Systemic Analytics

www.sa-journal.org

Syst. Anal. Vol. 3, No. 1 (2025) 20–26.

Paper Type: Original Article

Adoption of Likert-Type Scales for Airline Service Quality Assessment

Adetayo Olaniyi Adeniran[1], Olutayo Sunday Fakunle[2,*]

[1] Department of Transport Planning and Logistics, University of Ilesa, Ilesa, Osun State, Nigeria; adetayo_adeniran@unilesa.edu.ng.
[2] Department of Sociology, Redeemers's University, Ede, Osun State, Nigeria; fakunles@run.edu.ng.

Citation:

Received: 18 May 2024 *Revised: 27 June 2024* *Accepted: 03 August 2024*	Adeniran, A., O., Fakunle, O., S. (2025). Adoption of likert-type scales for airline service quality assessment. *Systemic Analytics, 3*(1), 20-26.

Abstract

Likert scales are helpful in social science and perception studies. High-quality tests are essential to examine the dependability of the data provided in a particular evaluation. One often-used measure of test reliability is Cronbach alpha. The dimensionality and test duration have an impact on it. The fundamentally tau-equivalent approach's (a statistical method that measures how consistent a set of items are in a test) presumptions should be followed by alpha as a reliability indicator. If these presumptions are not met, a low alpha is shown. The Airline Service Quality (ASQ) test measures the quality of services offered by a specific airline using service attributes provided by the airline, which influence passenger satisfaction and enhance more than one-time patronage. The original instrument used measured opinions of a four-point Likert scale with fifteen airline service items, and the allowable Cronbach's Alphas for the original test ranged from 0.70 to 0.95. By first adding a "3 represents undecided" option and then adding four-word items to the instrument, a five-point Likert scale was developed from the original instrument. The test was piloted with 33 participants.This pilot study's Cronbach's alpha was 0.85; following its employment in a larger research study, the instrument's Cronbach's alpha was 0.86, resulting in a tool with good internal consistency. Since reliability test depends on the extent of test, Cronbach alpha does not only evaluate test homogeneity or unidimensionality. Whether a test is homogeneous or not, its reliability is increased by its extent. A high alpha score (> 0.90) might indicate that the test should be shorter and may indicate redundancy.

Keywords: Coefficient alpha, Cronbach's alpha, Multidimensionality, Reliability, Unidimensionality.

1 | Introduction

Perception studies, particularly with attitude scores and rating scales, are frequently utilized. A Likert-type scale is commonly used in such instruments. According to a Likert-type scale, a person must indicate whether they Strongly Agree (SA), Agree (A), Disagree (D), are Neutral (N), or Strongly Disagree (SD) with a series

Corresponding Author: fakunles@run.edu.ng
10.31181/sa31202537

of items. Every response has a point value, and the sum of the point values for all the assertions determines a person's score [1].

To improve the accuracy of assessments and evaluations, social and scientific researchers try to develop tests and questionnaires that are both valid and reliable [2], [3]. Validity and reliability are two essential components in evaluating a measurement tool. Instruments can be conventional knowledge, skill, or attitude tests, social and experimental simulations, or survey questionnaires [4–7]. Instruments can measure concepts, psychomotor skills, or affective values.

Validity is concerned with how well an instrument measures what it is supposed to measure, and reliability is concerned with how consistently an instrument measures [8]. It should be noted that an instrument is closely related to its validity [6], [9]. However, the reliability of an instrument does not depend on its validity [10]. The reliability of an instrument can be measured objectively [11], and the definition of Cronbach's alpha, the most used objective reliability metric, was provided in this study.

When multiple-item assessments of a concept or construct are used in social and scientific researches, calculating alpha has become a standard procedure [12], [13]. This is because it only has to be administered once, making it simpler to utilize than other estimates (such as test and retest reliability estimates) [9], [11], [14]. Nevertheless, despite the extensive usage of Cronbach's alpha in literature, its definition, appropriate application, and interpretation remain unclear [15], [16]. To encourage its efficient use, it is crucial to provide more context for the underlying presumptions of alpha. It should be noted that the sole emphasis of this study is Cronbach's alpha as a dependability indicator [17], [18]. The quality of social and scientific evaluations may be monitored and enhanced by using alternative reliability measurement techniques based on other psychometric methods, such as item-response theory or generalisability theory [19].

This study explains how the Likert scale and the number of items for the existing instrument were modified for use and how data were gathered to confirm the reliability of the modified instrument. Rensis Likert [9] developed this type of scale and subsequently developed this technique for assessing attitudes. A Likert rating scale measurement can be useful and reliable for measuring service quality.

