

Scale adaptation and redevelopment: A review on validity and reliability

Esra Akaydın Gültürk^{1*}

¹Sivas Cumhuriyet University,
Department of Biostatistics, Turkey

*Author for correspondence:
Email: esra0709.eg@gmail.com

Received date: July 24, 2024
Accepted date: October 17, 2024

Copyright: © 2024 Gültürk EA. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Summary

Scales are data collection tools that can measure characteristics such as knowledge, emotion, interest, perception, attitude, belief, disposition, risk, quality of life and behavior. The scale development process includes determining the theoretical foundations, creating the items, pilot study, validity and reliability analysis, and final implementation. In the stage of determining the theoretical foundations, the definitions of the construct that the scale wants to measure in the literature are examined and a conceptual framework is created. At the stage of creating the items, expert opinions are taken and their suitability in terms of language and meaning is evaluated. In the pilot study phase, the developed items are applied to a small sample group to test the comprehensibility and functionality of the scale. In validity analyses, content, construct and criterion validity are evaluated. In reliability analyses, internal consistency, test-retest reliability and item-total correlation are measured. In the final implementation phase, the final version of the scale is administered to a large sample group and the validity and reliability analyses are repeated and the final scale is created. This process ensures the scientific accuracy and reliability of the scale. In addition, translation, cultural compatibility tests and pilot applications play an important role in the process of adapting scales to different cultures.

Keywords: Scale development, Validity, Reliability, Scale, Cultural adaptation

Introduction

Scales are data collection tools that can measure characteristics such as knowledge, emotion, interest, perception, attitude, belief, disposition, risk, quality of life, behavior. Often the rationale for using a test is that it covers more than one purpose. They can be used to assess achievement, changes in performance over two different time periods, classification, diagnosis, instruction, grading, prediction, program evaluation, etc. The data obtained from measurements form the building blocks of scientific research. Therefore, using, modifying, adapting and developing scales is critical in all human-based sciences [1,2]. Validity and reliability are two fundamental concepts in the evaluation of tests and measurement instruments. These concepts are used to determine whether a test gives accurate and consistent results. Validity refers to a test's ability to measure in accordance with its purpose, while reliability describes the consistency and repeatability of the test. Understanding these two concepts in the development and implementation of tests ensures that tests are used accurately and effectively [1].

When the Data Type is Numerical the Criteria Used for Validity Can be Given as Follows

Content validity

It determines whether the scale items cover all dimensions of the concept to be measured. It is

evaluated by expert opinions and is ensured by experts examining each item and item groups. There are different approaches to determining content validity. These approaches include seeking expert opinion and calculating the correlation coefficient between the new test and another test known to measure the same content. The most frequently used approaches for expert opinion are the Lawshe method and the Davis method. In the Lawshe method; an expert panel is formed to evaluate the validity of the scale items. This panel is selected from 5 to 40 individuals who are familiar with the subject [3]. Experts are asked to evaluate each item according to three categories.

1. It is necessary and should remain in the substance pool
2. The item is useful but not sufficient. The item contributes to the concept being measured but is not essential
3. Unnecessary.

Content Validity Ratio (CVR): CVR value is calculated for each item. Where N_e is the number of experts who rated the relevant item as “necessary” and N is the total number of experts [3].

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

A positive CVR value indicates that the number of experts who rated the item as “Required” is more than half, while a negative CVR value indicates that it is less than half, and equal when the CVR value is zero. Zero and negative ones are eliminated from the item pool. The items with a positive CVR are then compared with the table value calculated according to the number of experts at a certain significance level [3,4]. Another approach used in consulting expert opinion is the Davis method. In this method, experts rate their opinions as “the item is appropriate”, “the item should be slightly revised”, “the item should be seriously revised” and “the item is not appropriate”. Then, CVR values are calculated by summing the number of experts who say that the item is appropriate and should be slightly revised and dividing by the total number of experts and compared with 0.80 to decide whether the item should remain in the scale or not [4].

