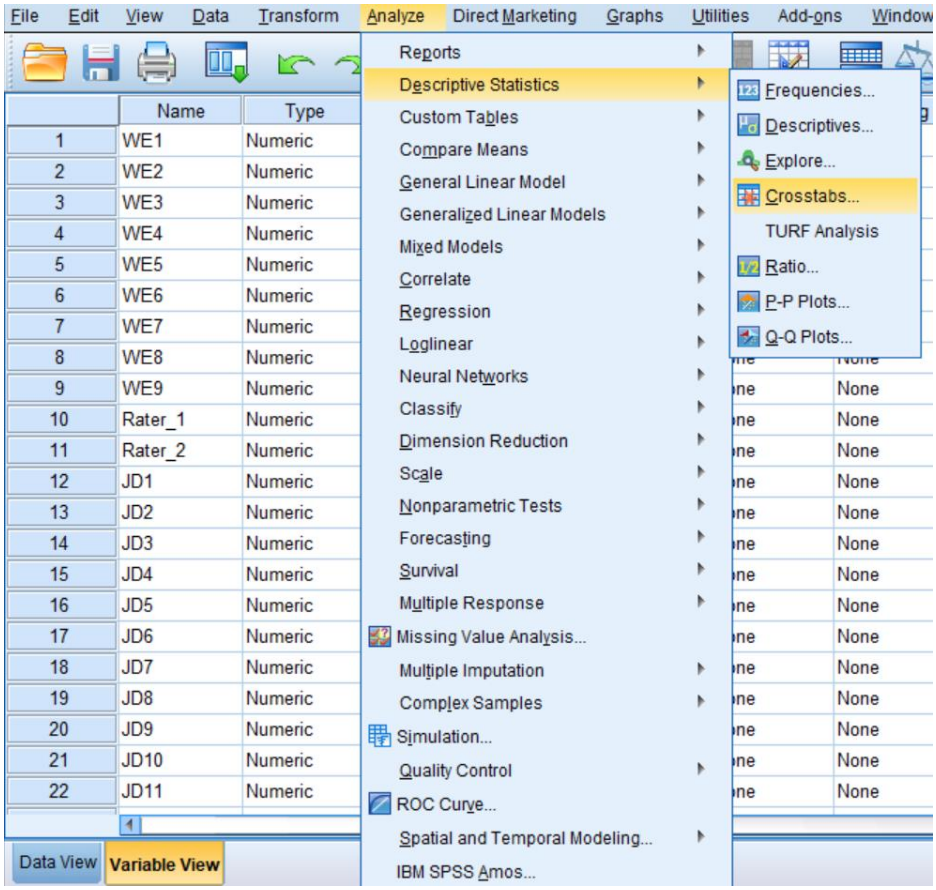
SPSS Output Viewer Window 4: Test-Retest Reliability

Inter-Rater Reliability

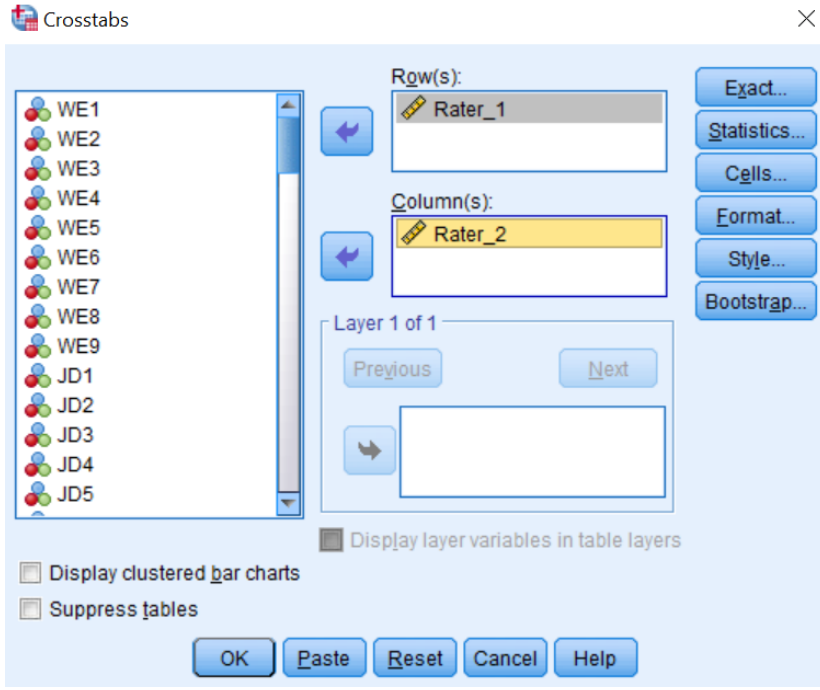
Inter-rater reliability is relevant when subjective judgment plays a role in data collection, such as when raters assess open-ended questionnaire responses, observe behavior, or code qualitative data in mixed-methods research (Cole, 2024). It measures the degree of agreement or consistency between different raters who are independently scoring or interpreting the same data (Tulliez et al., 2025, Ricci-Bonot et al., 2024). Common statistical methods for evaluating inter-rater reliability include Cohen's kappa, which accounts for agreement that might happen by chance, and the intraclass correlation coefficient, which is used for continuous data. A high degree of agreement between raters boosts the credibility and objectivity of the findings.

For example, in an employee work engagement survey where supervisors or managers are responsible for rating employees' engagement behaviors, ensuring high inter-rater reliability is essential for obtaining consistent and objective data. This type of reliability measures how similarly different raters evaluate the same individual, helping to minimize measurement errors that arise from subjective interpretation. To improve inter-rater reliability, researchers should use a standardized behavioral rating scale with clearly defined items that align with key engagement dimensions (e.g., vigor, dedication, absorption). Providing thorough training for raters, establishing clear coding guidelines, and conducting a pilot test are crucial steps to align evaluations and identify any inconsistencies. By following these strategies, researchers can greatly enhance the accuracy and reliability of their engagement assessments, ultimately strengthening the validity and robustness of their quantitative findings.

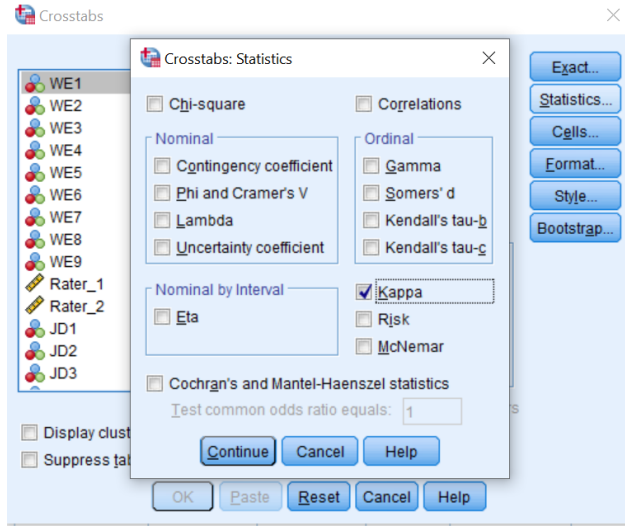
The process for assessing inter-rater reliability is described as follows:



