Bilad Alrafidain Journal of Humanities and Social Science Vol. 8, No. 1 **ISSN: 2788-6026**
Doi: 10.54720/bajhss/2026.080120 **Pages: 22-38**

Tutorial Methods on How to Test the Observation Techniques' Validity and Reliability in Social Science and Applied Research

Lect. Shurooq Talab Jaafar
Department of English, College of Education for Humanities, University of Diyala, 32001 Baqubah, Diyala, Iraq
shurooq.talab@gmail.com

Abstract
This study is presented as a set of guiding methods intended to assist researchers in applying validity and reliability testing procedures systematically. It provides an in-depth examination of observation techniques and procedures used in the social sciences and applied research, with a particular focus on how their validity and reliability can be verified. Unlike many previous studies, this review offers a clear methodological framework for evaluating the accuracy of data extraction, supported by the use of precise and reliable tools and techniques designed to meet the needs of novice researchers. The study emphasizes the fundamental methods required for constructing observation instruments, designing reliability tests, and analyzing and interpreting results, in addition to applying these procedures in real research contexts. The findings indicate that strict adherence to these steps contributes significantly to enhancing the validity and reliability of data derived from observation. This study seeks to bridge the gap between theory and practice by providing accurate and sound observation methods grounded in solid scientific principles within research environments.
Keywords: observation technique protocols, validity of observation techniques, reliability of observation techniques, Lawshe's method (CVR), Cohen's Kappa Index (CKI), Cronbach's alpha (α), Scott's Pi (π), Fleiss' Kappa (κ).

طرق تعليمية حول كيفية اختبار صدق وثبات تقنيات الملاحظة في العلوم الاجتماعية والبحوث التطبيقية

م. شروق طلب جعفر
قسم اللغة الإنجليزية، كلية التربية للعلوم الإنسانية، جامعة ديالى، 32001 بعقوبة، ديالى، العراق
shurooq.talab@gmail.com

الملخص
تقدم هذه الدراسة بوصفها طرق إرشادية تهدف إلى مساعدة الباحثين على تطبيق إجراءات اختبار الصدق والثبات بصورة منهجية. وتتناول بشكل معمق تقنيات وإجراءات الملاحظة المستخدمة في العلوم الاجتماعية والبحوث التطبيقية. كما انها تركز على كيفية التحقق من صدقها وثباتها. وعلى خلاف العديد من الدراسات السابقة، تطرح هذه المراجعة إطارا منهجيا واضحا لتقويم دقة استخراج البيانات، مدعوما باستخدام أدوات وتقنيات دقيقة وموثوقة، ومصممة بما يتلاءم مع احتياجات الباحثين المبتدئين. تركز هذه الدراسة على الطرق الأساسية اللازمة لبناء أدوات الملاحظة، وتصميم اختبارات الثبات، وتحليل النتائج وتفسيرها. إضافة إلى توظيف هذه الإجراءات في السياقات البحثية الواقعية. وتظهر النتائج أن الالتزام الصارم بهذه الخطوات يسهم بشكل ملحوظ في تعزيز صدق وثبات

البيانات المستمدة من الملاحظة. وتسعى هذه الدراسة إلى ردم الفجوة بين الجانب النظري والتطبيق العملي، من خلال إتاحة طرق ملاحظة دقيقة وسليمة قائمة على أسس علمية راسخة داخل البيئات البحثية.
الكلمات المفتاحية: بروتوكولات تقنيات الملاحظة، صدق تقنيات الملاحظة، ثبات تقنيات الملاحظة، طريقة لاوشي (CVR)، معامل كابا لكوهين(CKI) ، معامل ألفا كرونباخ(α) ، معامل باي لسكوت(π) ، معامل كابا لفليس.(κ)

1. Introduction

Observation has long been recognised as a central method of inquiry in social science and applied research, particularly in studies concerned with human behaviour, educational practices, and language use in real-world contexts. Unlike self-report instruments or experimental measures, observation allows researchers to document phenomena as they occur, capturing contextualised actions, interactions, and processes that might otherwise remain inaccessible. For this reason, observation techniques are frequently employed in educational and linguistic research, where understanding naturally occurring practices is essential. However, the methodological strength of observation-based research depends not only on careful data collection but also on the extent to which the procedures used can produce valid and reliable findings [1-4].

The validity and reliability of the observation technique remain critical concerns for scholars and language researchers [5-8]. Furthermore, the observation technique is very important in teaching and learning assessments. It is also one of the most significant research methods in linguistics and social sciences, and at the same time, one of the ultimate complexes. It may be the essential technique in one project or one of the various complementary qualitative research methods [9-16]. Systematically, the observation technique will be carried out with a focus on specific research questions, as in any scientific research method. Thus, from this optimum importance, a plethora of previous and recent studies focus on supporting the validity and reliability of observation techniques [17-18].

As a matter of reliability (consistency across time or observers), Practitioners should follow several protocols. If only one observer is present, the criteria of multiple observers are disregarded. However, consistency over time has become more important in assessing inter-observer agreement. In this case, the best way to test the reliability of the observation technique is through time intervals. Language teacher observes the phenomena under investigation through several intervals, and each interval lasts for three months or more. Inter-rater reliability is employed to ensure data reliability during the observation method by comparing agreement between three categories or time intervals. In other forms of research, consistency across observers (inter-observer agreement) is preferred [19-21]. On the other hand, data validity determines whether a study measures or observes the intended variables or phenomena. The validity of the observation technique data is tested according to two types of validity (internal validity and external validity) by a language teacher [22-24].

Although observation is widely used in the study of social science and applied research, the validity and reliability of much observational research are addressed only superficially, either conceptually or with a brief discussion of methodological procedures as applicable to studies conducted by practitioners and novices. As a result of the absence of systematic guides for using statistical and methodological procedures, ambiguity in how these protocols should be applied remains common, especially in educational & linguistic literatures.

The significance of this study lies in its practical and methodological contribution to social science and applied research. Unlike many previous studies that address the validity and reliability conceptually. This research provides a structural step-by-step tutorial framework supported by established quantitative indices. By translating abstract measurement principles into applicable procedures, the study offers practical value for practitioners, novice researchers, and language educators. Seeking to enhance the accuracy, consistency, and credibility of observation-based research.

2. Research Objectives

This study aims to:

1. To provide a comprehensive tutorial on how to test observation methods in terms of validity and reliability within the context of social science and applied research.
2. To present precise procedures for confirming the validity and reliability of observation technique protocols, providing valuable perceptions for language teaching and learning assessments.
3. To delve into various equations such as Lawshe's method (CVR), Cronbach's alpha (a), Cohen's Kappa Index (CKI), Scott's pi (π), and Fleiss' kappa (k) to test the validity and reliability of observation techniques.
4. To offer practitioners and language researchers a well-organised guide for improving observation technique protocols and obtaining reliable and accurate results in social science and applied research.
5. To emphasise the importance of reliability across time or observers and to discuss the best practices for testing the reliability of observation techniques in various research forms.
6. To examine the validity of observation technique data and ensure that the study reflects or observes what is desired to reflect or observe.