Construct validity

Determines the extent to which the scale accurately measures a theoretical construct or concept. It is tested by factor analysis; in Exploratory Factor Analysis (EFA), the relationships between items are examined and it is evaluated whether the scale conforms to the theoretical construct [5]. Factor loadings, eigenvalues and rotation techniques (varimax, promax) are frequently used in this process [6]. Confirmatory Factor Analysis (CFA) tests how well a predetermined factor structure fits the data. CFA uses various fit indices to assess model fit. The fit indices used in CFA are: RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), TLI (Tucker-Lewis Index), SRMR (Standardized Root Mean Square Residual), Chi-Square (χ^2) Test, GFI (Goodness-of-Fit Index), and AGFI (Adjusted Goodness-of-Fit Index). A RMSEA and SRMR below 0.08 and CFI, TLI, GFI and AGFI above 0.90 indicate good fit [7,8]. Exploratory Factor Analysis (EFA) is used to explore the factor structure of the scale and determine the number of factors. The main methods used in this analysis are known as PCA (Principal Components Analysis) and PAF (Principal Axis Factoring). EFA is particularly preferred when the scale is newly developed or has not been adapted into another language [9,10].

Criterion validity

Determines how well the scale relates to a relevant external criterion. There are two types: concurrent validity and predictive validity. Concurrent validity tests how well the scale relates to another valid measure measured at the same time, while predictive validity tests how well the scale predicts future performance or behavior [10]. Cross-validation assesses how well a regression model performs on different data sets and measures its predictive accuracy. The cross-validation method is used to determine the validity of the regression model and to test the accuracy of the study results. In this method, a group of 200 people randomly selected from the population is randomly divided into two groups. A prediction equation is created using the data of the first group. In this equation, the values of the predictor independent variables in the second group are substituted and criterion scores are calculated for the second group. These benchmark scores are compared with the actual scores of the second group. If the scores are close to each other, this indicates the accuracy of the prediction. In this process, the standard error of the estimation is used to provide information about the average error of the criterion scores obtained from the estimation equation [4,11]. The Bland-Altman plot is used when a new test is developed against a reference, to assess whether a new instrument, thought to measure the same trait, performs similarly to the classical measurement method. This graphical approach is applied to examine the agreement between two measurement methods, as long as the units are the same. It is also used to assess test-retest reliability. In addition to correlation and regression analyses, the Bland-Altman plot is used as a stand-alone method to determine how different a new test is from a reference test [12].

Validity Measures When Data Type is Qualitative

It is given by $\text{Sensitivity} = a / (a + c)$. It shows what % of those who actually succeeded were found to have succeeded as a result of the test developed.

Selectivity is given by $= d / (b + d)$. It shows how many % of those who actually failed, failed as a result of the test developed.

The false negative rate is denoted by $= c / (a + c)$. It is given as the percentage of the new test falsely classified as failing among true passers.

False positive rate = $b / (b + d)$ Defined as the percentage of true failures who are mistakenly labeled as successful as a result of the test developed.

The positive predictive value is given by $= a / (a + b)$ and gives the probability that those who are sick as a result of the new test developed are actually sick.

The negative predictive value is given by $= d / (c + d)$ and gives the probability that those who are robust as a result of the new test developed are actually robust [4,13].

Performance measures combining sensitivity and selectivity

To compare the performance of diagnostic tests, combined measures such as correct classification rate, youden index, likelihood ratio, odds ratio are used, which are calculated by combining sensitivity and selectivity values.

Correct classification rate

The closer it is to 1, the higher the delivery performance is said to be, if the c value is 0.50 and below 0.50, it is interpreted that the classification made with the developed test is by chance. If the number of individuals in the study is not close to each other, it is recommended to use sensitivity and selectivity values instead of the correct classification rate [14].

$$C = (a + d) / (a + b + c + d)$$

Youden index

Gives an overall value for the performance of the test and is also used to compare multiple tests. It is desired to be close to 1.