1.1 | Likert-Type Scales

A variety of answers to a statement or set of assertions are offered by Likert scales. Responses are often divided into five categories, with 5 denoting SA, 1 denoting Strong Disagree, and 3 denoting Undecided [24]. On the other hand, academics D over how many options a Likert-type scale should include. According to [2], [14], some researchers choose scales with seven items or an even number of answer items. According to Symonds [20], a seven-point rating system has the best dependability. If there are more, the reliability gains would be so negligible that it would not be worthwhile to create the instrument or study the difference [2].

The topic of Likert scale items or categories [9] has been the focus of a lot of research, with many seemingly contradicting results. For instance, according to Guilford [21], the ideal number of categories depends on the circumstances and must be empirically determined. However, the number of scale points used for the items has no bearing on the validity and reliability of an instrument [4]. In response, Ray [22] questioned the suitability of Mattell's studies, which employed unmatched student groups for sampling. Thus, if a sub-sample were exceptionally diverse, the response format responded to may appear to have artificially low dependability [2].

Additionally, Ray [22] found a substantial difference between the variously designed Likert scales. There was an increase in discriminating power and internal reliability [9] when the number of Likert items was increased from three to five. For internal consistency reliability, the researcher must compute and report the Cronbach's alpha coefficient when employing Likert-type scales. Cronbach's alpha calculates an instrument's internal consistency reliability by analyzing how each item relates to every other item in the instrument. Internal consistency reliability is the degree to which items in an instrument are consistent with each other and the instrument as a whole [1].

1.1.1 | What is cronbach alpha?

To quantify the internal consistency of a test or scale, Lee Cronbach created alpha in 1951. It is represented by a number ranging from 0 to 1 [18], [19], [23]. The degree to which each item in a test measures the same notion or construct is known as internal consistency, and it is related to the test's item interrelationships [19], [24]. To ensure validity, a test's internal consistency should be established before it is used for research or examination [17], [20]. Furthermore, reliability estimates display a test's measurement error. Simply put, this understanding of dependability is the correlation of the test with itself [25], [26].

This correlation is squared, and the index of measurement error is obtained by subtracting from 1.00. When a test's reliability is 0.80, for instance, its scores have a 0.36 error variance (random error) ($0.80 \times 0.80 = 0.64$; $1.00 – 0.64 = 0.36$), as shown in [8], [9], [27]. The percentage of a test score that may be attributed to a mistake will decrease as the reliability estimate rises. It is important to remember that test reliability shows how measurement error affects observed values rather than a specific participant [21], [28]. The measurement error standard must be computed to determine how measurement error affects a participant's observed value [29], [30].

A test's alpha score rises as its elements are connected. Nevertheless, a high degree of internal consistency is not necessarily indicated by a high coefficient alpha [31]. This is because the length of text also impacts alpha [32]. The alpha value decreases if the test duration is too short. More items that test the same idea should thus be added to the test to raise alpha [1], [32]. It's also critical to remember that alpha is a characteristic of test measurement data from a particular participant sample [33]. As a result, researchers should assess alpha each time the test is given rather than depending on published alpha estimates.

For data analysis, the researcher should add up the scales and not bother about examining each item separately. For individual items, Cronbach's alpha does not yield reliability estimates [23]. Even-numbered Likert scales, which lack a N point, enable participants to adopt a stance even if they are unsure of their viewpoint [31]. Likert scales with odd numbers allow for neutrality or indecision. Response bias, or the propensity to prefer one response over another, may be lessened by providing respondents with a N response choice, which frees them from having to decide on a given topic [29].

If participants donot have an opinion, they do not feel compelled to express it. It has been demonstrated that statistics is impacted by the usage of mid-point items. Considering the context of preliminary findings is important since the inclusion or exclusion of a midpoint might significantly change the results of a population poll intended to determine opinion. The explicit usage of a mid-point is still up for discussion, and it depends more on the preference of researchers [32]. To prevent response bias, it has long been recommended that survey instruments have both favorably and negatively phrased items [11], [18].

To serve as "cognitive speed bumps that require participants to engage in more controlled, as opposed to automatic, cognitive processing," negatively phrased questions are added to the scale [13]. The essential premise that items with opposing wording measure the same notion that positively worded items measure is the foundation for using negatively phrased questions to reduce response bias [13]. Comparing any of the three circumstances where negatively worded stems were utilized, Barnette [5] discovered that Cronbach's alpha was greater and explained at least 10%, and in one instance, 20%, higher internal consistency.