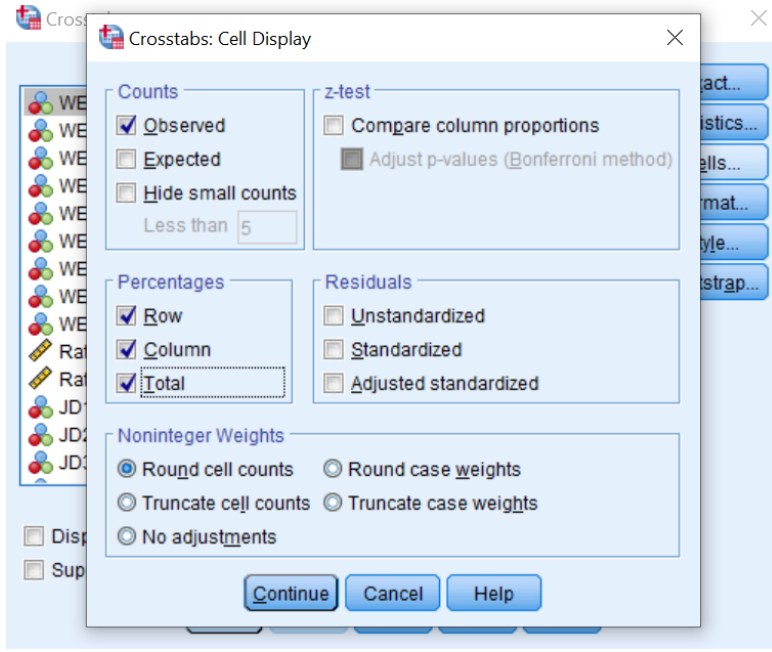
Step 1. Inter-Rater Reliability (Analyze> Descriptive Statistics> Crosstabs)



Step 2. Inter-Rater Reliability (Row and Column)



Step 3. Inter-Rater Reliability (Statistics: Kappa)



Step 4. Inter-Rater Reliability (Counts: Observed> Percentages: Row, Column, Total)

The inter-rater reliability analysis was conducted using Cohen's Kappa to assess the level of agreement between pairs of raters. For Rater_1 and Rater_2, the Kappa value was 0.093, indicating a slight level of agreement. However, this result was not statistically significant ($p = 0.236$), suggesting that the agreement observed could be due to chance.

	% of Total	0.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0
9.00	Count	0	0	0	0	0	0	1	1	
	% within Rater_1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	33.3%	33.3%	3
	% within Rater_2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	33.3%	50.0%	5
	% of Total	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.0%	5.0%	5
10.00	Count	0	0	0	1	0	0	0	0	
	% within Rater_1	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0
	% within Rater_2	0.0%	0.0%	0.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0
	% of Total	0.0%	0.0%	0.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0
Total	Count	1	5	1	2	1	3	3	2	
	% within Rater_1	5.0%	25.0%	5.0%	10.0%	5.0%	15.0%	15.0%	10.0%	1
	% within Rater_2	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	10
	% of Total	5.0%	25.0%	5.0%	10.0%	5.0%	15.0%	15.0%	10.0%	1

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.093	.100	1.186	.236
N of Valid Cases		20			

[a] Not assuming the null hypothesis.

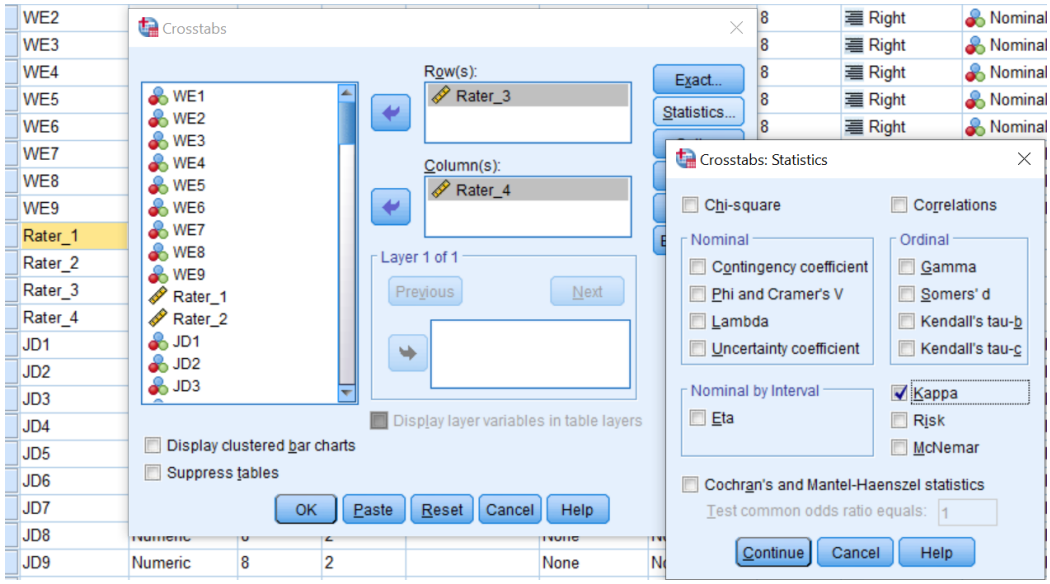
[b] Using the asymptotic standard error assuming the null hypothesis.

SPSS Output Viewer Window 5: Inter-Rater Reliability

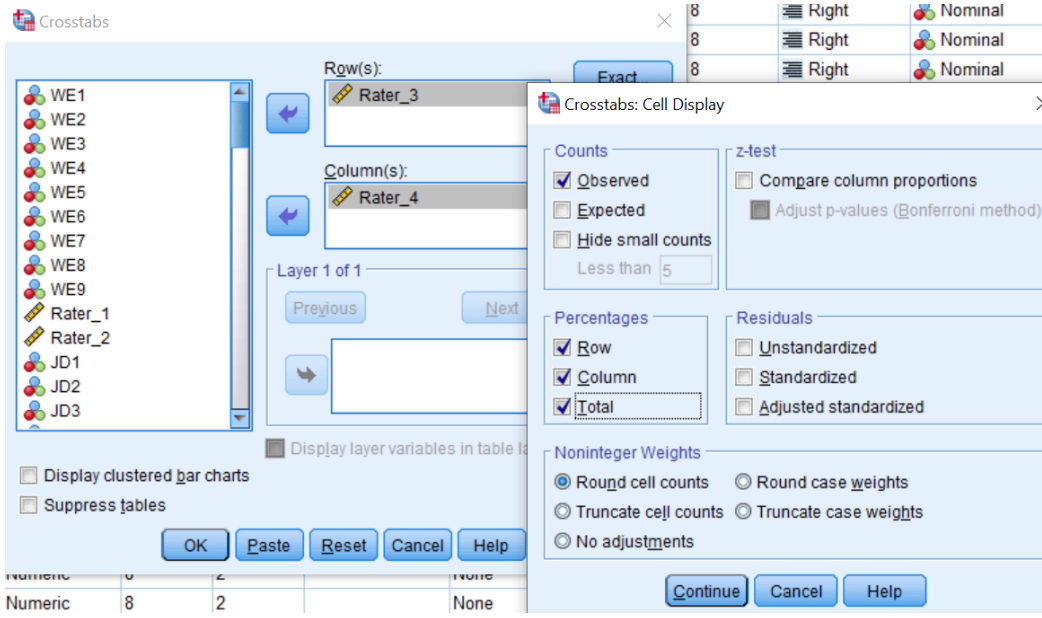
The Kappa scores range from -1 to 1, where 0 represents agreement by chance and 1 represents 100% agreement between screeners (Table 1). Achieving high inter-rater reliability is crucial for ensuring the validity and generalisability of research findings or evaluation results.