3. Methodological Framework and Literature Review

As practitioners or practitioners in various educational settings, we are responsible for ensuring the validity and reliability of our teaching and assessment methods. Below is a comprehensive literature review of the observation technique's validity and reliability:

3.1. Validity of Observation Technique Protocols Used by Practitioners

However, several practitioners stated that the observation technique is considered internally valid according to three main criteria: 1) if you can exclude different explanations of the results or findings. 2) The results or outcomes are valid over time. 3) The results or outcomes are valid when they are replicated [25-32].

Furthermore, the data validity of the observation technique was tested according to generalizability criteria. The observation technique is said to be externally valid when we can generalize the results of the observation technique beyond the sample under investigation [33-39]. Other studies demonstrated four main kinds of validity in observation protocols [40,41]:

1. Construct validity: comes to answer the question "Does the test measure the concept that it's intended to measure?"
2. Content validity: comes to answer the question "Is the test fully representative of what it aims to measure?"

3. Face validity: comes to answer the question "Does the content of the test appear to be suitable to its aims?"
4. Criterion validity: comes to answer the question "Do the results accurately measure the concrete outcome they are designed to measure? The four types of validity in observation protocols are obtained using different equations:

3.1.1. Lawshe's Method (CVR)

Lawshe's method (1975) is employed to maintain the population data validity ratio (content validity ratio). It decides the number of items of a measuring instrument that epitomises the whole content field [42]. Specialists who are familiar with the content field of the instrumented assessment decide if the items being used are valid or not. A content validity ratio (CVR) is a numerical value representing the device's ratio of validity, which is calculated from expert assessments of content validity [43]. This process is constructed on the question: (Are the content measured by this pilot study essential, useful but not essential, or not necessary to the performance of the construct?) Furthermore, the Specialists responded to this method of measuring and contributed their assessment regarding the scale of Lawshe's 'essential, useful but not essential, or not necessary' [44].

$$CVR = \frac{ne - \frac{N}{2}}{\frac{N}{2}} \qquad (1)$$

From this method, you perceive that (CVR) is a symbol of the population data validity ratio. (ne) is a symbol of the total sum of panellists who replied 'corresponding'. (N) is the total sum of the counselled panellists. In explaining the outcome of this method, concerning Lawshe's Method, the CVR is rated from (-1) to (+1). The closer the ratio is to the positive one (+1), the more vital the CVR becomes. The closer the ratio is to the negative one (-1), the less vital CVR becomes [45-46].

3.1.2. Cohen's Kappa Index (CKI)

Cohen's Kappa Index (CKI) is used to achieve Face validity in a selected measuring instrument [47]. It is used in analysing the data collected from the panellists' responses to the Yes-No dichotomous scale [48]. This type of answer dichotomous scale, does not allow the specialists to be neutral in their response to the question "Do you think this pilot study has face validity?" [49]. In other words, the measuring technique or instrument has 'face validity' if its contents basically look appropriate to the participants under assessment. Face validity assesses the presence of the measuring technique concerning readability, clarity, feasibility, and consistency of the formatting and style [50]. Thus, CKI was used to evaluate the collected data from the specialists' responses, which were on a Yes-No dichotomous scale. According to CKI (Cohen's Kappa Index), the satisfactory proportion of the collected data should not be beneath (0,60%). The counsellor specialists must be between (10-30). Responding on the scale of Yes-No dichotomous should be 'Yes, I think this measurement technique has face validity' or 'No, I do not' [51-53].

3.1.3. Cronbach's alpha (a)

This method is employed to evaluate the internal consistency of observation procedures and outcomes [54]. It illustrates how closely correlated a set of objects is in a group [55]. It is closely connected with a set of questions in a group to measure consistency. A "high" ratio of Cronbach's alpha does not mean

the rate has one dimension [56]. Thus, Cronbach's alpha is not a 'statistical test'; it is a 'coefficient of reliability or consistency'. This coefficient can be expressed as the average intercorrelation among five test questions [57]. Below is the equation for the coefficient of Cronbach's alpha:

$$a = \frac{N.\overline{c}}{\overline{v}\,(N-1)\,\overline{c}} \qquad (2)$$

Where (N) is the number of questions or items, ($\overline{c}$) is the mean value among the questions or items and ($\overline{v}$) is the average variance [58-60].

3.2. Reliability of Observation Technique Protocols Used by Practitioners

The data reliability of the observation technique demonstrates the consistency of the assessment tool. The observational assessment tool produces similar outcomes in similar conditions. Stating that the most well-known method of measuring the reliability of 'the observation technique' is by associating the agreements of two or more observation statements over observers or time [61]. Inter-rater reliability helps preserve collected data reliably in observation techniques by associating the agreement in two- or three-time patterns or intervals over time. Each interval or pattern should have lasted for three months (less or more). In the case of one observer, the best way to test the reliability is by using observation agreements over time intervals. While in other types of research, the best method to test the reliability of observation techniques is to use inter-rater reliability among observers [62-65].

However, there is another way to test the reliability of the observation technique by exposing the procedures of the observation technique to the specialists and then applying their responses to the following equations:

3.2.1. Interrater Reliability (the kappa statistic)

Cohen's Kappa Index (CKI) can be used again, like in face validity, but this time to test the interrater reliability of the observation technique. It differs from face validity in testing the data collected. Face validity is used to test the validity of observation technique procedures [66]. The kappa coefficient is commonly used to examine interrater reliability. It is a crucial method used to test how well the data collection from the observation technique accurately represents the variables under investigation and maintains interrater reliability [67]. Conventionally, CKI was used to measure the percentage agreement by calculating the sum of the agreement marks divided by the overall sum of marks. Similar to a plethora of correlation coefficients, the kappa index ranges from −1 to +1 [68-70]. Cohen's kappa index could be calculated according to the following formula:

$$k = \frac{Pr(a) - Pr(e)}{1 - \Pr(e)} \qquad (3)$$

In this formula, the 'Pr(a)' is the actual 'observed agreement', while 'Pr(e)' is 'the chance agreement'. Furthermore, Cohen precisely deliberated on two types of raters. One is based on the 'chi-square table, and the other is based on 'the Pr(e)', see the following formula:

$$\Pr(\mathrm{e})\ \frac{\frac{cm1x\ rm1}{n} + \frac{cm2x\ rm2}{n}}{n} \qquad (4)$$

In this formula, the 'cm1' signifies 'column 1 marginal', 'cm2' signifies 'column 2 marginal', 'rm1' signifies 'row 1 marginal', 'rm2' signifies 'row 2 marginal', and finally 'n' signifies 'the number of observations (not the number of raters)' [71,72].