1.1.2 | Use of cronbach's alpha

When alpha is improperly used, it can result in instances where a test or scale is incorrectly rejected or when the test is blamed for producing unreliable findings [31]. Understanding the related ideas of internal consistency, homogeneity, or unidimensionality can help to optimize the usage of alpha and prevent this scenario. While homogeneity relates to unidimensionality, internal consistency focuses on how a sample of test items relate to one another [25]. If a measure's items assess only one latent characteristic or concept, it is unidimensional. To measure homogeneity or unidimensionality in a sample of test items, an assessment of internal consistency is required, but it is an insufficient prerequisite [8].

Reliability is fundamentally based on the assumption that a sample of test items is unidimensional; if this condition is not met, dependability is significantly underestimated [8]. A multidimensional test does not always have a lower alpha than a unidimensional test, as has been well shown [28]. Accordingly, alpha cannot be only understood as a measure of a test's internal consistency, according to a more strict interpretation [29]. The dimensions of a test can be found via factor analysis. Therefore, alpha may be used to verify if a sample of things is truly unidimensional, rather than just measuring the unidimensionality of a group of objects [33].

However, if a test contains several concepts or constructs, it might not be a good idea to state the test's alpha overall since the higher number of questions will inevitably cause the alpha value to be inflated [1]. Therefore, in theory, alpha should be computed for each idea rather than for the test or scale as a whole [33]. Alpha should be determined for every instance in a summative exam with diverse, case-based questions. More significantly, alpha is based on the statistical presumption that every item in a test has the same real score, known as the "tau equivalent model" [5]. Tau-equivalent reliability is a statistical method that measures how consistent a set of items are in a test. It's also known as Cronbach's alpha, it measures internal consistency and reliability. Alternatively, it presumes that every test item assesses the same latent characteristic on the same scale [4].

Consequently, this assumption is broken because alpha understates the test's reliability if a scale's items are based on many factors or qualities, as shown by factor analysis [5]. The assumption of tau-equivalence will also be broken and dependability will be understated if there are too few test items [17]. Alpha gets closer to a more accurate level of dependability when test items satisfy the tau-equivalent model's assumptions [15]. Since diverse test items would go against the tau-equivalent model's presumptions, Cronbach's alpha is a lower-bound estimate of reliability [25]. An additional evaluation of the tau-equivalent measurement in the data could be necessary if the "standardized item alpha" calculation in SPSS is higher than the Cronbach's alpha.

1.1.3 | Numerical values of cronbach's alpha

As previously mentioned, the alpha value is influenced by the quantity of test items, item interrelatedness, and dimensionality [1]. The allowable alpha ranges from 0.70 to 0.95, according to various studies. A low alpha value may result from diverse constructs, limited questions, or weak item interrelatedness. For instance, certain elements should be changed or eliminated if a low alpha results from a poor correlation between them. The simplest way to identify them is to calculate each test item's correlation with the test's overall score; those with low correlations (around zero) are eliminated. An excessively high alpha might indicate that several items are redundant since they test the same question under a different name. It has been suggested that the alpha value should not exceed 0.90.

2 | Method

Airline Service Quality (ASQ) was modified for this study. These changes were made in light of the study that has been done on the topic. The original ASQ is a private, self-reporting survey used to gauge how satisfied customers are with airline services. In the ASQ, participants are often asked to answer 10 service items using a 4-point Likert scale. Cronbach's alpha has shown that the ASQ is internally consistent. Cronbach's alphas in samples from 23 countries varied from 0.70 to 0.95, with the majority falling in the high.80s, according to Schwarzer's [33] findings. This study's final version of the modified ASQ is a 15-question survey with a 5-point Likert scale.

Five questions were selected at random from the ten that were already in the survey to be reworded and positioned after every two appropriate questions. The scale was modified to include a midpoint choice, resulting in the following scale: According to accepted survey use criteria, the following labels are used: 1= Poor, 2= Fair, 3= Undecided, 4= Good, and 5= Excellent [7]. The following values were obtained by reversing the replies on the worded items to score the instrument: 5= Poor, 4= Fair, 3= Undecided, 2= Good, and 1= Excellent.

3 | Data

A pilot group of thirty-three passengers was used to test this device. The survey consisted of 15 questions with a 5-point Likert scale. Analysis of responses using the SPSS statistical software revealed a Cronbach's alpha of 0.85, indicating good internal consistency. Three months later, 70 passengers participated in a pre-test and post-test study using the same instrument. For this bigger sample, the Cronbach's alpha was 0.86.