Table 1. Kappa score range and interpretation of agreement

Values	**Interpretation**
Smaller than 0.00	Poor Agreement
0.00 to 0.20	Slight Agreement
0.21 to 0.40	Fair Agreement
0.41 to 0.60	Moderate Agreement
0.61 to 0.80	Substantial Agreement
0.81 to 1.00	Almost Perfect Agreement



Step 5. Inter-Rater Reliability (Row and Column; Statistics: Kappa)



Step 6. Inter-Rater Reliability (Counts: Observed> Percentages: Row, Column, Total)

The comparison between Rater_3 and Rater_4 yielded a higher Kappa value of 0.167, which also falls within the range of slight agreement. Importantly, this result was statistically significant (p = 0.019), indicating that the level of agreement between Rater_3 and Rater_4 is unlikely to have occurred by chance.

		% of Total	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.0%	5.0%
	9.00	Count	0	0	0	0	0	0	1	0
		% within Rater_3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
		% within Rater_4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	20.0%	0.0%
		% of Total	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.0%	0.0%
	10.00	Count	0	0	0	0	0	0	1	1
		% within Rater_3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%	50.0%
		% within Rater_4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	20.0%	33.3%
		% of Total	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.0%	5.0%
Total		Count	1	2	3	2	3	1	5	3
		% within Rater_3	5.0%	10.0%	15.0%	10.0%	15.0%	5.0%	25.0%	15.0%
		% within Rater_4	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	5.0%	10.0%	15.0%	10.0%	15.0%	5.0%	25.0%	15.0%

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.167	.097	2.353	.019
N of Valid Cases		20			

[a] Not assuming the null hypothesis.

[b] Using the asymptotic standard error assuming the null hypothesis.

SPSS Output Viewer Window 6: Inter-Rater Reliability

Overall, while both pairs of raters demonstrated only slight agreement, the findings highlight that Rater_3 and Rater_4 were more consistent in their ratings than Rater_1 and Rater_2.

Types of Validity in Quantitative Research

Establishing validity is essential in quantitative research, as it defines how accurately an instrument measures the theoretical construct it is designed to assess. Validity ensures that the conclusions drawn from questionnaire data are both meaningful and reflective of reality. Three main types of validity are typically addressed in quantitative research: content validity, construct validity, and criterion validity. Each plays a unique role in reinforcing the reliability of measurement instruments.

Content Validity

Content validity refers to how well a measurement instrument represents all aspects of the construct being studied. This type of validity ensures that the items in a questionnaire fully capture the scope of the domain they are meant to assess (Moreira dos Santos et al., 2024). For example, a questionnaire aimed at measuring employee work engagement should include items that cover different dimensions, such as vigor, dedication, and absorption.

Establishing content validity usually involves a thorough literature review followed by an evaluation from subject matter experts. These experts assess whether the questionnaire items are appropriate, relevant, and adequately represent the construct being measured (Szakály et al., 2024). Content validity is considered a qualitative judgment (Ayanwale et al., 2024) and does not depend on statistical methods. As Choi and Pak (2005) note, using expert

panels to review items for clarity and comprehensiveness is a common practice to strengthen content validity in fields like the pure sciences, business management, and social science research.

The following steps outline how to assess content validity:

Step 1: Expert Evaluation of Item Relevance

The initial phase involved collecting evaluations from a panel of domain experts deemed most qualified to assess the instrument's content. In this context, Experts were selected based on their expertise and direct experience with the subject matter under investigation (Alsulami et al., 2024). Each panelist was asked to evaluate individual items in the instrument by categorizing them as either "essential," "useful but not essential," or "not necessary" for representing the construct. This classification approach is grounded in Lawshe's method for quantifying expert consensus. Items receiving greater consensus as "essential" are considered to demonstrate stronger content validity.

Step 2: Computation of the Content Validity Ratio (CVR)

Following expert evaluation, the CVR was calculated for each item using the formula:

$$CVR = \frac{n_e - (N/2)}{N/2}$$

Here, n_e = number of experts indicating an item as "essential," and N = is the total number of experts

The resulting CVR values range from −1 to +1, where values greater than 0 indicate that more than half of the experts rated the item as essential. For example, if four out of five experts deemed an item essential, the CVR would be (see Table 2):

$$CVR = \frac{4 - (5/2)}{5/2} = 0.6$$

It is critical to interpret these values against established minimum thresholds (critical values), which vary by panel size. Values below the critical CVR suggest inadequate consensus and thus limited content validity for the item.

Table 2. Content Validity Ratio

Question	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	CVR
1	✔	✔	✔	✔		0.60
2	✔	✔			✔	0.20
3	✔		✔			-0.20
4	✔	✔		✔		0.20
5			✔		✔	-0.20
6	✔	✔	✔	✔		0.60
7	✔	✔	✔	✔	✔	1.00

Step 3: Calculation of the Content Validity Index (CVI)

To evaluate the overall content validity of the instrument, the CVI was computed as the average CVR across all items. This composite index provides a global assessment of content validity, with values closer to 1 indicating stronger agreement among experts. For instance, given seven items with respective CVRs of 0.6, 0.2, −0.2, 0.2, −0.2, 0.6, and 1.0, the CVI would be:

$$CVI = \frac{0.6 + 0.2 - 0.2 + 0.2 - 0.2 + 0.6 + 1.0}{7} \approx 0.31$$

Table 3. Content Validity Index

# of Panelists	Critical Value
5	0.99
6	0.99
7	0.99
8	0.75
9	0.78
10	0.62
11	0.59
12	0.56
20	0.42
30	0.33
40	0.29

A CVI of 0.31 indicates a low level of agreement among experts regarding the essentiality of the items within the instrument, suggesting poor overall content validity. Since the CVI is calculated as the average of the CVRs of all items, a value of 0.31 reflects that most items were not consistently rated as "essential" by the panel of subject matter experts. This implies that the instrument may not sufficiently capture the intended construct and therefore lacks theoretical soundness and practical applicability. A CVI of 0.31 falls significantly below the commonly accepted threshold of 0.80 or the critical value (Table 3) associated with the number of raters (e.g., 0.99 for five experts), indicating the need for substantial revision. Items with low CVR scores should be refined, clarified, or replaced based on expert feedback to enhance the overall CVI and ensure the instrument's relevance and validity.

Construct Validity

Construct validity assesses whether an instrument genuinely measures the theoretical construct it is designed to capture. This form of validity is particularly important when working with abstract concepts, such as employee work engagement. Construct validity is typically divided into two subtypes: convergent validity and discriminant validity (Monteiro & McConnell, 2023).