3.2.2. Scott's pi

This index is similar to CKI (Cohen's Kappa Index) in that it is based on the observed agreement in the extent to which agreement might be estimated by chance [73]. However, in each index, the predictable agreement (chance agreement) is calculated a little differently. Scott's pi coefficient assumes that experts have a similar distribution of answers, which makes 'Cohen's kappa index' a little more informative [74,75]. Scott's pi equation is similar to 'Cohen's Kappa Index'.

$$\pi = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)} \quad (5)$$

The only difference between the two equations is that 'Pr(e) in Scott's pi is calculated by using squared 'joint proportions' which are 'squared arithmetic means of the marginal proportions' whereas Cohen's uses 'squared geometric means of them' [76-77]. On the other hand, Fleiss' kappa extended Scott's pi coefficient to more than two raters. Scott's pi equation is a coefficient of agreement designed for two annotations of ordinal or nominal scale data, while Fleiss' kappa is a coefficient of agreement designed for more than two annotations [78-80].

3.2.3. Fleiss' kappa

This rarely used method is employed to test the agreement reliability between a constant number of observers who assign 'categorical ratings' to several items or several classifying items. Fleiss' kappa is different from other kappas, like Cohen's kappa, which work only when measuring the agreement with two raters, and intra-rater reliability (for one appraiser versus themselves) [81-83]. Fleiss' kappa {\displaystyle \kappa \,}can be illustrated as,

$$k = \frac{\bar{P} - \bar{P}e}{1 - \bar{P}} \quad (6)$$

Whereas, $1 - \overline{P}$ Provides the agreement degree accessible above chance. $\overline{P} - \overline{P}e$ provides the actual agreement degree achieved over chance. If all observers agree $k = 1$. On the other hand, if the observers are in complete disagreement $k \leq 1$ {\displaystyle \kappa \leq 0} [84-87].

To consolidate the methodological procedures, Figure 1 presents a conceptual framework that illustrates the sequential steps researchers can follow to confirm validity and reliability in observation techniques. Figure 1. Conceptual framework for testing the validity and reliability of observation techniques in social science and applied research.

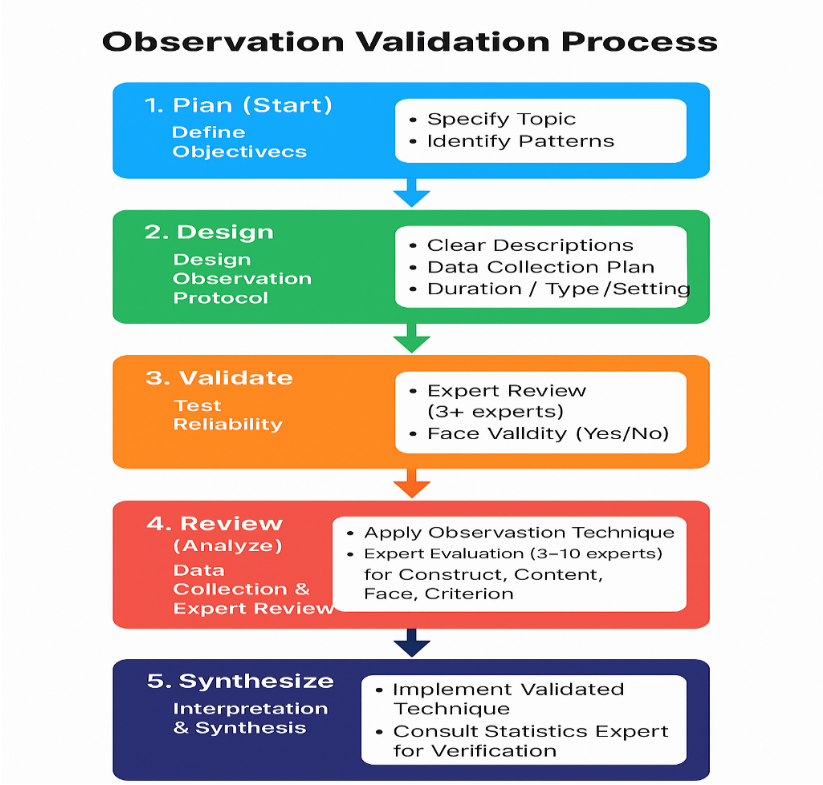

Figure 1: Conceptual framework: Testing Validity and Reliability of observation Techniques.

4. Findings and Discussions

Given the aforementioned equations, a critical discussion of their respective advantages and disadvantages is warranted. Regarding observation technique validation, Lawshe's method (CVR) offers a quantitative measure of content validity that is easy to compute and scalable to large panels of experts. Still, it may lead to a high rate of false positives and assumes that the experts' ratings are independent and uncorrelated. While Cohen's Kappa Index (CKI) offers a measure of inter-rater reliability that corrects for chance agreement and is scalable to two or more raters, it may be affected by the prevalence of the observed phenomenon and may not capture the severity of disagreements. Cronbach's alpha (α) offers a measure of internal consistency among protocol items that is easy to compute, but it may be affected by the number of items and may not capture the complexity of the measured construct.

Concerning the reliability of observation techniques, Scott's π (π) offers a measure of inter-rater reliability that corrects for chance agreement and is scalable to two or more raters, but it may be less reliable than other measures and may not capture the severity of disagreements. On the other hand, Fleiss' kappa (κ) offers a measure of inter-rater reliability that corrects for chance agreement and is scalable to three or more raters, but may be inappropriate for small sample sizes and may not capture the severity of disagreements. To mitigate the disagreement about observation technique protocols' validity and reliability, leading to more robust and trustworthy observation technique results or data, practitioners in schools and other educational settings can use a combination of equations, each with its advantages and disadvantages. However, all the above-mentioned is fully illustrated in the following table:

Table (1): Limitations of Common Equations for Testing Validity and Reliability of Observation Techniques

Method / Index	Limitations
Lawshe's Content Validity Ratio (CVR)	– Outcomes depend heavily on the number and expertise of panelists – Assumes independence of expert judgments – May overestimate content validity when the panel size is small
Cohen's Kappa Index (CKI)	– Sensitive to prevalence and marginal distributions – Basic form limited to two raters only – Does not capture the type or severity of disagreement
Cronbach's Alpha (α)	– Inflated by a large number of items – Not a true indicator of one-dimensionality – May misrepresent reliability when constructs are heterogeneous
Scott's Pi (π)	– Applicable only to two raters – Assumes equal distribution of responses, which is rarely realistic – Less robust than kappa with skewed data
Fleiss' Kappa (κ)	– Requires the same number of raters across all items – Can be unstable with small sample sizes – Interpretation may be difficult for non-statistical audiences