4 | Discussion

The modified ASQ's 15 items were dependable and consistent and could be utilized in a study that assessed domestic airline customers' satisfaction levels. The passenger's evaluations could have been impacted by the question order; however, this was not tested by rearranging the questions.

Alreck and Settle [34] state that it would not have been prudent to group all of the questions or to place the questions close to each other. Although a procedure for developing a Likert scale from scratch was described by Maurer and Pierce [24] and Adeniran [2], the survey utilized in this study was based on earlier research. Establish the focus first. Because Likert scales are unidimensional, it's critical to concentrate on the precise thing you're attempting to assess. After that, create a list of possible scale items and ask a group of judges to score them [35].

It was suggested that items with weak ratings and poor linking should be eliminated, or to reduce the number of items. Additionally, the average rating for judges' top and bottom quarters may be obtained, and the difference between the two can then be tested using a t-test. Higher t-value items are good discriminators and need to be retained [36].

There was a limited time to choose and utilize a survey, even if this was a legitimate approach for creating survey topics.The updated ASQ survey was conducted on the same type of passengers (participants) in the original study to identify any unexpected issues with the question. The poll respondents appeared to comprehend the questions and provided insightful responses.

5 | Conclusion

It was satisfying to develop a Likert scale tool that demonstrated internal dependability. High internal consistency was obtained with this modified instrument, created for the Air Service Quality (ASQ) scale. Although it would take longer to create a survey from the start using Trochim's principles than to use an existing instrument, it is possible. Numerous resources are available for those who want to design an instrument for specific research. It is hoped that others will employ this modified ASQ in airline satisfaction or service quality studies.

When assessing tests and surveys, Cronbach's alpha is essential to improve the validity and accuracy of data interpretation. However, it has often been presented without rigorous analysis and without sufficient comprehension and interpretation.

The assumptions used to calculate alpha, the variables affecting its size, and the possible interpretations of its value were all covered in this study. It is anticipated that future researchers will disclose alpha values from their investigations with greater skepticism.

Funding

This paper received no external funding.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Gay, L.R., Mills, G.E. and Airasian, P. (2009). Educational research competencies for analysis and applications. *Pearson, columbus.* https://www.scirp.org/reference/referencespapers?referenceid=1811842

[2] Olaniyi, A. A. (2019). Application of likert scale's type and cronbach's alpha analysis in an airport. *Scholar journal of applied sciences and research*, *2*(4), 1–5. https://innovationinfo.org/articles/SJASR/SJASR-4-223.pdf

[3] Tavakol, M., Mohagheghi, M. A., & Dennick, R. (2008). Assessing the skills of surgical residents using simulation. *Journal of surgical education*, *65*(2), 77–83. https://doi.org/10.1016/j.jsurg.2007.11.003

[4] Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study i: reliability and validity. *Educational and psychological measurement*, *31*(3), 657–674. https://doi.org/10.1177/001316447103100307

[5] Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and psychological measurement*, *60*(3), 361–370. https://doi.org/10.1177/00131640021970592

[6] Wendy K. Adams, C. E. W. (2010). Development and validation of instruments to measure learning of expert-like thinking. *International journal of science education*, *9*, 1289–1312. https://doi.org/10.1080/09500693.2010.512369

[7] Pamela L. Alreck, R. B. S. (1985). *The survey research handbook*. https://books.google.com/books/about/The_Survey_Research_Handbook.html?id=yxFZkB29y60C

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

[9] Likert, R. (1932). A technique for measurement of attitudes. Archives of Psychology. *Archives of psychology*, *22*, 5–55. https://legacy.voteview.com/pdf/Likert_1932.pdf

[10] Barnette, J. J. (2001). Likert survey primacy effect in the absence or presence of negatively-worded items. *Research in the schools*, *8*(1), 77–82. https://eric.ed.gov/?id=EJ663803

[11] Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Sage Publications. https://doi.org/10.4135/9781412986038

[12] B Brown, J. D. (1997). Reliability of surveys. *Testing & evaluation sig newsletter*, *1*(2), 18–21. https://teval.jalt.org/test/bro_2.htm

[13] Chen, Y.-H., Rendina-Gobioff, G., & Dedrick, R. F. (2007). Detecting effects of positively and negatively worded items on a self-concept scale for third and sixth grade elementary students. *Online submission*. ERIC. https://eric.ed.gov/?id=ED503122