1. **Convergent validity** assesses the degree to which two measures of constructs that theoretically should be related are actually related (Chou et al., 2024, Favieri et al., 2024).
2. **Discriminant validity** evaluates whether concepts or measurements that are supposed to be unrelated are, in fact, distinct (Chou et al., 2024).

Statistical techniques like exploratory and confirmatory factor analysis are commonly used to test construct validity. These methods analyze the underlying structure of the questionnaire to ensure that the items are correctly aligned with the intended factors. Tavakol and Dennick (2011) highlight the importance of factor analysis as an effective tool for establishing construct validity, especially in fields like psychology and education.

Table 4. Construct Validity

Construct	Convergent validity	Discriminant validity			
	Average variance extracted (AVE)	Heterotrait-Monotrait ratio (HTMT)			
		Y1	Y2	Y3	Y4
Job demands (Y1)	0.739				
Job resources (Y2)	0.682	0.635			
Personal demands (Y3)	0.73	0.783	0.362		
Personal resources (Y4)	0.841	0.741	0.736	0.537	
Work engagement (Y5)	0.538	0.362	0.273	0.825	0.461

The analysis of construct validity (Table 4), focusing on convergent and discriminant validity (HTMT), reveals meaningful results (Nigatu et al., 2024; Đajić et al., 2024). For convergent validity, the Average Variance Extracted (AVE) values for each construct are all above the acceptable threshold of 0.50 (Hossan et al., 2020), indicating that the constructs, such as Job Demands (Y1), Job Resources (Y2), Personal Demands (Y3), Personal Resources (Y4), and Work Engagement (Y5), effectively capture variance in their respective indicators. Regarding discriminant validity, the Heterotrait-Monotrait ratio (HTMT) values remain below the typical thresholds of 0.85 or 0.90 for most of the construct pairs, suggesting that the constructs are sufficiently distinct from one another (Hossan et al., 2020). Specifically, the HTMT values between constructs such as Job Demands (Y1) and Job Resources (Y2) (0.635), or Work Engagement (Y5) and Personal Demands (Y3) (0.461), all fall within acceptable ranges, further supporting the discriminant validity of the model. Overall, the analysis demonstrates that the constructs exhibit both strong convergent and discriminant validity, reinforcing the robustness of the measurement model.

Criterion Validity

Criterion validity evaluates how well a measure correlates with an outcome or an external criterion (Mukred et al., 2024). It reflects the ability of a questionnaire to predict a variable that is theoretically related (Song et al., 2023). Criterion validity is typically divided into two types:

1. **Concurrent validity** refers to the extent to which the results of a particular test correspond to those of a previously established measure administered at the same time (Fu et al., 2024, Jung et al., 2024).
2. **Predictive validity** evaluates how well the instrument forecasts future outcomes (Yusoff et al., 2024).

For example, a newly developed employee work engagement scale might be validated by comparing its results with those obtained from an established and widely used engagement instrument, demonstrating concurrent validity. Similarly, predictive validity can be assessed by examining how well the engagement scores forecast key organizational outcomes such as job performance, employee retention, or absenteeism over time. Choi and Pak (2005) emphasize that establishing criterion validity is particularly important in applied fields, including organizational and human resource research, where the ability of a measure to accurately predict real-world outcomes can directly influence management strategies and workplace interventions. While criterion validity involves rigorous statistical testing, its success largely depends on the availability and appropriateness of established benchmarks for comparison. Together, concurrent and predictive validity provide a thorough framework for assessing the accuracy, relevance, and practical utility of engagement questionnaires in quantitative research. Each type plays a crucial role in developing reliable and scientifically robust measurement tools.

Table 5. Concurrent Validity Assessment (Inter-Item Pearson Correlation Matrix for Work Engagement)

	WE1	WE2	WE3	WE4	WE5	WE6	WE7	WE8	WE9
WE1	1	.201**	.348**	.295**	.155*	-.112	.189**	.412**	.329**
WE2	.201**	1	.651**	.150*	.374**	.019	.365**	.454**	.230**

WE3	.348**	.651**	1	.398**	.231**	.006	.425**	.610**	.502**
WE4	.295**	.150*	.398**	1	.077	.197**	.381**	.634**	.423**
WE5	.155*	.374**	.231**	.077	1	-.148*	-.014	.243**	.257**
WE6	-.112	.019	.006	.197**	-.148*	1	.017	.169*	.057
WE7	.189**	.365**	.425**	.381**	-.014	.017	1	.265**	.190**
WE8	.412**	.454**	.610**	.634**	.243**	.169*	.265**	1	.503**
WE9	.329**	.230**	.502**	.423**	.257**	.057	.190**	.503**	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 5 presents the Pearson correlation matrix assessing the inter-item relationships among the nine items (WE1 to WE9) of the Work Engagement scale. The majority of item pairs show positive and statistically significant correlations at the 0.01 or 0.05 levels, indicating that the items are measuring a common underlying construct. For example, WE3 is strongly correlated with WE8 ($r = .610$, $p < .01$) and WE9 ($r = .502$, $p < .01$), while WE8 also shows strong positive correlations with multiple items, including WE4 ($r = .634$, $p < .01$). These strong associations support the scale's concurrent validity by demonstrating that the items consistently reflect the same conceptual dimension of work engagement. However, item WE6 exhibits weaker and sometimes negative correlations with other items, such as with WE1 ($r = −.112$) and WE5 ($r = −.148$*, $p < .05$), which may indicate that it is not as well aligned with the construct and could be reviewed for potential revision or removal. Overall, the results provide empirical support for the concurrent validity of the work engagement scale based on internal item relationships.

Table 6. Predictive validity

Construct	Q^2 predict	RMSE	MAE
Employee Work Engagement	0.771	0.483	0.384

Table 6 presents the predictive validity results generated using the PLS-Predict procedure in SmartPLS 4. The Q^2_predict value for Employee Work Engagement is 0.771, which indicates strong predictive accuracy (Ramayah et al., 2018). In addition, the RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) values are reasonably low, supporting the model's out-of-sample predictive validity. According to Shmueli et al. (2019), a model demonstrates predictive validity when its prediction errors are lower than those of a benchmark model, such as linear regression.

Validity and Reliability in Qualitative Research

In qualitative research, validity and reliability are understood differently than in quantitative research. Rather than emphasizing measurement precision or replicability, qualitative researchers focus on ensuring the trustworthiness of their findings. Lincoln and Guba's (1985) framework offers a valuable structure for assessing the credibility, transferability, dependability, and confirmability of qualitative research by interview, observations, discussions, and records (Chin et al., 2024).

Trustworthiness Framework

Lincoln and Guba's (1985) framework outlines four key criteria that serve as qualitative counterparts to validity and reliability:

1. **Credibility** refers to the truthfulness or believability of the research findings, ensuring that the research accurately represents the phenomena being studied. This is similar to internal validity in quantitative research (Bahrami & Nasiri, 2024).