5. Tutorial Modules on How to Test Observation Technique Procedures:

To test the validity and reliability of an observation technique, it is important to follow specific procedures to ensure accurate and dependable outcomes. By following these steps, researchers can establish the trustworthiness of their observations and make informed conclusions based on their findings:

5.1 Module 1: Pre-Observation Procedures

There are several steps to be followed before applying an observation technique procedure to avoid any ambiguity or vague outcomes, which are illustrated by the following:

1- Define your objectives: choose and decide your topic or area of interest to be observed.
2- Use Clear descriptions: the observer(s) should specify the details to be observed, like interactions, behaviours, and the selected environment.
3- Minimising the observer bias: The observer(s) should consider their biases and attempt to observe without preconceived concepts or thoughts.
4- Develop a plan for data collection: Decide the method of recording the observations (notes, checklists, etc.).
5- Determine the Duration of Observation: specify the frequency of the observation technique sessions (either a one-time interval with different observers or different time intervals with one observer).
6- Determine the type of observation (complete observation, participant observation, or non-participant observation).
7- Recording and Documentation: decide the Recording type of the observation technique, like accurate timestamps in one observer's method or a recording format in multiple observers' methods.
8- Determine the observation setting: specify the time, the place, and the participants. For example, the observation technique was applied to 2nd-year students in the English department at the University of Diyala for the academic year 2025-2026.
9- Before starting the observation procedures, send your observation (objectives, descriptions, method of recording, etc.) to three or more experts to ensure the face validity of the observation technique

procedures to avoid wasting your time in obtaining wrong outcomes due to a misconception of one or two criteria. Make sure that everything is fine before starting.

10- To test the face validity of the observation technique, ask the experts on the Yes-No dichotomous scale (Do you think this observation technique's pre-procedures own face validity? If not, please mention the weak point to enhance.). If the experts' responses demonstrate several weak points, rebuild your observation technique and resubmit it. If not, consider their response to Cohen's Kappa index (CKI) to obtain a CKI agreement.

The aforementioned steps are crucial and will guide your next procedures. If an expert questions any observation criteria, it could undermine the observation technique.

5.2 Module 2: Post-Observation Procedures

After determining the face validity of the observation procedures and ensuring that all the criteria of the method are well-constructed. Now, it is time to apply the technique according to these steps:

1- Applying the observation technique procedures with one observer (different time intervals) or multiple observers (with one time interval).

2- Use one type of observation method: complete observation, participant observation, or non-participant observation

3- Obtaining the observation technique by using notes, checklists, etc.

4- Evaluate the findings of observation technique validity by asking 3-10 experts to evaluate (Construct validity, Content validity, Face validity, and Criterion validity). In construct validity, the experts answer the question, "Does the observation technique measure the concept that it is intended to measure? Content validity: comes to answer the question "Is the observation technique fully representative of what it aims to measure? Face validity: comes to answer the question "Does the content of the observation technique appear to be suitable for its aims? In criterion validity, it comes to answering the question, "Do the results accurately measure the concrete outcome they are designed to measure?"

5- Applying the experts' responses to one of these equations (Lawshe's Method, Cohen's Kappa Index, Cronbach's alpha) to obtain the validity of the observation technique findings.

6- To evaluate the reliability of an observation technique's findings, use the inter-rater reliability method. This involves asking 3 to 10 experts to compare the ratings of two or more observation agreements over different observers or time period. Collect the responses of the experts using a Yes-No dichotomous scale (Do you think these observation technique findings have inter-rater reliability?) to gauge their opinion on whether the observation technique findings demonstrate inter-rater reliability.

7- Applying the experts' responses to one of these equations (Cohen's Kappa Index, Scott's pi, Fleiss' Kappa) to obtain the reliability of the observation technique findings.

8- The observer(s) specialised in social science and applied research, but not in mathematics, should seek advice from a(n) expert(s) in this field to ensure the validity and inter-rater reliability of the observation techniques, equations and then include their contribution in the appendices.

6. Worked Example on how to test observation technique validity and reliability

Following a definite methodology, a significant step toward achieving dependable and precise observation results is taken both before and after the data has been collected. This ensures that the study has well-defined objectives, the required observations can be made, any possible biases can be reduced, and the results can be checked against the goals for correctness. To illustrate how these steps work, consider the study on the morphological borrowings of Iraqi undergraduates by Jaafar Dzakiria and Singh (2022). In the subsequent table, every step is presented in parallel with its implementation to illustrate the method of both the reliability and the validity of the observations.

Table (2): Step-by-Step Implementation of Observation Technique Validity and Reliability in the Study of Morphological Borrowings

Step	Procedure	Case Study Application
Pre-Observation	Define Objectives	Identify patterns of morphological borrowing in the college student speaking process
	Clear Descriptions	Focus on the word formations, prefixes, and suffixes in the speech of the college students.
	Minimize Observer Bias	Specify a type of data collection, for example, tape recording, to record data (students' speech) objectively without any assumptions.
	Data Collection Plan	Planned a checklist for systematic recording.
	Observation Duration	Decide the duration of observation of one observer. Three different time intervals by one observer. Each time interval lasted for three months
	Type of Observation	Non-participant observation.
	Recording & Documentation	A checklist was used along with a tape recording.
	Observation Setting	100 second-year students in the Department of English Language for the academic year 2021–2022.
	Expert Review & Face Validity	Procedures reviewed by three experts. Then, Cohen's Kappa was used to confirm agreement.
Post-Observation	Apply Observation Technique	Checklists are applied systematically during the observations.
	Observation Method Consistency	Non-participant observation maintained.
	Obtain Observations	Borrowings are recorded in structured tables. After the observations are completed, the data collection is listed in tables
	Expert Evaluation of Findings	5 experts assessed construct, content, face, and criterion validity.
	Validity Calculations	Applying CVR and CKI.
	Reliability Evaluation	Inter-rater reliability assessed via expert agreement.
	Reliability Calculations	Cohen's Kappa Index measured agreement consistency.
	Statistical Consultation	Statistics expert verified calculations; documented in appendices.