Youden Index

$$\text{sensitivity} + \text{selectivity} - 1$$

Likelihood ratios

LR (+) = sensitivity / (1-selectivity) Defined as the ratio of the probability that the new test will detect a patient in the truly sick to the probability that the new test will detect a patient in the truly healthy. It indicates how many true positives the diagnostic test yields against one false positive. The higher the ratio, the better the true patients are separated. LR (-): (1-sensitivity)/Selectivity, the smaller this ratio, the better the true intact individuals are discriminated, in other words, the smaller the ratio, the better the negative success of the test [14] (Table 1).

Table 1. Display of validity measures when there is Qualitative Data.

Test Values	True Values	
	Positive	Negative
Positive	A (TP)	B (FP)
Negative	C (FN)	D (TN)
Total	A+C	B+D

TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative

Reliability Measures in Digital Data Types

Test-retest reliability

It is the measurement of the correlation between the results by applying the same test on the same group at different times. Pearson correlation coefficient is usually used in this method. A high correlation coefficient indicates that the test gives consistent results over time [15].

Internal consistency

It shows how consistent the items in the measurement tool are with each other. Cronbach's Alpha is the most commonly used coefficient in internal consistency assessments and takes values between 0 and 1. Cronbach's Alpha values of 0.70 and above are generally considered acceptable [1,16]. Item-total correlation measures the correlation of each item with the total score. A high correlation indicates that the item reflects the overall scale well. Item-total correlation is used to examine the relationship between each item in a test and the overall test score [17].

Kuder-Richardson coefficients (KR-20 and KR-21)

It measure internal consistency for dichotomous items. KR-20 is used when items have different degrees of difficulty, while KR-21 is used under the assumption that all items have the same degree of difficulty [17]. He states that the KR-20 coefficient is calculated by the following formula:

$$KR-20 = \frac{K}{K-1} \left(1 - \frac{\sum p_i q_i}{s^2} \right)$$

Here, K is the number of items, p_i is the correct answer rate of item i, q_i is the incorrect answer rate of item i, and s^2 is the variance of test scores. This method is used to measure the internal consistency of true/false tests [18]. Alternate Forms Reliability: It is the application of two different forms of tests to the same group and measuring the correlation between the results. This method is assessed using Pearson's correlation coefficient and can be used to measure the correlation between two different test forms measuring the same level of knowledge [19]. Accuracy Analysis assesses how close the measurement results are to the actual values. Metrics such as calculation of the mean error, mean absolute error (MAE) and mean square error (MSE) are used. Cohen *et al.* show that these methods are used to assess the accuracy of a measuring device [20].

Adaptation of a Scale to a Different Culture

The first step in the scale adaptation process is to carefully select the scale to be adapted. The selected scale should have high validity and reliability values in the original language. In addition, the process of obtaining cultural permission for the concepts that the scale wants to measure in the target population is the beginning of the scale adaptation process [21].

Literature review

To obtain information about the context in which the scale is used and similar studies in the language/culture to be adapted.

Obtaining permission

Obtaining the necessary permissions for adaptation from the original author or publisher of the scale.

Forward translation

Translation of the scale from the original language into the target language by at least two independent translators [22].

Expert panel

Bringing together translators and subject matter experts to review translations and ensure language and cultural alignment [21,22].

Back translation

The scale translated into the target language is retranslated in the original language by an independent translator.

Back translation comparison

Comparing the back translation with the original scale and identifying differences in meaning [23].

Pre-application

Application of the scale in 30-40 samples for a small sample group in the target language [24].

Feedback collection

Obtaining feedback from participants about the comprehensibility and cultural appropriateness of the scale items [25].

Edits

After revising the scale based on feedback and making the necessary adjustments, the scale is administered to a group of 300-400 individuals on the target sample [26].

Validity analysis

Construct validity, criterion validity and content validity analyses as a result of the scale application [25].

Reliability analysis

Internal consistency, test-retest reliability and item analysis [25].

Language and cultural equivalence

Evaluation of the item-based equivalence of the original and target language versions of the scale [27].

Final edits and reporting

The final version of the scale is created according to the results of the psychometric evaluation, the adaptation process and the findings obtained are reported in detail [28].