2. **Transferability** relates to how well research findings can be generalized or applied to other contexts or settings, which mirrors external validity in quantitative research (Shahzeydi et al., 2025).
3. **Dependability** addresses the stability of the findings over time, much like reliability in quantitative studies. It involves providing a clear audit trail that allows others to track the research process and replicate the study (Bahrami & Nasiri, 2024).
4. **Confirmability** refers to how much the findings are shaped by the participants' perspectives, rather than by the researcher's biases or preconceptions. This is akin to objectivity in quantitative research (Bahrami & Nasiri, 2024).

Table 7. Measuring Validity and Reliability in Qualitative Research Using the Trustworthiness Framework

Criteria	Purpose	Strategies to Enhance	Tools / Methods	Evidence in Report
Credibility	Internal Validity	- Triangulation (data, investigator, theory, method)	NVivo, cross-source coding	Description of triangulated sources or analysts Participant quotes
		- Member Checking	Feedback summaries, transcripts	Verbatim responses from participants validating findings
		- Peer Debriefing	Field notes, peer review documentation	Notes or memos from discussions with colleagues
		- Prolonged Engagement & Persistent Observation	Field logs, observation records	Evidence of sustained involvement in the field
Transferability	External Validity	- Thick Description of context, participants, and setting	NVivo memoing, detailed narrative reporting	Demographics, environmental context, sample characteristics

		- Purposive Sampling	Sample matrix, NVivo attribute tables	Rationale for selection, linked to research objectives
Dependability	Reliability	- Audit Trail	Research journal, coded audit files	Transparent record of research steps and coding process
		- Code-Recode Strategy	Manual coding comparison or NVivo matrix	Coding reliability over time with comparison of coding outputs
		- External Peer Examination	Coding validation logs	Peer-reviewed coding consistency, discussion outcomes
Confirmability	Objectivity	- Reflexivity (positionality & bias awareness)	Reflexive journal, memos in NVivo	Reflections on personal biases and their potential impact
		- Audit Trail & Data-Findings Link	Codebook, direct quotes in findings	Clear linkage between codes, themes, and raw data
		- Triangulation (again supports confirmability)	Coding matrices, source comparison in NVivo	Data consistency across sources

Enhancing Credibility

Enhancing credibility is crucial for establishing trustworthiness in qualitative research. Several strategies can be used to ensure that the findings are credible and accurately reflect participants' perspectives:

- **Member Checking**: One of the most widely used methods to boost credibility is member checking, where the researcher presents the findings to participants to confirm that their views are represented correctly. This process allows participants to validate or clarify their responses, improving the accuracy and trustworthiness of the results (Golafshani, 2003).

For example, in a study exploring employee work engagement, preliminary themes such as "sustained energy at work" (vigor), "emotional investment in

organizational goals" (dedication), and "deep focus during tasks" (absorption) were shared with participants in follow-up interviews. Participants confirmed that these themes accurately reflected their experiences, and in some cases, provided further clarification on how their sense of dedication varied during organizational changes.

- **Prolonged Engagement**: Prolonged engagement refers to spending enough time in the field to build rapport with participants and gain a deep understanding of the research context. This strategy helps the researcher develop a more comprehensive view of the phenomenon and reduces the risk of misinterpretation due to superficial involvement (Golafshani, 2003).

The researcher might, for example, spend a long period within the HR departments of several firms, participating in informal conversations, team meetings, and day-to-day observations. This immersion helps the researcher to capture subtle aspects of engagement, such as how morning briefings influenced employees' vigor, rather than relying solely on self-reported experiences.

- **Triangulation**: Triangulation involves using multiple data sources, methods, or researchers to cross-check and validate findings. This approach enhances credibility by offering a broader, more diverse perspective on the research topic and helps minimize biases or limitations that may arise from relying on a single source or method.

For instance, data are collected through semi-structured interviews, daily reflection diaries from employees, and internal HR performance reports. The convergence of insights from these sources strengthened the credibility of

themes, such as employees feeling most absorbed during task-focused work environments and least engaged during mandatory administrative duties.

Ensuring Dependability

Dependability refers to the consistency and reliability of both the research process and its findings over time (Bahrami & Nasiri, 2024). To ensure dependability, researchers must maintain transparency by thoroughly documenting each step of the research process. This includes providing a detailed audit trail that outlines the decisions made throughout the study, such as how data were collected, analysed, and interpreted. An audit trail enables others to trace the research steps, ensuring that the findings are not arbitrary and can be replicated. In qualitative research, this level of documentation helps demonstrate the stability of the findings and strengthens their dependability. Additionally, researchers may use peer debriefing, where an external reviewer examines the research process and findings to confirm that the methodology is rigorous and consistent.

For example, the study normally uses a clear protocol for interview coding, with codebooks updated iteratively as new data emerge. For instance, any code (such as "loss of vigor") might be added after repeated references to burnout during discussions of policy changes. An audit trail is maintained, documenting why such changes are made and how they influence thematic development. Peer debriefing further enhances dependability; an external qualitative researcher periodically reviews transcripts and coding decisions to ensure interpretations are grounded in the data and not shaped by the lead researcher’s expectations.

Confirmability and Transferability

Confirmability ensures that research findings are shaped by the data rather than the researcher's biases or assumptions (Shahzeydi et al., 2025). This can be achieved by techniques such as keeping a reflexive journal, where the researcher notes personal thoughts and reflections throughout the study. The journal helps the researcher stay aware of potential biases, ensuring that they do not inadvertently influence the findings. In addition to maintaining an audit trail, confirmability can be strengthened by involving external auditors who review whether the findings are supported by the data and free from researcher bias. This process enhances the credibility and objectivity of the research.

The researcher, for instance, maintains a reflexive journal to track personal assumptions, such as expecting that older employees would show less engagement, and how these assumptions are challenged during interviews. The journal entries, along with a detailed audit trail and external auditing of data interpretations, help confirm that the findings are data-driven.

Transferability refers to the extent to which research findings can be applied to other contexts (Shahzeydi et al., 2025). Although qualitative research does not aim for broad generalizability, transferability highlights the importance of providing enough context and detail for others to assess whether the findings are relevant to their own settings. The use of thick description (rich, detailed accounts of the research context, participants, and settings) is crucial in ensuring transferability. By offering thick description, researchers enable others to make informed judgments about how the findings may apply to different environments or populations.

Employees, for example, in a start-up describe engagement in terms of flexibility and creativity, while those in a corporate bank link engagement to structured recognition programs. These detailed contexts allow other researchers and practitioners to judge whether the findings are applicable to their own organizations.

Strengthening Qualitative Rigor through Member Checking and Inter-Coder Reliability

To further enhance the rigor and trustworthiness of qualitative research, incorporating member checking and inter-coder reliability is essential. Member checking allows participants to review and validate the accuracy of the transcribed data or emerging themes, ensuring that their perspectives have been authentically interpreted. This process can occur during data collection, after transcription, or at the analysis stage, and it serves as a powerful tool for increasing credibility and transparency. In parallel, inter-coder reliability involves multiple researchers independently coding a subset of data, then comparing their results to assess consistency and reduce subjective bias. Discrepancies are discussed and resolved collaboratively, refining the coding framework and strengthening the dependability of the analysis. Utilizing qualitative data analysis software such as NVivo can facilitate this process by quantifying coder agreement and visualizing coding overlaps. These strategies contribute significantly to the rigor, accountability, and reproducibility of qualitative research.