7. Conclusions and Recommendations for Further Studies

7.1 Conclusions

Different methods can be employed to assess the validity and reliability of the observation protocol. Practitioners can use a combination of equations to mitigate disagreement and produce more robust and trustworthy results. The best way to test the validity of observation techniques is by using the combination of any two equations of Lawshe's methods (CVR), Cohen's Kappa Index (CKI), and

Cronbach's alpha (a). The best method to test the reliability of observation techniques is by using the combination of any two equations, like the Kappa statistic, Scott's pi, and Fleiss' Kappa. All these methods or protocols, besides others, come to test the validity and reliability of observation techniques, and this is not the end of the path. Practitioners will keep digging deeper to create other protocols for better observation techniques and assessment to establish their global perspectives and identities. However, for practical application, Table 3 presents a step-by-step checklist that novice researchers can follow to ensure observation validity and reliability in their research:

Table (3): Conclusion and Checklist for Novice Researchers

Step	Purpose	Tips
Define Objectives	Clarify what you aim to observe	Be specific; focus on patterns or behaviors
Describe Observations	Ensure consistent data recording	Specify elements like word forms, prefixes, and suffixes
Minimize Observer Bias	Improve objectivity	Use tape/video recordings; avoid assumptions
Plan Data Collection	Maintain a systematic approach	Prepare checklists or coding sheets
Maintain Observation Consistency	Ensure comparability across sessions	Stick to the same method (e.g., non-participant)
Expert Review	Confirm the validity of procedures and findings	Involve multiple experts for content, face, and construct validity
Assess Reliability & Validity	Quantify agreement and accuracy	Use Cohen's Kappa, Lawshe's CVR, or other statistical measures
Practical Note	Quick guidance for novice researchers	Follow the steps systematically, document carefully, and consult experts whenever possible

Overall, the paper is valuable for practitioners and language researchers looking to improve their observation technique protocols. It is the result of 17 years of working on these equations and testing their strengths and weaknesses. Leading to the present, a well-organised and detailed guide on ensuring the validity and reliability of observation techniques, which is crucial for obtaining reliable and accurate results in social science and applied research.

7.2 Recommendations for Further Studies

Future research may build on the present tutorial framework by empirically examining its applicability across diverse disciplinary contexts, including psychology, sociology, and health-related research. Further studies could investigate the comparative effectiveness of different validity and reliability indices under varying research conditions, such as sample size, number of observers, and observation settings. In addition, future studies could examine the use of generalizability theory alongside traditional reliability indices to determine whether it provides more stable estimates across different observation conditions. Researchers may also consider adopting mixed-method validation designs when quantitative indicators alone do not sufficiently explain inconsistencies in observation outcomes. Such directions would allow for a more precise evaluation of observation protocols and support their methodological refinement in social science and applied research.

References

[1] Gong, Y. F., Lai, C., & Gao, X. A. (2022). Practitioners' identity in teaching intercultural communicative competence. *Language, Culture and Curriculum, 35*(2), 134–150. https://doi.org/10.1080/07908318.2021.1995220

[2] Aronoff, M., & Rees-Miller, J. (Eds.). (2020). *The handbook of linguistics*. John Wiley & Sons.

[3] Ren, J. (2021). Variability and functions of lexical bundles in research articles of applied linguistics and pharmaceutical sciences. *Journal of English for Academic Purposes, 50,* 100968. https://doi.org/10.1016/j.jeap.2021.100968

[4] Thi Ngu, D., Huong, D. T., Huy, D. T. N., Thanh, P. T., & Dongul, E. S. (2021). Language teaching application to English students at master's grade levels on history and macroeconomic-banking management courses in universities and colleges. *Journal of Language and Linguistic Studies, 17*(3), 1457–1468. https://doi.org/10.52462/jlls.89

[5] Jaafar, S. T., Dzakiria, H., & Singh, M. K. S. (2022). Morphological Descriptive Study of Borrowings In Iraqi Undergraduate Students. *International Journal of Psychosocial Rehabilitation*, *26*(1), 163–175. https://doi.org/10.37200/IJPR/V26I1/PR340011

[6] Jaafar, S. T., Dzakiria, H., & Singh, M. K. S. (2023). The Use of English and Arabic Blended Syllables in the Plural Form by Iraqi Undergraduate Students. *Diyala Journal For Human Researches - مجلة ديالى للبحوث الانسانية, 1*(98), 564–582. http://creativecommons.org/licenses/by/4.0/

[7] Jaafar, S. T., & Ugla, R. L. (2022). Tutorial Overview by Using Self-Evaluation and Self-Correction Techniques. *AL-Yarmouk Journal*, *16*(3), 199–224.

[8] Ugla, R. L., & Jaafar, S. T. (2023). Role Models in Language Teaching: A Case of Iraqi EFL University Lecturers. *Bilad Alrafidain Journal of Humanities and Social Science*, *5*(1), 198–207 https://doi.org/10.54720/bajhss/2023.050116

[9] McKuin, B., Zumkehr, A., Ta, J., Bales, R., Viers, J. H., Pathak, T., & Campbell, J. E. (2021). Energy and water co-benefits from covering canals with solar panels. *Nature Sustainability, 4*(7), 609–617. https://doi.org/10.1038/s41893-021-00693-8

[10] Fernandes, G., & O'Sullivan, D. (2021). Benefits management in university-industry collaboration programs. *International Journal of Project Management, 39*(1), 71–84. https://doi.org/10.1016/j.ijproman.2020.09.008

[11] Jaafar, S. T. (2013). Listening Comprehension for first Grade Students of Department of English Language Arts at AL-Yarmouk University College. *Diyala Journal*, *58*, 799–819.

[12] Jaafar, S. T. (2014). The Effect of Gender on the Achievement of Third Year Students in the Area of English Drama. *Arts Journal*, *110*, 61–82.

[13] Jaafar, S. T. (2017a). *Iraqi EFL College Stduents' Performance in the Area of Perfect Tense, Methodology of Teaching English as a Foreign Langauge*. LAP LAMBERT Acadmic Publishing.

[14] Jaafar, S. T. (2017b). Second Year Students' Problems in Mastering English Sonnet. *Al-Yarmouk Journal*, *9*(9), 101–113.

[15] Jaafar, S. T., Buragohain, D., & Haroon, H. A. (2019). Differences and Classifications of Borrowed and Loan Words in Linguistics Context: A Critical Review. In D. Suryani, I. and Buragohain (Ed.), *International Languages and Knowledge: Learning in a Changing World* (2nd ed., pp. 95–112). Universiti Malaysia Perlis Press.

[16] Jaafar, S. T., Dzakiria, H., & Singh, M. K. S. (2021). Survey Study of Borrowings in the Arabic Language Based on The Hierarchy of Linguistics Branches. *The Asian EFL Journal*, *25*(1), 1–23.