Scale Development Stages

At this stage, the definitions of the concept or construct to be measured by the scale in the literature and the dimensions of this construct are determined. In this context, a literature review is conducted and a theoretical framework is created. Then, an item pool is created by experts in the field. Based on the theoretical framework, the items of the scale are written. These items should cover the concept or construct that the scale wants to measure [29,30]. In the literature, it is reported that the sample size for the applicability of factor analysis is “50 very bad, 100 bad, 200 appropriate, 300 good, 500 very good and 1000 excellent”. It is stated that a sample size of at least 300 will give good results for the general applicability of factor analysis [31,32]. Studies emphasize the importance of adequate sample sizes in factor analysis. For example, MacCallum *et al.* emphasize that a sample size of at least 300 is generally considered sufficient for reliable factor analysis results. Their findings suggest that smaller sample sizes may lead to unstable factor solutions that may jeopardize the validity of the scale [33]. Another important criterion for factor analysis is the correlation between items. The item-total correlation coefficient is calculated by evaluating the correlation between each item and a scale score that excludes that item. If the inter-item correlation coefficient is below 0.30, it is reported that the inter-item correlation is insufficient and items should be removed [34,35]. A correlation coefficient below 0.30 is considered as low correlation; between 0.30 and 0.50 as weak correlation; between 0.50 and 0.70 as medium-good correlation; between 0.70 and 0.90 as high correlation and above 0.90 as very high correlation [36]. If the item-total correlation score is below 0.30, it is expected to be removed from the scale, resulting in a significant improvement in the scale Cronbach Alpha coefficient. If the item-total correlation value is negative or very low (below 0.30), the statement item is invalid. However, if the item-total correlation value is greater than the critical or cut-off value (0.30), the statement

is considered valid [37]. Bartlett's Test of Sphericity tests whether the observed correlation matrix is a unit matrix with all off-diagonal values zero. Bartlett's Test of Sphericity tests whether the observed correlation matrix is a unit matrix with the property of having all off-diagonal zero values [13]. If Bartlett's test of sphericity is significant, the results indicate that the analyzed data are not unit matrices and are suitable for CFA. However, this test can identify problematic data sets and should be performed before CFA [15]. Although Bartlett's test of sphericity is accepted as significant in many data set analyses, some authors in the literature have found the result of Bartlett's test of sphericity to be statistically insignificant [38]. The validity of the scale is done to determine how accurately the scale measures the construct it is intended to measure. In this process, various types of validity (content validity, construct validity, criterion validity) are evaluated [39,40]. The reliability of a scale determines its ability to provide consistent and stable measurements. Reliability analyses are usually performed by methods such as internal consistency (Cronbach's Alpha), test-retest reliability and item-total correlation. [16] The final version of the scale is applied to a large sample group and validity and reliability analyses are repeated and the scale is finalized. The use of appropriate analyses for data types ensures that validity and reliability analyses are performed accurately and reliably in the scale development process [40].

Dichotomous data

Binary data is often analyzed using methods such as logistic regression. This analysis helps to test whether the scale accurately measures a particular construct or behavior. Methods such as the chi-square test contribute to tests of content validity and construct validity by examining the relationships between scale items. In binary data, Kuder-Richardson (KR-20 and KR-21) coefficients are used specifically to measure internal consistency. These coefficients determine whether reliability is high by assessing the consistency between the items of the scale. In ordinal data, ordinal logistic regression or ordinal probit analyses contribute to construct validity by testing whether the scale items accurately measure the theoretical construct. When used in combination with methods such as Confirmatory Factor Analysis (CFA) and Exploratory Factor Analysis (EFA), it can show whether the scale has the correct factor structure [43]. Reliability measures such as Cronbach's Alpha and item-total correlation are used for ordinal data types. These methods, which measure the internal consistency of ordinal data, evaluate the contribution of items to the overall consistency of the scale [43].