Member checking strengthens validity (credibility), while inter-coder reliability enhances reliability (dependability) in qualitative research. For example, in a study exploring how employees engage with their work, researchers carried out a series of in-depth interviews to understand personal experiences of vigor, dedication, and absorption. To ensure the findings truly reflected participants' perspectives, the team used member checking by sharing an early summary of emerging themes with the interviewees. Some participants offered thoughtful clarifications, for instance, suggesting that what the researchers had initially described as "task immersion" was more in line with how they perceived absorption, and that "commitment" didn't quite capture the emotional drive they associated with dedication. These insights prompted the researchers to refine their thematic labels, improving the study's credibility. At the same time, to maintain analytical consistency, two researchers independently coded several transcripts using the same codebook. When they encountered differences, like one tagging a section as "vigor" while the other saw it as "task engagement", they discussed their reasoning, adjusted the code definitions, and used NVivo to track and resolve overlaps. This process of inter-coder reliability helped ensure the findings were not only coherent but also dependable across researchers.

Challenges and Ethical Considerations

Challenges

The design and implementation of questionnaires present several challenges that need to be carefully addressed. One major challenge is cultural sensitivity. Questions that are valid in one culture may be misinterpreted or

irrelevant in another, so it's important to ensure the questionnaire is culturally appropriate for all respondents. This may involve translating the questionnaire and conducting pre-tests in different cultural contexts to ensure clarity and relevance.

Another challenge is respondent fatigue, especially in lengthy surveys. If participants find the questionnaire too long or repetitive, they may lose interest or rush through their responses, which can compromise the quality of the data. To address this, researchers can design shorter questionnaires or break long surveys into smaller, more manageable sections.

Ambiguity in questions is also a common issue. Vague or unclear questions can lead to misinterpretation, which undermines the reliability of the data. To avoid this, it is essential to pre-test the questionnaire and ensure that questions are specific, clear, and unambiguous, so they capture the intended responses accurately.

Ethical Considerations

Ethical considerations are crucial in any research involving human participants. One of the most important requirements is informed consent, which ensures that participants are fully aware of the study's purpose, procedures, potential risks, and benefits before agreeing to participate. This transparency allows participants to make an informed decision about their involvement.

Anonymity is another key ethical concern. Researchers must take steps to protect participants' identities throughout the study and when disseminating results. This can be achieved by assigning codes or pseudonyms to

participants and ensuring that no personally identifiable information is included in the final analysis.

Minimizing participant burden is also an important ethical consideration. Researchers should design questionnaires that are not overly time-consuming or mentally taxing, as excessive demands can lead to disengagement or distress. Providing clear instructions and offering appropriate incentives can help reduce participant burden while upholding ethical standards.

Recommendations and Best Practices

Use Established Scales with Documented Validity and Reliability When Possible

One of the key practices for ensuring the validity and reliability of questionnaires is to use established scales that have already been rigorously tested. These scales come with documented evidence of their psychometric properties, confirming that they are both reliable (providing consistent results over time) and valid (measuring what they are intended to measure). By using established scales, researchers can minimize the need for extensive validation processes, as their validity and reliability have already been demonstrated in previous studies. This approach not only saves time but also enhances the credibility of the research.

Conduct Pilot Studies to Refine Instruments

A pilot study is an essential step before launching a full-scale survey or questionnaire. It allows researchers to test their instruments on a small

group of participants, helping to identify any issues with clarity, comprehension, or potential ambiguities in the questions. A pilot study can reveal whether certain questions need to be revised, if the length of the scale is appropriate, or if there are any technical problems with administering the survey. The feedback gathered from a pilot study enables researchers to make adjustments before beginning the full data collection process, ultimately improving the validity and reliability of the final instrument.

Engage Experts for Content Validation

Content validation is the process of ensuring that the questions in a questionnaire accurately represent the construct being measured. To achieve this, researchers should involve subject matter experts who can assess the relevance and appropriateness of the content. These experts offer valuable feedback on whether the items effectively capture the concept being studied. Their input also helps eliminate irrelevant items and ensures that the survey comprehensively measures the intended constructs, ultimately enhancing the construct validity of the questionnaire.

Exploratory Factor Analysis (EFA) is primarily used for construct validity: a part of questionnaire validity, but not directly for assessing questionnaire reliability

EFA is mainly used to check construct validity, which means it helps see if the questionnaire items that are supposed to measure the same idea actually group together under one factor. This is useful to confirm that the structure of the questionnaire makes sense. However, questionnaire validity is a

broader concept that includes content validity (whether the questions fully cover the topic), construct validity (supported by EFA), and criterion validity (whether the questionnaire matches real-world outcomes). EFA only helps with construct validity, not the other types. On the other hand, questionnaire reliability means how consistent the responses are across items, and it is usually measured with Cronbach's alpha, not EFA. So, while EFA is helpful for checking part of the questionnaire's quality, other tools are needed to fully test its validity and reliability.

Apply Mixed-Method Strategies to Enrich Data Quality

A mixed-method approach, which combines both quantitative and qualitative methods, can greatly enrich the data collection process. Quantitative methods provide numerical data that can be analyzed statistically, while qualitative methods offer deeper insights into the reasons, thoughts, and motivations behind respondents' behaviors. By blending these two approaches, researchers can cross-validate their findings, leading to a more comprehensive and nuanced understanding of the research topic. For example, qualitative interviews can help explain patterns discovered during the quantitative phase, thereby strengthening the overall validity of the research conclusions.

Maintain Transparency and Reflexivity in Qualitative Designs

In qualitative research, transparency and reflexivity are crucial for ensuring the trustworthiness of the findings. Transparency involves providing a clear explanation of the research process, including how data were collected, analyzed, and interpreted. Reflexivity, on the other hand, requires researchers to reflect critically on their own role in the study, recognizing any biases or

assumptions they may bring to the research. This reflection helps ensure that the conclusions drawn are shaped by the participants' experiences rather than the researcher's personal perspectives, ultimately enhancing the validity and reliability of the study.

Conclusion

Validity and reliability are fundamental pillars of research design, and their significance cannot be overstated. While these concepts are assessed differently in quantitative and qualitative research, they both aim to ensure that the findings are accurate, trustworthy, and meaningful. In quantitative research, reliability concerns the consistency of results, while validity ensures that the instrument truly measures what it is intended to. In qualitative research, reliability is often linked to the consistency of data interpretation, and validity focuses on the authenticity and credibility of the findings. To achieve strong research outcomes, it is essential to design questionnaires that undergo thorough testing for both validity and reliability. A well-designed questionnaire, built on established scales, piloted for refinement, and validated through expert feedback and mixed-method approaches, boosts the credibility of the research. By upholding transparency and reflexivity, researchers can ensure their qualitative designs meet the same rigorous standards, contributing to the overall validity and reliability of the study. Ultimately, evaluating both validity and reliability strengthens the research's credibility and maximizes its impact on the field.