[17] Labdaoui, K., Mazouz, S., Moeinaddini, M., Cools, M., & Teller, J. (2021). The Street Walkability and Thermal Comfort Index (SWTCI): A new assessment tool combining street design measurements and thermal comfort. *Science of the Total Environment, 795,* 148663. https://doi.org/10.1016/j.scitotenv.2021.148663

[18] Wright, D. J., Frank, C., & Bruton, A. M. (2022). Recommendations for combining action observation and motor imagery interventions in sport. *Journal of Sport Psychology in Action, 13*(3), 155–167. https://doi.org/10.1080/21520704.2020.1842442

[19] Jansen, M., Doornebosch, A. J., de Waal, M. W., Wattel, E. M., Visser, D., Spek, B., & Smit, E. B. (2021). Psychometrics of the observational scales of the Utrecht Scale for Evaluation of Rehabilitation (USER): Content and structural validity, internal consistency and reliability. *Archives of Gerontology and Geriatrics, 97,* 104509. https://doi.org/10.1016/j.archger.2021.104509

[20] Rohde, A., McCracken, M., Worrall, L., Farrell, A., O'Halloran, R., Godecke, E., … Doi, S. A. (2022). Inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane Evidence-Based Language Test. *Disability and Rehabilitation, 44*(4), 637–645. https://doi.org/10.1080/09638288.2020.1845832

[21] Wilson, M. H., Ashworth, E., Hutchinson, P. J., & British Neurotrauma Group. (2022). A proposed novel traumatic brain injury classification system: An overview and inter-rater reliability validation on behalf of the Society of British Neurological Surgeons. *British Journal of Neurosurgery,* 1–6. https://doi.org/10.1080/02688697.2022.2061456

[22] Peeters, M. J. (2021). Moving beyond Cronbach's alpha and inter-rater reliability: A primer on generalizability theory for pharmacy education. *Innovations in Pharmacy, 12*(1), 12. https://doi.org/10.24926/iip.v12i1.3555

[23] Martin, A. R., Jentzsch, T., Wilson, J. R., Moghaddamjou, A., Jiang, F., Rienmueller, A., … Fehlings, M. G. (2021). Inter-rater reliability of the modified Japanese Orthopedic Association score in degenerative cervical myelopathy: A cross-sectional study. *Spine, 46*(16), 1063–1069. https://doi.org/10.1097/BRS.0000000000003963

[24] Lindenschot, M., Koene, S., Nott, M. T., Nijhuis-van der Sanden, M. W., de Groot, I. J., Steultjens, E. M., & Graff, M. J. (2022). The reliability and validity of the perceive, recall, plan and perform assessment in children with a mitochondrial disorder. *Disability and Rehabilitation,* 1–14. https://doi.org/10.1080/09638288.2022.2100284

[25] Flannelly, K. J., Flannelly, L. T., & Jankowski, K. R. (2018). Threats to the internal validity of experimental and quasi-experimental research in healthcare. *Journal of Health Care Chaplaincy, 24*(3), 107–130. https://doi.org/10.1080/08854726.2018.1480854

[26] Gupta, R. K., Marks, M., Samuels, T. H., Luintel, A., Rampling, T., Chowdhury, H., … & Noursadeghi, M. (2020). Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: An observational cohort study. *European Respiratory Journal, 56*(6), 2003498. https://doi.org/10.1183/13993003.03498-2020

[27] Mungroop, T. H., Van Rijssen, L. B., Van Klaveren, D., Smits, F. J., Van Woerden, V., Linnemann, R. J., … Besselink, M. G. (2019). Alternative fistula risk score for pancreatoduodenectomy (a-FRS): Design and international external validation. *Annals of Surgery, 269*(5), 937–943. https://doi.org/10.1097/SLA.0000000000002610

[28] Pound, P., & Ritskes-Hoitinga, M. (2018). Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. *Journal of Translational Medicine, 16*(1), 304. https://doi.org/10.1186/s12967-018-1678-1

[29] Rahman, T., Khandakar, A., Abir, F. F., Faisal, M. A. A., Hossain, M. S., Podder, K. K., … Chowdhury, M. E. (2022). QCovSML: A reliable COVID-19 detection system using CBC biomarkers by a stacking machine learning model. *Computers in Biology and Medicine, 143,* 105284. https://doi.org/10.1016/j.compbiomed.2022.105284

[30] Siedlecki, S. L. (2020). Understanding descriptive research designs and methods. *Clinical Nurse Specialist, 34*(1), 8–12. https://doi.org/10.1097/NUR.0000000000000493

[31] Sileyew, K. J. (2019). Research design and methodology. In K. Ghezzar (Ed.), *Contemporary topics in science and engineering research* (pp. 1–12). IntechOpen. https://doi.org/10.5772/intechopen.85731

[32] Xie, J., Hungerford, D., Chen, H., Abrams, S. T., Li, S., Wang, G., … Toh, C. H. (2020). Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. *medRxiv.* https://doi.org/10.1101/2020.03.28.20045997

[33] Hao, T., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2020). Testing whether ensemble modelling is advantageous for maximizing predictive performance of species distribution models. *Ecography, 43*(4), 549–558. https://doi.org/10.1111/ecog.04871

[34] Garbern, S. C., Nelson, E. J., Nasrin, S., Keita, A. M., Brintz, B. J., Gainey, M., … & Leung, D. T. (2022). External validation of a mobile clinical decision support system for diarrhea etiology prediction in children: A multicenter study in Bangladesh and Mali. *Elife, 11*, e71705. https://doi.org/10.7554/eLife.71705

[35] Gerry, S., Bonnici, T., Birks, J., Kirtley, S., Virdee, P. S., Watkinson, P. J., & Collins, G. S. (2020). Early warning scores for detecting deterioration in adult hospital patients: Systematic review and critical appraisal of methodology. *BMJ, 369*, m1501. https://doi.org/10.1136/bmj.m1501

[36] Kamran, F., Tang, S., Otles, E., McEvoy, D. S., Saleh, S. N., Gong, J., … Wiens, J. (2022). Early identification of patients admitted to hospital for COVID-19 at risk of clinical deterioration: Model development and multisite external validation study. *BMJ, 376,* e068576. https://doi.org/10.1136/bmj-2021-068576

[37] Kernbach, J. M., & Staartjes, V. E. (2022). Foundations of machine learning-based clinical prediction modeling: Part II—Generalization and overfitting. *Machine Learning in Clinical Neuroscience,* 15–21. https://doi.org/10.1007/978-3-030-93311-1_3

[38] Reps, J. M., Schuemie, M. J., Suchard, M. A., Ryan, P. B., & Rijnbeek, P. R. (2018). Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association, 25*(8), 969–975. https://doi.org/10.1093/jamia/ocy032

[39] Takeshita, H., Tachibana, K., Sugiyama, H., Kagawa, M., Yano, A., Okada, Y., … Kawakami, S. (2022). Nomogram predicting testicular torsion in Japanese patients with acute scrotal pain using physical examination findings and environmental conditions: Development and prospective external validation. *International Journal of Urology, 29*(1), 42–48. https://doi.org/10.1111/iju.14734

[40] Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single-item measures in psychological science: A call to action. *European Journal of Psychological Assessment, volume 38, 1-5.*

[41] Zenker, S., Braun, E., & Gyimothy, S. (2021). Too afraid to travel? Development of a pandemic (COVID-19) anxiety travel scale (PATS). *Tourism Management, 84,* 104286. https://doi.org/10.1016/j.tourman.2020.104286

[42] Upadhyay, V., Saoji, A. A., Verma, A., & Saxena, V. (2022). Development and validation of 20-min yoga module for reducing burnout among healthcare workers. *Complementary Therapies in Clinical Practice, 46,* 101543. https://doi.org/10.1016/j.ctcp.2022.101543

[43] Semancik, B., Schmeler, M. R., Schein, R. M., & Hibbs, R. (2021). Face validity of standardized assessments for wheeled mobility & seating evaluations. *Assistive Technology,* 1–9. https://doi.org/10.1080/10400435.2021.1949051

[44] Almanasreh, E., Moles, R. J., & Chen, T. F. (2022). A practical approach to the assessment and quantification of content validity. In J. W. Smith & T. F. Chen (Eds.), *Contemporary research methods in pharmacy and health services* (pp. 583–599). Academic Press.