Continuous data

For continuous data, regression analysis and ANOVA are used to assess whether the scale accurately measures a construct. In addition, predictive validity can be tested with a continuous data construct, which reveals whether the scale accurately predicts future performances or behaviors. Alpha is also often used with continuous data types. In addition, test-retest reliability is tested using Pearson correlation coefficient in continuous data. In this way, it is checked whether the scale gives consistent results over time [44].

Conclusion

Scales are important data collection tools that can measure characteristics such as knowledge, emotions, interests, perceptions, attitudes, beliefs, dispositions, risks, quality of life and behaviors. They form the basic building blocks of scientific research and are

used for a variety of purposes such as achievement assessment, classification, diagnosis, grading and program evaluation. In order for scales to be used accurately and effectively, it is critical to analyze their validity and reliability. Validity refers to a test's ability to measure in accordance with its purpose, while reliability defines the consistency and reproducibility of the test. Various types of validity, such as content validity, construct validity and criterion validity, ensure that scales measure accurately. Reliability analyses are performed by methods such as internal consistency, test-retest reliability, and item-total correlation. The scale development process starts with the determination of the theoretical foundations and continues with stages such as item development, pilot testing, validity and reliability analyses. In this process, the validity and reliability of the scale items are tested using methods such as expert opinions and factor analyses. In addition, steps such as translation, back translation, and pilot testing are followed in the process of adapting scales to different cultures. Today, the methods and analyses used in scale development and adaptation processes are constantly evolving and renewed. Therefore, it is important for researchers to follow the current literature and use best practices in scale development processes. As a result, accurate and consistent scales with validity and reliability analyses increase the quality of scientific research and ensure the reliability of the data obtained. Likert-type scales play an important role in the data collection process in scientific research. There are many important points to be considered in the use, modification, adaptation and development of these scales. Researchers should pay attention to the validity and reliability analysis of the scales, and adapt and develop the scales by taking into account the characteristics of the target group. Carrying out these processes meticulously will improve the quality of scientific research and contribute to obtaining more reliable data. In this review study, the methods in the literature on scale development and adaptation processes were examined and the difficulties encountered in the validity and reliability analysis of these processes were discussed. Sample size, types of data used and cultural adaptation processes are among the methodological limitations of the reviewed studies. In addition, it was stated that methods based on expert opinions in validity analyses may include subjective differences and this may limit the generalizability of the findings. Despite these limitations, our review study made important contributions to the existing approaches to scale development and adaptation in the literature and provided a guiding framework for future studies in this field with more comprehensive and diverse data groups.