[14] Cohen, L., Manion, L., & Morrison, K. (2002). Research methods in education. *Routledge*. https://doi.org/10.4324/9780203224342

[15] Chirayu Auewarakul, Steven M. Downing, R. P. & U. J. (2005). Item analysis to improve reliability for an internal medicine undergraduate OSCE. *Springer link*, *10*, 105–113. https://doi.org/10.1007/s10459-005-2315-3

[16] Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality1. *Educational and psychological measurement*, *37*(4), 827–838. https://doi.org/10.1177/001316447703700403

[17] Cohen, R. J., Swerdlik, M., & Sturman, E. (2012). *Psychological testing and assessment: an introduction to tests and measurement.* McGraw hill. https://perpus.univpancasila.ac.id/repository/EBUPT181396.pdf

[18] Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill. https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1867797

[19] Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied psychological measurement*, *27*(3), 159–203. https://doi.org/10.1177/0146621603027003001

[20] Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *APA psycnet*, *7*(6), 456–461. https://psycnet.apa.org/doi/10.1037/h0074469

[21] Guilford, J. P. (1955). Psychometric Methods. *Cambridge university press*, *20*(2), 163–165. https://doi.org/10.1007/BF02288988

[22] Ray, J. J. (1980). How many answer categories should attitude and personality scales use? *South african journal of psychology*, *10*(1–2), 53–54. https://doi.org/10.1177/008124638001000108

[23] Gliem, J. A., Gliem, R. R., & others. (2003). Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. *Midwest research-to-practice conference in adult, continuing, and community education* (Vol. 1, pp. 82–87). http://pioneer.chula.ac.th/~ppongsa/2900600/LMRM08.pdf

[24] Maurer, T. J., & Pierce, H. R. (1998). A comparison of Likert scale and traditional measures of self-efficacy. *Journal of applied psychology*, *83*(2), 324. https://psycnet.apa.org/doi/10.1037/0021-9010.83.2.324

[25] Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98. https://psycnet.apa.org/doi/10.1037/0021-9010.78.1.98

[26] Leganger, A., Kraft, P., & Roysamb, E. (2000). Perceived self-efficacy in health behaviour research: Conceptualisation, measurement and correlates. *Psychology and health*, *15*(1), 51–69. https://doi.org/10.1080/08870440008400288

[27] Green, S. B., & Thompson, M. S. (1989). Structural equation modeling in clinical psychology research. , 138 *Handbook of research methods in clinical psychology* (Vol. 138). http://dx.doi.org/10.1002/9780470756980.ch8

[28] Cronbach, L. J. (1949). *Essentials of psychological testing.* Harper. https://psycnet.apa.org/record/1950-00647-000

[29] Randall, D. M., & Fernandes, M. F. (1991). The social desirability response bias in ethics research. *Journal of business ethics*, *10*, 805–817. https://doi.org/10.1007/BF00383696

[30] Bland, J. M., & Altman, D. G. (1997). Statistics notes Cronbach's alpha. *Bmj*, *314*(7080), 572. https://doi.org/10.1136/bmj.314.7080.572

[31] Brown, J. D. (2000). What issues affect Likert-scale questionnaire formats. *Shiken: JALT testing & evaluation sig newsletter*, *4*(1). 27-30. https://teval.jalt.org/test/PDF/Brown7.pdf

[32] Garland, R. (1991). The mid-point on a rating scale: is it desirable. *Marketing bulletin*, *2*(1), 66–70. http://scorevoting.net/MB_V2_N3_Garland.pdf

[33] Schwarzer, R. (2002). The general self-efficacy scale (GSE), department of health. *Psychology*. http://userpage.fu-berlin.de/~health/engscal.htm

[34] Alreck, P.L. and Settle, R. B. (2003). Survey research handbook. 3rd edition, *McGraw-Hill/Irwin, New York.* https://www.scirp.org/reference/referencespapers?referenceid=2361388

[35] Punuri, S. B., Kuanar, S. K., Kolhar, M., Mishra, T. K., Alameen, A., Mohapatra, H., & Mishra, S. R. (2023). Efficient Net-XGBoost: An implementation for facial emotion recognition using transfer learning. *Mathematics*, *11*(3), 1–24. https://doi.org/10.3390/math11030776

[36] Guru, A., Mohanta, B. K., Mohapatra, H., Al-Turjman, F., Altrjman, C., & Yadav, A. (2023). A survey on consensus protocols and attacks on blockchain technology. *Applied sciences*, *13*(4), 1-21. https://doi.org/10.3390/app13042604