[45] Ishanuddin, N. M., Sukadarin, E. H., Nawi, N. S. M., Widia, M., Rashid, A. A. A., Aziz, H. A., … Jawi, Z. M. (2021). Design and implementation content validity study: Development of an instrument for measuring consumers' perception of automatic emergency braking (AEB). In *International Conference on*

Applied Human Factors and Ergonomics (pp. 382–392). Springer, Cham. https://doi.org/10.1007/978-3-030-80012-3_47

[46] Singh, A., & Ali, A. (2022). Psychosocial inventory for caregivers (PIC) of persons with mental illness: Content validity and cognitive interviewing process. *Asian Journal of Psychiatry, 70,* 103020. https://doi.org/10.1016/j.ajp.2022.103020

[47] Lamping, C., Derks, M., Koerkamp, P. G., & Kootstra, G. (2022). ChickenNet—An end-to-end approach for plumage condition assessment of laying hens in commercial farms using computer vision. *Computers and Electronics in Agriculture, 194,* 106695. https://doi.org/10.1016/j.compag.2022.106695

[48] Hickman, C., Marks, E., Pihkala, P., Clayton, S., Lewandowski, R. E., Mayall, E. E., … & van Susteren, L. (2021). Climate anxiety in children and young people and their beliefs about government responses to climate change: A global survey. *The Lancet Planetary Health, 5*(12), e863–e873. https://doi.org/10.1016/S2542-5196(21)00278-3

[49] Battineni, G., Sagaro, G. G., Chintalapudi, N., Di Canio, M., & Amenta, F. (2021). Assessment of awareness and knowledge on novel coronavirus (COVID-19) pandemic among seafarers. *Healthcare, 9*(2), 120. https://doi.org/10.3390/healthcare9020120

[50] Abedini, Z., Khoramirad, A., Ahmari Tehran, H., & Saeedi, M. (2022). Psychometric evaluation of the perceived nursing students' incivility questionnaire. *Nursing Open, 9*(3), 1709–1714. https://doi.org/10.1002/nop2.1197

[51] Mittal, V., & Sharma, R. K. (2021). Deep learning approach for voice pathology detection and classification. *International Journal of Healthcare Information Systems and Informatics, 16*(4), 1–30. https://doi.org/10.4018/IJHISI.2021100101

[52] Siyam, N., Hussain, M., & Alqaryouti, O. (2022). Factors impacting teachers' acceptance and use of bring your own device (BYOD) in the classroom. *SN Social Sciences, 2*(1), 1–30. https://doi.org/10.1007/s43545-021-00237-0

[53] Koo, B. W., Guhathakurta, S., & Botchwey, N. (2022). Development and validation of automated microscale walkability audit method. *Health & Place, 73,* 102733. https://doi.org/10.1016/j.healthplace.2021.102733

[54] Kalkbrenner, M. T. (2021). Alpha, Omega, and H internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation, 12*(1), 1–12. https://doi.org/10.1080/21501378.2020.1827313

[55] Kumari, A., Ranjan, P., Chopra, S., Kaur, D., Upadhyay, A. D., Kaur, T., … Vikram, N. K. (2021). Development and validation of a questionnaire to assess knowledge, attitude, practices, and concerns regarding COVID-19 vaccination among the general population. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 15*(3), 919–925. https://doi.org/10.1016/j.dsx.2021.04.024

[56] Lechien, J. R., Chiesa-Estomba, C. M., Hans, S., Calvo-Henriquez, C., Mayo-Yáñez, M., Tucciarone, M., … Saibene, A. M. (2021). Validity and reliability of the COVID-19 symptom index, an instrument evaluating severity of general and otolaryngological symptoms. *Acta Oto-Laryngologica, 141*(6), 615–620. https://doi.org/10.1080/00016489.2021.1888811

[57] Akel, K. B., Masters, N. B., Shih, S. F., Lu, Y., & Wagner, A. L. (2021). Modification of a vaccine hesitancy scale for use in adult vaccinations in the United States and China. *Human Vaccines & Immunotherapeutics, 17*(8), 2639–2646. https://doi.org/10.1080/21645515.2021.1884740

[58] Aziz, A. A., Mamat, M. N., Salleh, D. M., Abdullah, S. F. S., & Nordin, M. N. (2021). An analysis of systematic literature review on the development of Islamic-oriented instruments. *Journal of Contemporary Issues in Business and Government, 27*(1).

[59] Mahamid, F. A., Veronese, G., Bdier, D., & Pancake, R. (2021). Psychometric properties of the COVID stress scales (CSS) within Arabic language in a Palestinian context. *Current Psychology,* 1–10. https://doi.org/10.1007/s12144-021-02059-5

[60] Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics, 9*(1), 4–11. https://doi.org/10.12691/ajams-9-1-2

[61] Wauters, J., Couckuyt, I., & Degroote, J. (2021). ESLA: A new surrogate-assisted single-loop reliability-based design optimization technique. *Structural and Multidisciplinary Optimization, 63*(6), 2653–2671. https://doi.org/10.1007/s00158-021-02864-4

[62] Mellado-Berenguer, J., & Monfort-Pañego, M. (2025). Development, Validation, and Reliability of the System for Observing Teaching Competencies in Physical Education (SOTC-PE). *Measurement in Physical Education and Exercise Science*, *29*(1), 66-85.

[63] Slanzi, C. M., Wang, Y., Yang, Y., Schieber, E., Ahmed, I. A., Ibrahim, A. M., ... & McKune, S. L. (2025). The use of behavioral observation in global health: a case study from Rural Ethiopia. *medRxiv*, 2025-05.

[64] Worden, R. E., Holladay, B. P., McLean, S. J., Cochran, H., & Reynolds, D. L. (2025). Systematic Social Observation of Police-Citizen Encounters: Coding and Measurement Through Body-Worn Cameras. *Justice Quarterly*, 1-34.

[65] Lindblade, K. A., Mpimbaza, A., Ngufor, C., Yavo, W., Atobatele, S., Akpiroroh, E., ... & Humes, M. (2025). Assessing the accuracy of the recording and reporting of malaria rapid diagnostic test results in four African countries: methods and key results. *Malaria Journal*, *24*(1), 206.