References

1. Nunnally J, Bernstein I. Psychometric Theory. 3rd edition. New York: MacGraw-Hill; 1994.
2. Anastasi A. Psychological testing. Upper Saddle River, NJ: Prentice Hall; 1997.
3. Lawshe CH. A Quantitative Approach to Content Validity. Personnel psychology/Berrett-Koehler Publishers. 1975.
4. Alpar R. Uygulamalı istatistik ve geçerlik-güvenilirlik: spor, sağlık ve eğitim bilimlerinden örneklerle. Detay Yayıncılık; 2010.
5. Brown TA. Confirmatory factor analysis for applied research. Guilford publications; 2015 Jan 7.
6. Costello AB, Osborne J. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research, and Evaluation. 2019;10(1):7.
7. Byrne BM. Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. International Journal of Testing. 2001 Mar 1;1(1):55-86.
8. Kline P. Handbook of psychological testing. London: Routledge; 2013 Nov 12.
9. Tabachnick BG, Fidell LS, Ullman JB. Using multivariate statistics. Boston, MA: pearson; 2013 Jul.
10. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods. 1999 Sep;4(3):272.
11. AYLAR F, Nagihan EV. DERLEME: ÖLÇEK GELİŞTİRME ÇALIŞMALARINDA DOĞRULAYICI FAKTÖR ANALİZİNİN KULLANIMI. The Journal of Social Sciences. 2019 Aug 22;4(10):389-412.
12. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet. 1986 Feb 8;327(8476):307-10.
13. Şencan H. Güvenilirlik ve geçerlilik. Hüner Şencan; 2005.
14. Sachs MC. plotROC: a tool for plotting ROC curves. Journal of Statistical Software. 2017 Aug;79.
15. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine. 2016 Jun 1;15(2):155-63.
16. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education. 2018 Dec;48:1273-96.
17. Field A. Discovering statistics using SPSS. 3rd Edition. London: SAGE Publications Ltd; 2009.
18. Tavakol M, Dennick R. Making sense of Cronbach's alpha. International Journal of Medical Education. 2011;2:53.
19. Cohen RJ, Swerdlik ME, Phillips SM. Psychological testing and assessment: An introduction to tests and measurement. Mayfield Publishing Co; 1996.
20. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. 2005 Dec 19;30(1):79-82.
21. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine. 2000 Dec 15;25(24):3186-91.
22. Cruchinho P, López-Franco MD, Capelas ML, Almeida S, Bennett PM, Miranda da Silva M, et al, and Validation of Measurement Instruments: A Practical Guideline for Novice Researchers. Journal of Multidisciplinary Healthcare. 2024 Dec 31:2701-28.
23. Van Teijlingen E, Hundley V. The importance of pilot studies. Social Research Update. 2001(35):1-4.
24. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. Journal of Evaluation in Clinical Practice. 2011 Apr;17(2):268-74.
25. Kömürlüoğlu A, Akaydın Gültürk E, Yalçın SS. Turkish Adaptation, Reliability, and Validity Study of the Vaccine Acceptance Instrument. Vaccines. 2024 Apr 29;12(5):480.
26. Miller LA, Lovler RL. Foundations of psychological testing: A practical approach. London: SAGE Publications Ltd; 2018 Dec 20.
27. Mota FR, Victor JF, Silva MJ, Bessa ME, Amorim VL, Cavalcante

- ML, et al. Cross-cultural adaptation of the Caregiver Reaction Assessment for use in Brazil with informal caregivers of the elderly. *Revista da Escola de Enfermagem da USP*. 2015 Jun;49(3):424-31.
28. Worthington RL, Whittaker TA. Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*. 2006 Nov;34(6):806-38.
29. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health*. 2018 Jun 11;6:149.
30. Knekta E, Runyon C, Eddy S. One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE—Life Sciences Education*. 2019;18(1):rm1.
31. Kyriazos TA. Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*. 2018 Aug 24;9(08):2207.
32. Yaşlıoğlu MM. Sosyal bilimlerde faktör analizi ve geçerlilik: Keşfedici ve doğrulayıcı faktör analizlerinin kullanılması. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*. 2017 Nov;46:74-85.
33. MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. *Psychological Methods*. 1999 Mar;4(1):84.
34. Büyüköztürk Ş. Sosyal bilimler için veri analizi el kitabı. Pegem Atf İndeksi. 2018:001-214.
35. Ramjit S. Primary Care Assessment Tool-adult edition (PCAT-AE) and the assessment of the primary care in South-West Trinidad (Doctoral dissertation).
36. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*. 2012;24(3):69-71.
37. Dirlik EM, Koç N. The analysis of the psychological tests using in educational institutions according to the testing standards. *Journal of Measurement and Evaluation in Education and Psychology*. 2017 Dec 1;8(4):453-68.
38. Plake BS, Wise LL. What is the role and importance of the revised AERA, APA, NCME Standards for Educational and Psychological Testing?. *Educational Measurement: Issues and Practice*. 2014 Dec;33(4):4-12.
39. Flake JK, Pek J, Hehman E. Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*. 2017 May;8(4):370-8.
40. McNeish D. Thanks coefficient alpha, we'll take it from here. *Psychological Methods*. 2018 Sep;23(3):412.
41. Hair J. Multivariate data analysis. *Exploratory factor analysis*. 2009.
42. Field A. *Discovering statistics using IBM SPSS statistics*. London: Sage Publications Limited; 2024 Feb 22.
43. Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. Hoboken, NJ: John Wiley & Sons; 2013 Feb 26.
44. Montgomery DC, Runger GC. *Applied statistics and probability for engineers*. Hoboken, NJ: John Wiley & Sons; 2010 Mar 22.