References

1. Adeniran, A. O. (2025). Understanding Cronbach's Alpha in Social and Management Studies. Current Science Research Bulletin, 2(02), 11-16.
2. Adeyemi, B. B. (2024). Reliability and validity in quantitative research. In Applied Linguistics and Language Education Research Methods: Fundamentals and Innovations (pp. 86-102). IGI Global.
3. Ahmed, S. K. (2024). The pillars of trustworthiness in qualitative research. Journal of Medicine, Surgery, and Public Health, 2, 100051.
4. Anderson, R., Taylor, S., Taylor, T., & Virues-Ortega, J. (2022). Thematic and textual analysis methods for developing social validity questionnaires in applied behavior analysis. Behavioral Interventions, 37(3), 732-753.
5. Bujang, M. A., Omar, E. D., Foo, D. H. P., & Hon, Y. K. (2024). Sample size determination for conducting a pilot study to assess reliability of a questionnaire. Restorative dentistry & endodontics, 49(1).
6. Chetwynd, E. (2022). Critical analysis of reliability and validity in literature reviews. Journal of Human Lactation, 38(3), 392-396.
7. Choi, B. C., & Pak, A. W. (2005). A catalog of biases in questionnaires. Preventing Chronic Disease—Public Health Research, Practice and Policy, 2(1).
8. Cole, R. (2024). Inter-rater reliability methods in qualitative case study research. Sociological Methods & Research, 53(4), 1944-1975.
9. Gelino, B. W., Schlitzer, R. D., Reed, D. D., & Strickland, J. C. (2024). A systematic review and meta-analysis of test–retest reliability and stability of delay and probability discounting. Journal of the experimental analysis of behavior, 121(3), 358-372.
10. Golafshani, N. (2003). Understanding reliability and validity in qualitative research. The qualitative report, 8(4), 597-607.
11. Hossan, D., Aktar, A., & Zhang, Q. (2020). A study on partial least squares structural equation modeling (PLS-SEM) as emerging tool in action research. LC International Journal of STEM, 1(4), 130-146.
12. Lim, W. M. (2024). What is qualitative research? An overview and guidelines. Australasian Marketing Journal, 14413582241264619.
13. Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills, CA: Sage.
14. Lopez, A., Caffò, A. O., Tinella, L., & Bosco, A. (2023). The four factors of mind wandering questionnaire: content, construct, and clinical validity. Assessment, 30(2), 433-447.

15. Millar, M. M., Taft, T., & Weir, C. R. (2022). Clinical trial recruitment in primary care: exploratory factor analysis of a questionnaire to measure barriers and facilitators to primary care providers' involvement. BMC primary care, 23(1), 311.
16. Monteiro, S., & McConnell, M. M. (2023). Evaluating the construct validity of competencies: a retrospective analysis. Medical Science Educator, 33(3), 729-736.
17. Nurhidayati, S., & Sukri, A. (2025). Validation of Students' Green Behavior Instrument Based on Local Potential Using Structural Equation Modeling With Smart Partial Least Squares. European Journal of Educational Research, 14(1).
18. Ramayah, T., Cheah, J., Chuah, F., Ting, H., & Memon, M. A. (2018). Partial least squares structural equation modeling (PLS-SEM) using smartPLS 3.0. An Updated Guide and Practical Guide to Statistical Analysis.
19. Shelley, L., Jones, C., Pearson, E., Richards, C., Crawford, H., Paricos, A., ... & Waite, J. (2024). Measurement tools for behaviours that challenge and behavioural function in people with intellectual disability: A systematic review and meta-analysis of internal consistency, inter-rater reliability, and test-retest reliability. Clinical Psychology Review, 102434.
20. Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J. H., Ting, H., Vaithilingam, S., & Ringle, C. M. (2019). Predictive model assessment in PLS-SEM: guidelines for using PLSpredict. European journal of marketing, 53(11), 2322-2347.
21. Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. Assessment, 30(5), 1662-1671.
22. Taherdoost, H. (2022). Designing a questionnaire for a research paper: A comprehensive guide to design and develop an effective questionnaire. Asian Journal of Managerial Science, 11(1), 8-16.
23. 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
24. Tristan-Lopez, A. (2025). Enhancing Cronbach's Alpha Applications through the Harmonised VOR Framework.
25. Tulliez, S., Karantzoulis, S., Marcus, J. C., Casamayor, M., Blanchard, C., Goenjian, H., ... & Keefe, R. S. (2025). Assessing the inter-rater reliability of the Schizophrenia Cognition Rating Scale: a non-interventional quantitative study. Schizophrenia, 11(1), 1-8.
26. Vilagut, G. (2024). Test-retest reliability. In Encyclopedia of quality of life and well-being research (pp. 7180-7184). Cham: Springer International Publishing.

Appendix:

1. Abdu, S. M. M., Gebreal, A., Alsaleem, S. A., Aljohani, M. S., Abdel-Rahman, S., Hussein, M. F., ... & Ghazy, R. M. (2024). Arabic validation and cross-cultural adaptation of climate anxiety scale. Clinical Epidemiology and Global Health, 30, 101835.
2. Alsulami, A., Fairbrass, J., Botelho, T., & Assadinia, S. (2024). Renewable energy and innovation in Saudi Arabia: An exploration of factors affecting consumers' intention to adopt Solar PV. Technological Forecasting and Social Change, 204, 123430.
3. Arslanlı, S. E., Çelebioğlu, A., Çelik, İ., & Uzun, N. B. (2024). Development and Psychometric Testing of the Fathers' Self-Efficacy Scale for Newborn Care. Asian Nursing Research, 18(2), 97-105.
4. Ayanwale, M. A., Molefi, R. R., & Liapeng, S. (2024). Unlocking educational frontiers: Exploring higher educators' adoption of google workspace technology tools for teaching and assessment in Lesotho dynamic landscape. Heliyon, 10(9).
5. Bahrami, M., & Nasiri, A. (2024). Behind the scenes of caregiving in patients with advanced cancer: A qualitative study on family caregivers. Asia-Pacific Journal of Oncology Nursing, 11(1), 100330.
6. Bauer, F., King, D. R., Friesl, M., Schriber, S., & Weng, Q. (2024). Acquisition integration capabilities and organizational design. Long Range Planning, 57(6), 102479.
7. Chin, T., Shi, Y., Arrigo, E., & Palladino, R. (2024). Paradoxical Behavior toward Innovation: Knowledge sharing, knowledge hiding, and career sustainability interactions. European Management Journal.
8. Chou, C. C., Iamtrakul, P., Yoh, K., Miyata, M., & Doi, K. (2024). Determining the role of self-efficacy in sustained behavior change: An empirical study on intention to use community-based electric ride-sharing. Transportation Research Part A: Policy and Practice, 179, 103921.
9. Contreras-Villalobos, T., Escudero, E. B., & Leiva, V. L. (2024). Digging and building: How transformative mixed-methods research contributes to explaining and responding to educational exclusion and school dropout. Methods in Psychology, 11, 100170.
10. Cristache, N., Pricopoaia, O., Năstase, M., Șișu, J. A., Tîrnovanu, A. C., & Matiș, C. (2024). The metaverse, a new frontier for innovative business models. Technological Forecasting and Social Change, 209, 123838.
11. Đajić, M. J., Lalic, D. C., Vujičić, M. D., Stankov, U., Petrovic, M., & Đurić, Ž. (2024). Development and validation of the project manager skills scale (PMSS): An empirical approach. Heliyon, 10(3).