[66] Tago, M., Katsuki, N. E., Yaita, S., Nakatani, E., Yamashita, S., Oda, Y., & Yamashita, S. I. (2021). High inter-rater reliability of Japanese bedriddenness ranks and cognitive function scores: A hospital-based prospective observational study. *BMC Geriatrics, 21*(1), 305. https://doi.org/10.1186/s12877-021-02261-y

[67] Chandra, S., Rasheed, R., Sen, P., Menon, D., & Sivaprasad, S. (2022). Inter-rater reliability for diagnosis of geographic atrophy using spectral domain OCT in age-related macular degeneration. *Eye, 36*(2), 392–397. https://doi.org/10.1038/s41433-021-01814-0

[68] Injeyan, H. S., Hogg-Johnson, S., Abdulla, S., Chow, N., Cox, J., Ridding, A., & Jacobs, C. (2021). Intra- and inter-rater reliability of an electronic health record audit used in a chiropractic teaching clinic system: An observational study. *BMC Health Services Research, 21*(1), 1–11. https://doi.org/10.1186/s12913-021-06131-9

[69] Northam, K. A., Parker, W. F., Chen, S. L., Cicci, J. D., Lin, F. C., Rollins-Raval, M. A., & Kasthuri, R. S. (2021). Evaluation of 4Ts score inter-rater agreement in patients undergoing evaluation for heparin-induced thrombocytopenia. *Blood Coagulation & Fibrinolysis, 32*(5), 328–334. https://doi.org/10.1097/MBC.0000000000001042

[70] Rau, G., & Shih, Y. S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes, 53,* 101026. https://doi.org/10.1016/j.jeap.2021.101026

[71] Rimikis, S., & Buchwald, A. (2019). The impact of morphophonological patterns on verb production: Evidence from acquired morphological impairment. *Clinical Linguistics & Phonetics, 33*(1–2), 68–94. https://doi.org/10.1080/02699206.2018.1504911

[72] Yang, L., Driscol, J., Sarigai, S., Wu, Q., Lippitt, C. D., & Morgan, M. (2022). Towards synoptic water monitoring systems: A review of AI methods for automating water body detection and water quality monitoring using remote sensing. *Sensors, 22*(6), 2416. https://doi.org/10.3390/s22062416

[73] Landmann, J., Fröhling, L., Gieschen, R., Buck, B. H., Heasman, K., Scott, N., … Hildebrandt, A. (2021). Drag and inertia coefficients of live and surrogate shellfish dropper lines under steady and oscillatory flow. *Ocean Engineering, 235,* 109377. https://doi.org/10.1016/j.oceaneng.2021.109377

[74] Almqvist, V., Berg, C., & Hultgren, J. (2021). Reliability of remote post-mortem veterinary meat inspections in pigs using augmented-reality live-stream video software. *Food Control, 125*, 107940. https://doi.org/10.1016/j.foodcont.2021.107940

[75] Midamba, D. C., Ndolo, O. F., Chepkoech, B., Agbolosoo, J. A., Ouya, F. O., & Jjengo, A. (2025). Data collection methods in social sciences: A primer for novice researchers and students. *South Asian Journal of Social Studies and Economics, 22*(6), 217–229. https://doi.org/10.9734/sajsse/2025/v22i61049

[76] de Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. (2021). A comparison of reliability coefficients for ordinal rating scales. *Journal of Classification, 38*(3), 519–543. https://doi.org/10.1007/s00357-021-09405-x

[77] Vanacore, A., & Pellegrino, M. S. (2022). Robustness of κ-type coefficients for clinical agreement. *Statistics in Medicine, 41*(11), 1986–2004. https://doi.org/10.1002/sim.9301

[78] Andrés, A. M., & Hernández, M. Á. (2022). Multi-rater delta: Extending the delta nominal measure of agreement between two raters to many raters. *Journal of Statistical Computation and Simulation, 92*(9), 1877–1897. https://doi.org/10.1080/00949655.2022.2033035

[79] Silveira, P. S. P., & Siqueira, J. O. (2022). Better to be in agreement than in bad company. *Behavior Research Methods,* 1–22. https://doi.org/10.3758/s13428-022-01891-y

[80] Vergni, L., Todisco, F., & Di Lena, B. (2021). Evaluation of the similarity between drought indices by correlation analysis and Cohen's kappa test in a Mediterranean area. *Natural Hazards, 108*(2), 2187–2209. https://doi.org/10.1007/s11069-021-04744-6

[81] Lopes, R., Geffroy, L., Padiolleau, G., Ngbilo, C., Baudrier, N., Mainard, D., … Amouyel, T. (2021). Proposal of a new CT arthrographic classification system of osteochondral lesions of the talus. *Orthopaedics & Traumatology: Surgery & Research, 107*(6), 102890. https://doi.org/10.1016/j.otsr.2021.102890

[82] Monteith, S., & Glenn, T. (2021). Comparison of potential psychiatric drug interactions in six drug interaction database programs: A replication study after 2 years of updates. *Human Psychopharmacology: Clinical and Experimental, 36*(6), e2802. https://doi.org/10.1002/hup.2802

[83] Vetchinnikova, S., Konina, A., Williams, N., Mikušová, N., & Mauranen, A. (2022). Perceptual chunking of spontaneous speech: Validating a new method with non-native listeners. *Research Methods in Applied Linguistics, 1*(2), 100012. https://doi.org/10.1016/j.rmal.2022.100012

[84] Chhablani, J., & Behar-Cohen, F. (2022). Validation of central serous chorioretinopathy multimodal imaging-based classification system. *Graefe's Archive for Clinical and Experimental Ophthalmology, 260*(4), 1161–1169. https://doi.org/10.1007/s00417-021-05503-7

[85] Huang, M., Wang, Y., Sun, Y., Zhou, Y., Liu, Y., & Ye, H. (2022). The accuracies of three intraoral scanners with regard to shade determination: An in vitro study. *Journal of Prosthodontics*. https://doi.org/10.1111/jopr.13611

[86] Fatima, N., Mento, F., Zanforlin, A., Smargiassi, A., Torri, E., Perrone, T., & Demi, L. (2022). Human-to-AI interrater agreement for lung ultrasound scoring in COVID-19 patients. *Journal of Ultrasound in Medicine*. https://doi.org/10.1002/jum.16156

[87] Webb, J. M., Adusei, S. A., Wang, Y., Samreen, N., Adler, K., Meixner, D. D., … Alizad, A. (2021). Comparing deep learning-based automatic segmentation of breast masses to expert interobserver variability in ultrasound imaging. *Computers in Biology and Medicine, 139,* 104966. https://doi.org/10.1016/j.compbiomed.2021.104966