12. Favieri, F., Forte, G., Savastano, M., & Casagrande, M. (2024). Validation of the brief screening of Social Network Addiction Risk. Acta Psychologica, 247, 104323.
13. Fu, X., Shen, A., Zhang, L., Wang, Y., & Lu, Q. (2024). Development and psychometric testing of the lymphedema self-management support scale for breast cancer survivors. Asia-Pacific Journal of Oncology Nursing, 11(6), 100494.
14. Jung, J. Y., & Kim, J. Y. (2024). Validity and Reliability of the Korean Version of the Caregiver Burden Questionnaire for Family Caregivers of Hemodialysis Patients. Asian Nursing Research, 18(5), 509-515.
15. Kang, N. H., Yoo, J. E., & Kim, H. G. (2024). Development and Validation of a Scale of Competency of Digital Age Teaching (SCoDAT) for Preservice Teachers. Innovation and Education, 6(1), 58-77.
16. Kim, Y., & Zailani, S. (2024). Pull and push factors of Koreans' sustainable consumption behaviours from plant-based meat products. Environmental Challenges, 15, 100886.
17. Moreira dos Santos, E. R., Pereira, L. N., Pinto, P., & Boley, B. B. (2024). Development and validation of the new Resident Empowerment through Tourism Scale: RETS 2.0. Tourism Management.
18. Mukred, M., Asma'Mokhtar, U., Hawash, B., AlSalman, H., & Zohaib, M. (2024). The adoption and use of learning analytics tools to improve decision making in higher learning institutions: An extension of technology acceptance model. Heliyon, 10(4).
19. Nguyen, A. H. D., Le, T. T., Dang, T. Q., & Nguyen, L. T. (2024). Understanding metaverse adoption in education: The extended UTAUMT model. Heliyon, 10(19).
20. Nigatu, T., Degoma, A., & Tsegaye, A. (2024). Green practices and economic performance: Mediating role of green innovation in Ethiopian leather, textile, and garment industries—An integrated PLS-SEM analysis. Heliyon, 10(15).
21. Nong, N. M. T., Phuong, N. Q., & Duc-Son, H. (2024). The effect of employee competence and competence–job–fit on business performance through moderating role of social exchange: A study in logistics firms. The Asian Journal of Shipping and Logistics, 40(4), 187-197.
22. Ricci-Bonot, C., Romero, T., Nicol, C., & Mills, D. S. (2024). Mind your language! Lessons from the application of an English published version of a Japanese horse personality instrument to a French population. Applied Animal Behaviour Science, 273, 106237.
23. Sawaneh, I. A., Fan, L., & Sesay, B. (2024). Investigating the influence of residents' attitudes, perceptions of risk, and subjective norms on their willingness to engage in flood prevention efforts in Freetown, Sierra Leone. Nature-Based Solutions, 6, 100143.

24. Shahzeydi, A., Jazi, Z. H., & Babaei, S. (2024). Exploring the perceptions of nursing internship students and their faculty mentors regarding the complexities and hindrances associated with implementing the nursing process within the clinical setting: An in-depth qualitative investigation. Heliyon, 10(11).
25. Strandberg, C., & Ek Styvén, M. (2024). The multidimensionality of place identity: A systematic concept analysis and framework of place-related identity elements. Journal of Environmental Psychology, 95.
26. Szakály, Z., Gáthy, A. B., Kontor, E., Balogh, P., & Czine, P. (2024). Analysis of heterogeneity in consumer attitudes based on the Intuitive Eating Scale-2. International Journal of Gastronomy and Food Science, 38, 101039.
27. Ulum, Ö. G. (2024). Empowering voices: A deep dive into translanguaging perceptions among Turkish high school and university language learners. Heliyon, 10(20).
28. Yukhymenko-Lescroart, M. A., & Voiedilova, O. (2024). Validation of the Ukrainian versions of two life purpose scales: Revised sense of purpose in life scale (SOPS-2U) and specific purpose orientations (SPO-U). MethodsX, 13, 102848.
29. Yusoff, H. M., Ismail, K. I., Ismail, R., Khamis, N. K., Robat, R. M., & Bryce, J. M. (2024). Development and evaluation of a scale to measure nurses' unsafe driving behaviour while commuting. Heliyon, 10(1).

Dalowar Hossan is currently serving as a Personal Research Assistant to Prof. Dr. Bert Wolfs, a founding member, Academic Dean and DBA Program Manager at SBS Swiss Business School. He holds a PhD in Corporate Leadership and an MSc (by research) in Business Economics from the School of Business and Economics, Universiti Putra Malaysia (UPM). Dr. Hossan has published articles in several scientific journals, including those indexed in Scopus and Web of Science, and has presented papers at numerous international academic conferences. He has conducted many international workshops and webinars, particularly focusing on data analysis and research methodology. His research interests include management, work and organisational psychology, employee engagement and performance, leadership, entrepreneurship, artificial intelligence, corporate governance, and sustainability.

Bert Wolfs is a founding member of SBS Swiss Business School and holds the position of the Academic Dean and DBA Program Manager at the institution. His research and teaching focus on entrepreneurship, e-commerce, higher education, and digitalization. He holds advanced degrees from Harvard Graduate School of Education (MDP, 2009), UC Kentucky (Ph.D., 2015), and VSK India (D.Litt,

2014). Prof. Wolfs has international teaching experience in Riga, Moscow, Bengaluru, and Kiev, and industry experience with Master Foods, Mars, and consulting. He is a recipient of the Albert Einstein Peace Award (IAEWP, UN-NGO) and has been recognized by the Belgian-Luxembourg Chamber of Commerce and the Latvian Ministry of Education. His recent publications explore global trade, scientific methodology, emotional intelligence, and strategic resilience. He contributes regularly to academic journals, conferences, and media, and is active as a keynote speaker and advisor.

Noor-E-Medina Suraiya Jesmin is an Assistant Professor at the University of Asia Pacific, Bangladesh. She earned her PhD from Cardiff Metropolitan University, UK, and holds two LLM degrees, including a specialization in Human Rights from Queen Mary University of London, and a general LLM along with an LLB (Hons) from the University of Dhaka. Her research interests include human rights, environmental law, climate change and corporate sustainability. She has authored numerous scholarly articles published in reputed academic journals. In addition to the academic work, Dr. Jesmin is a member of the Bangladesh Bar Council.

Milos Petkovic is the Head of Research at SBS Swiss Business School in Zurich, Switzerland. His research fields include leadership, strategy, and human resources, with the ability to utilize both quantitative and qualitative research methods. Prof. Petkovic has published numerous scientific articles, book chapters, international conference papers, and more. For more than 10 years, he has been teaching at various academic institutions in Europe and the USA.